



# The Routledge Handbook of Artificial Intelligence and Philanthropy

Edited by Giuseppe Ugazio and Milos Maricic



# THE ROUTLEDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE AND PHILANTHROPY

*The Routledge Handbook of Artificial Intelligence and Philanthropy* acts as a catalyst for the dialogue between two ecosystems with much to gain from collaboration: artificial intelligence (AI) and philanthropy. Bringing together leading academics, AI specialists, and philanthropy professionals, it offers a robust academic foundation for studying both how AI can be used and implemented within philanthropy and how philanthropy can guide the future development of AI in a responsible way.

The contributors to this *Handbook* explore various facets of the AI-philanthropy dynamic, critically assess hurdles to increased AI adoption and integration in philanthropy, map the application of AI within the philanthropic sector, evaluate how philanthropy can and should promote an AI that is ethical, inclusive, and responsible, and identify the landscape of risk strategies for their limitations and/or potential mitigation. These theoretical perspectives are complemented by several case studies that offer a pragmatic perspective on diverse, successful, and effective AI-philanthropy synergies.

As a result, this *Handbook* stands as a valuable academic reference capable of enriching the interactions of AI and philanthropy, uniting the perspectives of scholars and practitioners, thus building bridges between research and implementation, and setting the foundations for future research endeavors on this topic.

**Giuseppe Ugazio** is an Associate Professor in Behavioral Philanthropy and Finance at the Geneva Finance Research Institute, faculty of economics and management of the University of Geneva. He holds two doctorates, one in philosophy and the other in neuro-economics and studies the neuropsychological mechanisms that influence complex human social behavior. He is pioneering research on the use of AI to unveil the potential of this technology to support the strategies and operations of philanthropy organizations. These projects include the development of a big data-driven tool using natural language processing to unveil synergies among philanthropic organizations in Switzerland.

**Milos Maricic** is an expert at the intersection of philanthropy, finance, and AI. As the Founder of the Altruist League, a Geneva-based global philanthropy consultancy, he pioneered the use of AI in investment sourcing, fund distribution, and portfolio management. In 2021, he co-authored the book *Fixing Philanthropy*, which proposes a more efficient framework for global giving to address systemic issues. He has represented the venture capital industry in key international discussions on AI regulation. As a former humanitarian executive, he has spoken at major events, advocating for grassroots, citizen-led organizations.



# Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# THE ROUTLEDGE HANDBOOK OF ARTIFICIAL INTELLIGENCE AND PHILANTHROPY

*Edited by Giuseppe Ugazio and Milos Maricic*



**HANDBOOK OF ARTIFICIAL INTELLIGENCE AND PHILANTHROPY**



**UNIVERSITÉ  
DE GENÈVE**

GENEVA SCHOOL OF ECONOMICS  
AND MANAGEMENT



**UNIVERSITÉ  
DE GENÈVE**

GENEVA CENTRE  
FOR PHILANTHROPY

 **Routledge**  
Taylor & Francis Group  
LONDON AND NEW YORK

Designed cover image: Getty Images/Leyn

First published 2025

by Routledge

4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by Routledge

605 Third Avenue, New York, NY 10158

*Routledge is an imprint of the Taylor & Francis Group, an informa business*

© 2025 selection and editorial matter, Giuseppe Ugazio and Milos Maricic;  
individual chapters, the contributors

The right of Giuseppe Ugazio and Milos Maricic to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

The Open Access version of this book, available at [www.taylorfrancis.com](http://www.taylorfrancis.com), has been made available under a Creative Commons Attribution-Non Commercial-No Derivatives (CC-BY-NC-ND) 4.0 license.

Any third party material in this book is not included in the OA Creative Commons license, unless indicated otherwise in a credit line to the material.

Please direct any permissions enquiries to the original rights holder.

*Trademark notice:* Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

*British Library Cataloguing-in-Publication Data*

A catalogue record for this book is available from the British Library

*Library of Congress Cataloging-in-Publication Data*

Names: Ugazio, Giuseppe, editor. | Maricic, Milos, editor.

Title: The Routledge handbook of artificial intelligence and philanthropy / edited by Giuseppe Ugazio and Milos Maricic.

Description: Abingdon, Oxon; New York, NY: Routledge, 2025. |

Includes bibliographical references and index.

Identifiers: LCCN 2024034942 (print) | LCCN 2024034943 (ebook) |

ISBN 9781032743011 (hardback) | ISBN 9781032743028 (paperback) |

ISBN 9781003468615 (ebook)

Subjects: LCSH: Charities. | Endowments. | Artificial intelligence.

Classification: LCC HV25 .R68 2025 (print) | LCC HV25 (ebook) |

DDC 361.7/4—dc23/eng/20240918

LC record available at <https://lcn.loc.gov/2024034942>

LC ebook record available at <https://lcn.loc.gov/2024034943>

ISBN: 978-1-032-74301-1 (hbk)

ISBN: 978-1-032-74302-8 (pbk)

ISBN: 978-1-003-46861-5 (ebk)

DOI: 10.4324/9781003468615

Typeset in Times New Roman  
by codeMantra

# CONTENTS

<i>Acknowledgments</i>	ix
<i>List of Figures</i>	x
<i>List of Tables</i>	xiii
<i>List of appendixes, table of interviews, and boxes</i>	xv
<i>List of Contributors</i>	xvi
<i>Foreword: From AI-empowered philanthropy to philanthropy-driven AI</i> Luciano Floridi	xxiii
Editors' introduction to the volume <i>Giuseppe Ugazio and Milos Maricic</i>	1
<b>PART I</b>	
<b>AI applications in the philanthropic sector</b>	<b>9</b>
1 The relationship between AI and philanthropy: From historical roots to modern convergence <i>Camilla Della Giovampaola and Maria Cristiana Tudor</i>	11
2 Measuring meaningful change: AI-enhanced impact measurement in philanthropy <i>Simone Bartalucci, Antonia Muhr, Sina Sauer and Volker Then</i>	29
3 PHIL4DEV: A text-based machine learning model to compare philanthropic funding across the world <i>Nelson Amaya, Harry de los Rios and Madeleine Lessard</i>	51
4 Fast-tracking the use of AI in everyday philanthropy <i>Stefan Schöbi</i>	61

Contents

5	Applying diverse AI tools to transform philanthropic operations: Insights from the for-profit sector <i>Prity Khastgir and Shweta Shalini</i>	76
6	The use of AI and technology by civil society organizations and its international implications <i>Anita Budziszewska and Oto Potluka</i>	94
7	The impact of artificial intelligence and information technology on philanthropic organizations: Case studies of non-profit and cultural associations <i>Luca Barzanti, Lia Benvenuti and Enrico Gaudenzi</i>	107
8	On the impact of AI-empowered, gaming-based virtual worlds on philanthropy <i>Marc Schipper, Manouchehr Shamsrizi and Adalbert Pakura</i>	132
9	Toward a framework for Responsible AI in storytelling for nonprofit fundraising <i>Marta Herrero and Shauna Concannon</i>	145
<b>PART II</b>		
<b>Philanthropies' regional AI adoption, readiness, and applications</b>		<b>157</b>
10	Artificial intelligence, machine learning, and data science philanthropy: Case studies of a purposive classification of philanthropic missions <i>Patricia Snell Herzog</i>	159
11	Data science and AI among philanthropic foundations in Europe <i>Sevda Kilicalp, Jack O'Neill and Daniel Spiers</i>	172
12	Digitalization of Swiss non-profit foundations: The potential role of AI from a cross-sectoral perspective <i>Aline Kratz-Ulmer and Hubert Halopé</i>	192
13	Technological readiness of Asia's social sector for the adoption and use of artificial intelligence <i>Kithmina V. Hewage</i>	205

14	Digital philanthropy in China: How internet fundraising platforms and artificial intelligence are transforming non-profit governance <i>Bertram Lang</i>	221
15	A case study on AI usage for collecting philanthropy data in the Western Balkans <i>Nikola Milinković and Marko Galjak</i>	240
16	Optimizing philanthropic investment with AI: A case study of the Altruist League <i>Milos Maricic</i>	257
17	The Spandows: Pioneering AI in family philanthropy and sustainable business <i>Malgorzata Smulowitz and Peter Vogel</i>	275
18	Digital stunt philanthropy: Mechanisms, impact, and ethics of using social media influencing for the greater good <i>Monica Lea and Lucia Gomez</i>	287
<b>PART III</b>		
<b>Philanthropy for AI development and regulation</b>		<b>307</b>
19	Navigating risk complexity associated with data philanthropy for AI <i>Rahul Jha</i>	309
20	From margin to mainstream: Moving philanthropy to reshape our AI-enabled future <i>Yolanda Botti-Lodovico and Vilas Dhar</i>	328
21	Altruistic collective intelligence for the betterment of artificial intelligence <i>Thomas Maillart, Lucia Gomez, Mohanty Sharada, Dipam Chakraborty and Sneha Nanavati</i>	344
22	Hand out or help out: A resource-based view of AI in philanthropy <i>Joe Wheeler</i>	361
23	Philanthropy's urgent opportunity to create the Interim International AI Institution (IIAII) <i>David Evan Harris and Anamitra Deb</i>	374



Contents

24	On foundations and foundation models: What lessons can AI and philanthropy learn from one another? <i>Diana Acosta-Navas</i>	393
25	The AI extreme risk mitigation philanthropic sector: A philanthropic ecosystem at the forefront of AI <i>Siméon Campos and Daniel S. Schiff</i>	408
<b>PART IV</b>		
<b>Ethics, AI, and philanthropy</b>		<b>431</b>
26	How can philanthropy promote ethical, inclusive, and responsible AI development? Lessons from impactIA Foundation <i>Laura Tocmacov</i>	433
27	Guided choices: The ethics of using algorithmic systems to shape philanthropic decision-making <i>Rhodri Davies</i>	451
28	Shaping the ethical and inclusive AI revolution: Five roles for philanthropies <i>Ulla Jasper, Siddhartha Jha and Stefan Germann</i>	471
29	Getting to heaven: What teaching AI teaches us about ourselves <i>Elizabeth A.M. Searing and Donald R. Searing</i>	486
30	Why philanthropy should embrace the ideological struggle shaping artificial general intelligence: A preliminary theological-political analysis <i>Ezekiel K. Takam</i>	503
31	AI disruptions in philanthropy: A multi-scale model of ethical vigilance <i>Charles Sellen and Joost Mönks</i>	520
32	AI and philanthropy: How can they elevate each other? <i>Ravit Dotan</i>	539
	Closing reflections and future directions <i>Camilla Della Giovampaola, Lucia Gomez, Hubert Halopé, Maria Cristiana Tudor and Giuseppe Ugazio</i>	554
	<i>Index</i>	559

# ACKNOWLEDGMENTS

Editing this *Handbook* was an incredibly rewarding yet challenging journey. This *Handbook* was made possible thanks to the collaborative effort and the immense support we received throughout the book's making from a tireless research team. Camilla Della Giovampaola's unwavering dedication to this project was matched only by her exemplary team leadership, creating a positive and motivating working environment. Her commitment was indispensable during the challenging phases of this journey. Lucia Gomez's passion for everything AI-related was instrumental in ensuring that this *Handbook* would cover as many angles of this technology as possible. Maria Tudor's unique attention to detail and desire to perfect every contribution was fundamental to ensuring the book could include diverse manuscripts from both academics and practitioners with consistent rigor. Hubert Halopé's private sector expertise and understanding of its dynamics helped identify several angles that greatly impacted and inspired the research-focused team. Working with this brilliant-minded group has been a pleasure and an honor. It has allowed us to bring together different skills and perspectives. We are also grateful to all the contributors who joined us on this journey and shared their expertise and knowledge.

The research and academic contributions are only one side of the coin. We could not have completed this project without the constant and thorough support of the Geneva Center for Philanthropy's splendid team—Henry Peter, Mara de Monte, and Margaux Casagrande. Of equal importance was the meticulous work of our editing project manager, Anne-Françoise Ritter, who revised and refined every page of this book.

We would also like to take this opportunity to thank the Routledge team, especially Kristina Abbotts and Christiana Mandizha, for their invaluable support and their kind availability.

Finally, all the activities entailed by this *Handbook* and other parallel efforts within the broader AI and philanthropy project were made possible by the generous financial support provided by the Botnar Foundation and another strategic partner of the Geneva Center for Philanthropy, who wishes to remain anonymous. Their support was a crucial investment in our shared vision and the future of philanthropy.

# FIGURES

1.1	Examples of AI for Social Good (adapted from Tudor et al., 2024)	23
2.1	Basic impact modeling steps (own work)	31
2.2	Impact chain structure according to the IOOI model (own work, based on Then et al., 2017. Social Return on Investment Analysis: Measuring the Impact of Social Investment, Springer International Publishing, reproduced with permission from SNCSC)	33
2.3	Logical scheme of the first approach (own work)	36
2.4	Logical scheme of the second approach (own work)	37
2.5	Synthetic impact chain flows and correlations according to the IOOI model (own work)	40
2.6	Process flow for identifying and selecting indicators to measure the quality of life of CPC stakeholders (own work)	41
2.7	Logical scheme for utilizing a similarity tool in the CPC case study (own work)	44
3.1	Overview of OECD CRS purpose code classification system as of 2021	53
3.2	PHIL4DEV workflow	54
3.3	PHIL4DEV data workflow	55
3.4	Histogram of PHIL4DEV performance across classifications	56
3.5	Public accessibility to PHIL4DEV model using Shiny App	57
4.1	Innovation adoption curve by Everett Rogers, see Rogers (2003, p. 281)	62
4.2	Four states of innovation diffusion, including confidence thresholds	64
4.3	Empirical mapping of AI innovation adoption stages in Swiss philanthropy: grantseeking vs. grantmaking organizations	67
5.1	Schematic representation of an automated electronic impact platform for sustainable investment in socially responsible endeavors (photograph from the patent application)	78
5.2	Block diagram illustrating the network architecture of the next-generation influence-based crowdfunding infrastructure (photograph from the patent application)	84
5.3	Block diagram illustrating the steps involved in implementing the influence-based crowdfunding method (photograph from the patent application)	85

## Figures

5.4	Flow diagram illustrating the method of developing, attracting, and scaling an innovation (photograph from the patent application)	87
5.5	Flow diagram of the artificial intelligence-based method for processing raw data (photograph from the patent application)	90
7.1	The giving pyramid. The pyramid explains the Donors' segmentation	109
7.2	The architecture of KFM. Each part of the system is highlighted, with the explication of all the interactions	112
7.3	The mathematical model of dynamic evaluation of strategies in KFM. In this linear algebra model, "card" denotes the cardinality (i.e., the number of elements) of the considered set	113
7.4	Fuzzy evaluation of the gift probability in KFM. The fuzzy structure of the variables is shown	113
7.5	Evaluation of the gift probability in FS. The role of each mathematical object of the evaluation is explicated	115
7.6	Evaluation of the amount in FS. The role of each mathematical object of the evaluation is explicated	115
7.7	The fuzzification process in FS. The mathematical process implemented by the fuzzy system is specified	116
7.8	Graphical user interface. The interaction mask between the system and the user is shown. In particular, the input parameters are classified by their meaning	117
7.9	First ranking phase. The expected gift probability and amount are computed for each Donor	118
7.10	Fuzzy aggregation and whole ranking. The score of each Donor is computed	118
7.11	Campaign results. The expected total return of the campaign and the robustness of the result are calculated	119
7.12	Campaign results with budget saving. The robustness of the result is enhanced	119
7.13	Supply and demand quantities as a function of price. The equilibrium point is shown	122
7.14	Implementation of <i>Problem 1</i> in Excel with the <i>Goal Seek</i> tool	123
7.15	Solving <i>Problem 1</i> using the <i>Goal Seek</i> tool	123
7.16	Model of <i>Problem 2</i> , point (a). The graphical model facilitates the resolution	124
7.17	Solving <i>Problem 2</i> , point (a) using Excel. Use of Excel functions <i>FV</i> and <i>SUM</i> is highlighted	125
7.18	Evolution of the model in relation to a new investment opportunity. The graphical model is very effective to capture financial changes	125
7.19	Solving <i>Problem 2</i> , point (b) using Excel. Use of Excel functions <i>FV</i> and <i>SUM</i> is highlighted	126
7.20	Model for overall interest rate computation. The graphical model highlights the meaning of the Internal Rate of Return (IRR)	126
7.21	Solving <i>Problem 2</i> , point (c) using Excel. The use of the Excel function <i>RATE</i> is highlighted	126
7.22	Modeling of the decision problem. The graphical model is very effective in showing the choice variable	127
7.23	Solving <i>Problem 2</i> , point (d) using Excel	127
7.24	The overall solution of <i>Problem 2</i> . An overview of the whole Problem's resolution is given	128
7.25	Interaction model between the involved software programs. Structure of the Demo Web Interactive Financial Lab	129

## Figures

8.1	Individual (A), co- (B), and joint (C) perception (adapted from Deroy & Longin, in press)	136
10.1	Relationships among AI, machine learning, and data science technology definitions	160
10.2	University building construction sign says philanthropy at work	165
10.3	Framework for the three roles of philanthropy in technology for good activities	168
11.1	Foundations' engagement with AI either for internal or external activities	179
11.2	Entry points for foundations to support AI enhancement and external initiatives	180
11.3	Overview of internal, external, and absent data roles in foundations	182
11.4	Thematic distribution of AI applications in foundations	184
13.1	AI readiness framework for the social sector	208
13.2	Percentage of SDOs with difficulties recruiting and retaining staff	213
13.3	Percentage of SDOs with a cybersecurity strategy	213
14.1	Oligopolistic structure of Chinese online fundraising	224
14.2	Distribution of Tencent Charity project launches over time	228
14.3	Fundraising trends on Tencent Charity: platform expansion and soaring inequality	230
14.4	Independent organizations and individuals outcompeted by GONGOs	230
15.1	Giving Balkans interactive data visualization and analysis web application	244
15.2	CiviGraph—a social network analysis tool built into the Giving Balkans app leveraging the relational data	244
15.3	The process of collecting data for Giving Balkans database	246
15.4	The AI-assisted process of collecting data for Giving Balkans database	249
16.1	Example API call for embeddings	269
17.1	Convergence during model training on the IFRC data	280
18.1	Schematic depicting the process of video selection for recommendation	289
18.2	Figure summarizing seven dimensions underlying audience engagement with social media content that aligns with philanthropy-like content	290
18.3	Natural Language Processing indicates that little overlap exists between MrBeast and Beast Philanthropy YouTube channels' content	298
18.4	Top differential content keywords reveal a common philanthropy content and a differing use of entertainment and marketing	299
18.5	Video Statistics for MrBeast and Beast Philanthropy YouTube channels	300
19.1	Collaboration governance framework	316
21.1	Bipartite network of relationships between teams and submissions as weighted by precision scoring	351
21.2	Network slice by submission rounds, in challenge and benchmark phases	352
21.3	Determinants of collective intelligence in AICrowd Food Recognition Challenge	353
21.4	Detail on the relation between the variables used for modeling	354
31.1	Three levels to apprehend AI in philanthropy	522
31.2	Four key reasoning steps to apprehend AI in philanthropy	523
31.3	Multi-scale model to apprehend AI in philanthropy	524
31.4	Different scales of disruptions brought by AI in philanthropy	525
31.5	Several levels of impacts on stakeholders caused by AI in philanthropy	530
31.6	Several levels of ethical questions raised by AI in philanthropy	532
31.7	Several levels of best practice and safeguards to properly handle AI in philanthropy	532
31.8	Multi-scale model to apprehend AI in philanthropy (with suggested use)	534

# TABLES

1.1	Primary, secondary, and regulatory AI4SG stakeholders	19
2.2	Criteria used to analyze CPC stakeholder quality of life indicators (own work)	42
2.3	Ranking of CPC's quality of life measurement frameworks based on analysis criteria scores by stakeholder (own work)	43
2.4	Dimensions of quality of life resulting from an analysis of existing frameworks using the similarity tool grouped by stakeholder (own work)	44
4.2	Expected use of AI, n = 89	65
4.3	Inhibitors, n = 89	66
4.4	Feasibility and desirability of the most common use cases for AI in everyday philanthropy, n = 20	71
4.5	Concerns and chances of technology implementation	73
6.1	Financial management in digital civil society in Switzerland	97
7.2	Some Donors' characteristics with respect to their segmentation	114
7.3	Statistics regarding Donors' profiles	114
10.1	Organizations by source	161
10.2	Tech for-good philanthropy case studies	168
13.1	Staff access to computers and/or tablets	210
13.2	Collection and storage of data digitally	211
13.3	Proportion of Asian SDOs that increased their use of digital technology and online tools	215
16.1	The principles of systemic philanthropy (adapted from Fernandez et al., 2020)	258
16.2	The criteria for inclusion of an organization in the Altruist League's dataset	260
16.3	Parameters tracked in the Altruist League's dataset	261
17.1	A synopsis of key learnings and recommendations from the Amesto case	284
18.1	Key components of digital stunt philanthropy	294
19.1	Chapter objectives and research questions	310
19.2	DP – outline of key DP scenarios, detailing stakeholder roles and how IOs facilitate these efforts (George et al., 2019a)	312

*Tables*

19.3	UNESCO ethics of AI principles ( <i>Recommendation on the Ethics of Artificial Intelligence, 2022</i> ) mapping to EIAI aspects	318
19.4	Spectrum of copyright levels through Creative Commons suite of licenses ( <i>About CC Licenses, 2020</i> )	320
19.5	Case study questions	321
19.6	Recommendations based on applying DP principles to the case study	323

# APPENDIXES, TABLE OF INTERVIEWS, AND BOXES

## **Appendixes**

21.1 A1	Food Recognition Challenge Case Study	358
21.2 A2	Relation between variables used for modeling	360
31.1	G20's Civil 20 worldwide working group	538

## **Table of interviews**

2.1 A	List of quality of life and outcome frameworks	45
2.2 B	Framework dimensions per stakeholder	47
2.3 C.1	Children's framework similarity	48
2.3 C.2	Informal carers' framework similarity	48
2.3 C.3	Siblings' framework similarity	48
14.1	Table of interviews	235

## **Boxes**

11.1	Internal Barriers	186
31.1	Stakeholders checklist	529
31.2	Ethics checklist	531
31.3	Best practice checklist	532



# CONTRIBUTORS

**Diana Acosta-Navas** is an Assistant Professor of Business Ethics at Loyola University Chicago. She holds a PhD in Philosophy from Harvard University and served as a Lecturer in Ethics and Public Policy at the Harvard Kennedy School.

**Nelson Amaya** is an Economist specializing in applied statistics and currently leads the research at the OECD Centre on Philanthropy in Paris. He holds a master's degree in Public Administration from Columbia University and a BA in Economics from Universidad de los Andes.

**Simone Bartalucci** is a Senior Associate at Fondazione AIS, working on projects related to impact measurement of philanthropic and financial investments. Simone holds an MSc in Management Engineering from the Politecnico di Milano, focusing on sustainability and social impact.

**Luca Barzanti** is an Associate Professor in the Department of Mathematics at the University of Bologna, Italy. His current research interests include mathematical models and decision support systems for nonprofit organizations, fundraising management and public value measurement, and IT tools for financial education.

**Lia Benvenuti** is an Adjunct Professor of Mathematics at the University of Bologna (Forlì School), Italy, and General Manager of Techne, a public nonprofit training body. Her current research interests include financial education and decision support systems for nonprofit organizations.

**Yolanda Botti-Lodovico** is the Policy and Advocacy Lead at the Patrick J. McGovern Foundation. Yolanda develops strategic recommendations for rights-based AI governance and digital justice. Her expertise lies at the intersection of human rights, policy, and innovation.

**Anita Budziszewska**, PhD, is a member of the Department of Diplomacy and International Institutions, Faculty of Political Science and International Studies, the University of Warsaw; Swiss Government Excellence Scholarship holder at the University of Geneva, Geneva Centre for Philanthropy. Research areas include culture and philanthropy in international relations and international organizations.

## *Contributors*

**Siméon Campos** is the CEO and Founder of SaferAI, a French nonprofit organization dedicated to AI risk management. In this context, he studies the risks of AI and how to govern them.

**Dipam Chakraborty** is a Data Scientist at H2O AI. He is experienced in Deep Learning, Computer Vision, and Generative AI. Previously, he worked at Alcrowd where he organized a large number of research competitions centered around machine learning.

**Shauna Concannon** is an Assistant Professor in Computer Science and Digital Humanities at Durham University, UK. Using an interdisciplinary approach, their work examines how humans interact with AI systems and emerging technologies' societal and ethical impacts.

**Rhodri Davies** is a Pears Research Fellow at the Centre for Philanthropy at the University of Kent and the Founder and Director of the think tank Why Philanthropy Matters. He is also the host of the “Philanthropisms” podcast.

**Anamitra Deb** is the Managing Director at Responsible Tech at Omidyar Network and a Senior Fellow at the Munk School of Global Affairs & Public Policy at the University of Toronto.

**Camilla Della Giovampaola** is a Doctoral Researcher at the Geneva Graduate Institute (IHEID), Switzerland, working on the institutionalization of philanthropy. Camilla holds a master's in Development Studies from IHEID and a Bachelor of Arts in Politics and International Relations from Royal Holloway University of London.

**Vilas Dhar** is President of the Patrick J. McGovern Foundation. A technologist, scholar, and human rights advocate, Vilas serves on the UN High-Level Advisory Body on AI and is committed to advancing civil society representation in global AI decision-making.

**Ravit Dotan**, PhD, is an AI ethics advisor, researcher, and speaker. Her specialty is helping tech companies, investors, and procurement professionals develop responsible AI approaches.

**Marko Galjak** is a Research Fellow at the Institute of Social Science and serves as the Technology Director for Catalyst Balkans. His background is in Demography and Computational Social Science.

**Enrico Gaudenzi** is an External Consultant of Techne, a public, nonprofit training body, and CEO of the Company Recare S.r.l. His current research interests include financial education and decision support systems for nonprofit organizations.

**Stefan Germann** is an Advisor to Fondation Botnar and former CEO (2017–2023). He is a global expert in children's health and well-being with over 30 years of experience, with a singular passion for the progressive realization of children's rights and the impact of the digital age.

**Lucia Gomez** is currently a Postdoctoral Researcher at the University of Geneva Finance Research Institute. Holding a Bachelor in Psychology and a PhD in Neurosciences, her interdisciplinary research expertise has Artificial Intelligence as its core.

**Hubert Halopé** is a PhD candidate at the University of Geneva, researching the value creation of AI across sectors. Hubert has held various roles in technology management and strategy, and

## *Contributors*

holds a dual master's degree in International Management from UCD Michael Smurfit Graduate Business School and Bocconi University (CEMS), and a Bachelor in Economics from Maastricht University.

**David Evan Harris** is a Chancellor's Public Scholar at the University of California, Berkeley and Senior Research Fellow at the International Computer Science Institute. He writes and speaks publicly about AI, misinformation, social media, elections and technology policy.

**Marta Herrero** is a Senior Lecturer and Head of Creative and Cultural Industries Management at the University of York's School of Arts and Creative Technologies. Her research focuses on using digital technologies to support nonprofit fundraising, and she has led projects on the use of VR and interactive documentaries.

**Kithmina V. Hewage** is a Senior Advisor at the Center for Asian Philanthropy and Society (CAPS). He works on policy engagement and design with governments, philanthropists, corporations, and social sector organizations across 17 Asian economies. Kithmina is a graduate of Johns Hopkins University and University College London (UCL).

**Ulla Jasper** is the Governance and Policy Lead at Fondation Botnar. She holds a PhD in Political Science from the University of St Gallen. At Fondation Botnar, she is responsible for managing policy work and coordinating external relations and partnerships.

**Rahul Jha** holds degrees in engineering and business administration and has diverse experience in various roles at an international organization in Geneva. An advocate for neurodiversity and specialist in cybersecurity and digital transformation, he is pursuing a PhD in Management at GSEM alongside his career.

**Siddhartha Jha** is the AI and Digital Innovation Lead at Fondation Botnar. He combines his engineering education from the Indian Institute of Technology and ETH Zurich with global experience in the technology innovation sector to support the foundation's grant-making and strategic initiatives.

**Prity Khastgir** is a Director at Tech Corp International Strategist (TCIS, India) and a registered Patent Attorney in India who specializes in IPRs, cyber laws, and international commercial mediation. With over 18 years of experience, she advises conglomerates worldwide on combating IP infringement and is a WIPO tutor for Blockchain and AI patents.

**Sevda Kilicalp** is Head of Research and Knowledge Development at Philea Philanthropy Europe Association. She has a PhD in Philanthropic Studies from the Indiana University Lilly Family School of Philanthropy and a master's degree in Philanthropic Studies and Social Entrepreneurship from the University of Bologna.

**Aline Kratz-Ulmer** is an Attorney-at-Law at her own law firm AKU Anwaltsbüro Kratz-Ulmer in Zurich (Switzerland), where she advises all types of foundations. She is also Academic Fellow at the University of Geneva since 2020. She holds a PhD and master's degree in Law from the University of Zurich and received her bar admission in Zurich in 2010.

## *Contributors*

**Ezekiel Kwetchi Takam** is a PhD candidate at the University of Geneva and a visiting PhD Researcher at the Leverhulme Centre for the Future of Intelligence, at the University of Cambridge.

**Bertram Lang** is a Political Scientist and Academic Coordinator of the Interdisciplinary Center for East Asian Studies (IZO) at Goethe University Frankfurt. His recent work centers on the transnational politics of philanthropy and civil society, with an empirical focus on China.

**Monica Lea** is a PhD candidate at the University of Nebraska at Omaha studying Public Administration. With an MS in Family, Youth, and Community Sciences, her research focuses on nonprofit and digital governance as well as emotional labor.

**Madeleine Lessard** is a Research and Policy Analyst currently working at the OECD Centre on Philanthropy in Paris. She holds a master's degree in International Development from the Institut d'études politiques de Paris (Sciences Po), and a BA in Liberal Arts from Thomas Aquinas College.

**Thomas Maillart** is a Senior Lecturer at the Geneva School of Economics and Management, specializing in complex techno-social systems and the dynamics of collective intelligence. Holding a PhD from ETH Zurich, Maillart has extensively researched and implemented collective intelligence in education and innovation.

**Milos Maricic** specializes in the intersection of finance and AI. He co-founded the Altruist League. As an investor, he represents the venture capital industry in AI regulation discussions in Brussels and Washington, DC. He leads Executive AI, a network of European CEOs committed to AI's effective and ethical use.

**Nikola Milinković** is an Automation Specialist at Catalyst Balkans, bringing solid mathematics expertise. He has contributed significantly to various research projects, such as devising numerical algorithms for the Department of Hydraulic and Environmental Engineering and the Faculty of Civil Engineering in Belgrade.

**Joost Mönks** is an Independent Philanthropy Advisor engaged in promoting the ethical use and development of AI. He most recently served as International Coordinator of the G20 India 2023 working group on Education and Digitalization. Previously, he lectured at the University of Geneva on regional philanthropy and was the Executive Director of a Humanitarian Quality Assurance Agency in Geneva.

**Antonia Muhr** works with Fondazione AIS as a Consultant on Impact Measurement in Philanthropy. She is a PhD student at the Institute for Nonprofit Management at the Vienna University of Economics and Business, focusing on philanthropy and social innovation.

**Sneha Nanavati** leads the Communications and Marketing efforts at AICrowd. In her role at AICrowd, Sneha is dedicated to supporting a platform that makes significant strides in making technology better and accessible while also highlighting stories from a vibrant community of thinkers.

**Jack O'Neil** is the Data Officer for the Philanthropy Europe Association (Philea). He holds an undergraduate degree in French, German, and Irish and a master's degree in Information Systems, both of which were attained in his home country of Ireland.

## *Contributors*

**Adalbert Pakura** is an Interdisciplinary Scientist and Practitioner. He has researched and taught in gaming, digitalization, and education, focusing on how these domains intersect to enhance learning and engagement in the digital age.

**Oto Potluka**, PhD, is a Senior Researcher at the Center for Philanthropy Studies, University of Basel (Switzerland) and Department of Management, Prague University of Business and Economics. His research relates to regional and economic development and civil society, as well as public expenditure programs in regional development, especially those co-financed by EU cohesion policy.

**Harry de los Rios** is a Data Scientist at Ernst & Young's Artificial Intelligence Innovation Centre. He is also a PhD student in Complex Systems Networks and holds a master's degree in Computational engineering and Mathematics and a bachelor's degree in Theoretical Physics.

**Sina Sauer** is a Senior Associate at Fondazione AIS. She works in the field of impact investing and impact measurement. She is pursuing a PhD in Economics at the University of Heidelberg, focusing on impact finance and impact measurement.

**Daniel S. Schiff** is an Assistant Professor of Technology Policy at Purdue University and Co-Director of the Governance and Responsible AI Lab (GRAIL) in the USA, where he studies the governance of AI, as well as AI's societal implications.

**Marc Schipper**, PhD, is a professor, psychologist, and neuroscientist. He earned his doctorate at the Center for Cognitive Sciences at the University of Bremen. Currently, he serves at the APOLLON University of Health Management, focusing on perceptual psychology, gerontopsychology, and counseling psychology.

**Stefan Schöbi** is the CEO of StiftungSchweiz, a Swiss philanthropy platform. Previously, he established and managed the Migros Pioneer Fund, one of Switzerland's largest corporate funds. He earned an MBA in Marketing from Zurich University of Applied Sciences and certificates from INSEAD and Stanford University.

**Donald R. Searing** founded multiple software development and consulting firms focusing on automating complex decision-making processes and workflows across industries. Dr. Searing is the Vice President of Product & Engineering at Sagility Health and Principal Scientist at Syncere Systems.

**Elizabeth A.M. Searing** is an Assistant Professor of Public and Nonprofit Management at the University of Texas in Dallas. Dr. Searing's primary research focus is the financial management of nonprofit and social enterprise organizations, but she also conducts work on resilience, charity data, and comparative social economy more broadly.

**Charles Sellen**, PhD, is the Founder of SmartPhil. He is affiliated with Carleton University (Ottawa) and PhiLab (UQAM Montréal) in Canada. He previously served as the inaugural "Global Philanthropy Fellow" at the Indiana University Lilly Family School of Philanthropy (2019–2021) and a Fulbright "NGO Leader" visiting from France to the USA.

**Shweta Shalini** is a Senior Associate at Tech Corp International Strategist (TCIS, India). With over 11+ years of experience in global technology laws, Shweta is a highly skilled technology lawyer specializing in AI and Blockchain consultancy.

**Manouchehr Shamsrizi** is “among the most publicly prominent voices of Germany’s younger generation” (*Washington Post*) and “well positioned to assess emerging trends” (*Monocle*). Shamsrizi co-founded Humboldt-Universität’s gamelab.berlin, was an Ariane de Rothschild Fellow at Cambridge, and a Global Justice Fellow at Yale.

**Mohanty Sharada** is the CEO and Founder of AICrowd. His research focuses on numerous problems at the intersection of AI and health, with a strong interest in reinforcement learning. In his current role, he focuses on building better engineering tools for AI researchers and making research in AI accessible to a larger community of engineers.

**Malgorzata Smulowitz** is a Research Fellow at IMD’s Debiopharm Chair for Family Philanthropy. She has published articles in peer-reviewed academic journals such as *Human Relations* and practitioner-oriented work on numerous topics. She co-authored the award-winning book *Family Philanthropy Navigator*.

**Patricia Snell Herzog**, PhD, is an Associate Professor at the Indiana University Lilly Family School of Philanthropy and Affiliate Faculty in the Human-Computer Interaction Department of the IU, the Luddy School of Informatics, Computing, and Engineering in Indianapolis, USA.

**Daniel Spiers** is a Program Manager (Peer-Exchanges and Knowledge) at Philanthropy Europe Association (Philea). With a background in European Studies, he focuses mainly on organizational development and data science in the context of emerging practices in foundations.

**Volker Then** is the CEO and a member of the Executive Board of Fondazione AIS, Bologna. He has published extensively on impact measurement and holds positions on several committees in the nonprofit and philanthropic sectors.

**Laura Tocmacov** is co-founder and Director of the impactIA Foundation, an institution that works to create organizations with hybrid intelligence optimized for collective well-being. She has been working in the field of transition and employability for over 20 years.

**Maria Cristiana Tudor** is a Researcher at the Geneva Finance Research Institute, University of Geneva (UniGe), and a PhD candidate at the Lemanic Neuroscience Doctoral School. She holds a master’s in Neuroscience from UniGe and a Bachelor of Arts and Sciences (BASc) from University College London, with a major in Mathematics-Physics and a minor in Management.

**Giuseppe Ugazio** is an Associate Professor in Behavioral Philanthropy and Finance at the Geneva Finance Research Institute, Geneva School of Economics and Management of the University of Geneva. He holds a PhD in Neuroeconomics and a PhD in Philosophy.

### *Contributors*

**Peter Vogel** is a Professor of Family Business and Entrepreneurship, holder of the Debiopharm Chair for Family Philanthropy, and the Director of IMD Global Family Business Center. Vogel is also the Director of the IMD Global Family Business Award.

**Joe Wheeler** has worked in philanthropy since 2011. He has launched programs for Dropbox, UNDP, and WhatsApp. A graduate of Whitman College and the London School of Economics, he is currently working on a PhD in Nonprofit, Philanthropic, and Social Enterprise at the University of Oregon.

# FOREWORD

## From AI-empowered philanthropy to philanthropy-driven AI

Artificial intelligence (AI) and philanthropy may seem like odd companions, yet this *Handbook* shows the relevant and profound connection between the two fields, providing an indispensable guide for those at the forefront of this dynamic intersection.

The swift advancement of AI has reshaped countless sectors, including philanthropy. This is one side of the coin, analyzed in detail in the following chapters: what AI can do for philanthropy. As charities seek to amplify their impact and address multidimensional societal issues, AI offers new approaches for innovation, efficiency, and effectiveness in allocating resources that are constantly and inevitably too limited. The *Handbook* investigates how AI can upgrade organizational strategies for problem-solving, resource distribution, and impact assessment, and examines its applications, consequences, and possibilities within the philanthropic domain. AI appears to be a significant agent of change, potentially invaluable in a sector occasionally prone to excessive caution or conventional approaches. In essence, the transformative power of AI might embolden philanthropy to aspire to greater and improved outcomes.

The other side of the coin is what philanthropy can do for AI. Here, the innovation may seem less conspicuous, but it is equally and, one may argue, perhaps even more significant. Philanthropy can play a crucial role in AI's development and ethical integration by providing the necessary support for AI research, innovation, development, and deployment while enabling a diverse field of practitioners. The philanthropic sector can fund unique initiatives beyond the scope or remit of business and state organizations and drive technological advancements in AI that benefit society and the environment. For example, philanthropic organizations can invest in tailored educational programs that equip future generations with the skills to excel in an AI-dominated landscape, emphasizing technical expertise, ethical considerations, sustainability, and societal impact. Furthermore, philanthropy can bridge gaps by supporting AI applications in public interest areas that may be overlooked by commercial and public actors, such as in healthcare, environmental conservation, and the cultural sector. Additionally, philanthropic efforts can support the development and application of ethical guidelines and legal frameworks that govern AI development. Funding research into AI's social implications, including bias, fairness, and transparency, can lead to more responsible and accountable design and use of AI systems. By focusing on these areas, philanthropy can ensure that AI advancements work toward the greater good.



Looking at both sides together, there is a crucial, cooperative spirit that may benefit from the interplay between AI and philanthropy. AI-empowered philanthropy and philanthropy-driven AI have the extraordinary opportunity to catalyze cross-sector collaboration, uniting academia, industry, and civil society to ensure that AI is aligned with human values, societal necessities, and environmental imperatives. Arguably, this binding function is among the most pressing requirements of our era, and philanthropy is well-placed to successfully fulfill this unique role by leveraging AI. By grasping AI's opportunities and managing its challenges, we can collectively strive to use its potential responsibly and inclusively for the greater societal benefit.

The *Handbook* offers a wealth of perspectives, thanks to contributions from distinguished experts across disciplines, including scholars, researchers, and philanthropic representatives. This fosters a comprehensive understanding of AI in philanthropy, drawing on fields such as computer science, social sciences, and law, to provide a multidisciplinary examination of the topic. In addition, the authors do not limit themselves to analyzing AI's current state within philanthropy but also look ahead to future trends and potential shifts. They consider scenarios where AI could redefine philanthropic strategies, including personalized giving through AI algorithms and augmented cooperation and knowledge sharing among philanthropic entities. Acknowledging the global dimension of AI and philanthropy, the authors offer perspectives and case studies from various regions, understanding that AI must address worldwide issues such as poverty, healthcare, and sustainability, while respecting diverse cultural and political contexts.

This volume starts by offering a more historical account of the relationship between AI and philanthropy. It traces the origins of both fields, highlighting their early interactions and the development of AI tools for philanthropic endeavors. The overview acknowledges the potential of AI to drive positive social change, while also considering the ethical quandaries and obstacles that accompany its use in philanthropy. The *Handbook* then explores multiple aspects of AI in philanthropy, evaluating its applications, effects, and broader implications. It examines AI's capacity to drive substantial change, from automating routines and improving resource distribution to refining impact measurement or enabling data-driven decision-making. The chapters underscore AI's potential to revolutionize philanthropic practices, making them more productive, impactful, and responsive to community and beneficiary needs. Additionally, the *Handbook* confronts the difficulties and risks of adopting AI in philanthropy. The authors analyze potential pitfalls, such as ingrained biases in AI systems, data privacy concerns, cybersecurity risks, and the necessity for responsible AI development and application. They emphasize the need to address these challenges with comprehensive governance structures, ethical principles and guidelines, legal frameworks, and inclusive approaches that uphold transparency, accountability, and stakeholder participation. Case studies that illustrate practical applications and outcomes of AI in philanthropy are included to demonstrate its tangible benefits. These examples offer insights into the use of AI in various philanthropic contexts, showcasing its practicality in predictive grant-making models, beneficiary engagement through AI-enhanced communication systems, and more. Finally, the *Handbook* focuses on the confluence of AI, philanthropy, and ethics, stressing the importance of responsible and inclusive AI integration. It tackles ethical concerns like algorithmic bias, openness, and accountability, and advocates for ethical frameworks and guidelines that align with philanthropic principles and prioritize community welfare.

As we consider AI's current and prospective roles in philanthropy, the *Handbook* is a thought-provoking resource that encourages debate and informed decision-making among all parties involved. It underscores the importance of capacity building and skills development and encourages philanthropic bodies to invest in digital literacy, data analytics, and AI expertise to leverage these technologies for the public good. The book is set to provide an invaluable resource

*Foreword*

for researchers, practitioners, policymakers, and philanthropic organizations dealing with the complex issues of AI in philanthropy. Rigorous academic analysis and actionable insights enrich our understanding of AI's potential, challenges, and ethical issues in philanthropic work. It is a comprehensive and stimulating guide, perfectly timed to advance the conversation on AI in philanthropy, spurring further research and cooperative innovation.

We stand at the threshold of a promising era where the synergies between AI and philanthropy present innovative solutions to some of humanity's most pressing problems. Realizing the full spectrum of possibilities AI offers to philanthropy hinges on the collective endeavors of researchers, practitioners, and organizations, equipped with insights gleaned from resources such as this *Handbook*. Thanks to similar efforts, we can steer AI's application in philanthropy toward shaping a more equitable, sustainable, and prosperous world for all.

**Luciano Floridi**

Yale Digital Ethics Center

Yale University

New Haven, CT

and

Department of Legal Studies

University of Bologna

Bologna, Italy



# Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# EDITORS' INTRODUCTION TO THE VOLUME

*Giuseppe Ugazio and Milos Maricic*

Artificial intelligence (AI) and philanthropy are two ecosystems that have spontaneously interacted for decades. This interaction, however, has been steered by a select group of influential philanthropic actors, mainly linked to the digital and tech sectors. Given the technical complexities and the fast pace of AI advancements, most non-involved philanthropic actors find it extremely difficult to start interacting with AI. To remove some of the barriers impeding a fruitful interaction between AI and philanthropy, this *Handbook* proposes a first-of-its-kind, multidisciplinary, academic contribution, aiming to facilitate philanthropic actors' involvement in AI implementation, development, and use, while also broadening awareness and knowledge of the potential positive impact of AI on the philanthropic sector. In this *Handbook*, AI is broadly defined as the branch of computer science dedicated to developing tools to enable machines to perform tasks that would otherwise require human intelligence or skill. With the development of quick and efficient ways to share information (e.g., the internet) and the exponential growth in computational resources, the past decades have seen AI becoming increasingly ubiquitous in our societies, influencing our social, professional, and personal lives (Sheikh, 2021). While previous technological revolutions have, arguably, had similar impacts on human societies and the environment, what is unprecedented in the case of AI is not only the pace at which the change is happening but also the potential to replace or even outperform human skill in ways never seen before (Gill, 2017). Very few sectors, if any, have not been impacted by the rapid development of AI technologies. Appropriate incentives, strategies, and tools to ensure that this technology is developed, regulated, and used responsibly to benefit society are urgently needed. At present, while it is undeniable that AI is contributing to fostering humans' well-being, removing barriers to collaboration, and augmenting the efficiency of our work, the forces that have shaped AI's development and use have also exacerbated current profound societal issues such as bias perseverance, widening economic inequalities, and increased exclusion of the most vulnerable. As we reach the cusp between one of the latest AI hype waves and approach what some call an AI winter (Floridi, 2020), it is now an ideal time to empower the increased presence of an actor capable of steering AI's future development and use in alignment with the wider interests of humanity: philanthropy.

The *Handbook* broadly defines philanthropy as the private giving of time or resources (money, security, property) for public purposes (Salamon, 2014) – the latter being codified by the United

Nations 2030 Sustainable Development Goals (U.N. SDGs)<sup>1</sup> – a definition proposed by the Geneva Center for Philanthropy of the University of Geneva.<sup>2</sup> Philanthropic organizations (POs) are in the unique position to advocate for and contribute to AI development that is ethical, inclusive, and responsible, as they hold as their core goal making humanity thrive while equally understanding private companies' logic. To do so, however, it is crucial that the philanthropic sector builds a solid knowledge of what AI is, what it can bring to society, and where the limits of this technology should be set to minimize its risks. To build capacity on AI's potential, as well as stimulate the ethical debate on its righteous use, in the present book, we identified two broad key areas of research addressed by multidisciplinary experts: (1) *unveiling the tools and adoptions of AI with the most potential to assist philanthropy's strategies and operations*; (2) *identifying the practical and ethical principles that should be kept in mind when developing and using AI*. By addressing these issues, we aim to provide academics, developers, practitioners, and researchers with the knowledge and expertise needed to leverage their respective unique roles, thereby successfully facilitating solutions-oriented dialogue, aiming to shape our digital (and AI) futures in a human-centric manner.

This *Handbook* is divided into four parts. The first, titled ***AI applications in the philanthropic sector***, explores how AI can be used to aid POs in becoming more efficient and increase the impact of their actions, which otherwise, without AI-based technologies, would be challenging, more limited, or inefficient. The second part, ***Philanthropies' regional AI adoption, readiness, and applications***, offers perspectives written by sector professionals and academics, illustrating the current status quo of AI readiness and adoption in the nonprofit sector across several world regions, identifying key barriers and recommendations for improvement, while equally presenting several applied case studies. Part three, ***Philanthropy for AI development and regulation***, focuses on the role of philanthropy in AI development, regulation, and policy-making. Finally, the fourth part, titled ***Ethics, AI, and philanthropy***, explores some of the ethical principles that POs should advocate for, to guide better AI implementation and development, in the quest for Ethical and Inclusive AI (EIAI).

## Part I: AI applications in the philanthropic sector

The *Handbook* opens with a comprehensive overview by **Camilla Della Giovampaola and Maria Cristiana Tudor** on the historical relationship between AI and philanthropy, from their earliest interaction in the mid-20th century up to the most relevant contemporary movements, such as *AI for Social Good*, alongside key stakeholders and concrete examples of how AI can be used in this domain. The authors offer insights into how the relationship between technology and philanthropy is neither novel nor static, with these two fields' interaction evolving across time and resulting in new forms of tech-driven philanthropy. The following chapter, by **Simone Bartalucci, Antonia Muhr, Sina Sauer, and Volker Then**, describes how AI can be used to accomplish one of the most demanding tasks in philanthropy: *measuring impact*. In this chapter, the authors argue that AI can harmonize impact measuring by focusing on several key steps with transferable and universal applications. Their contribution proposes a solution to the complex matter of establishing a unique framework to *measure activity impact in social and environmental contexts*, developed by a group of partners and philanthropic stakeholders. In a similar spirit, AI used to design a unified framework for monitoring philanthropic activities is the central theme of **Nelson Amaya, Harry De los Rios, and Madeleine Lessard**'s contribution. Their chapter showcases how *AI can monitor and compare the financial flows of philanthropic capital* used to support international development. The authors introduce *PHIL4DEV*, a machine learning tool used to classify philanthropic financing data collected and used by the OECD from POs within their network. Next, **Stefan Schöbi**

discusses the role of *large language models (LLMs)* and natural language processing (NLP) to aid philanthropic operations. LLMs are particularly strong where wording plays a significant role in a given organization's success, for example, in *drafting reports or grant applications*. It further remarks the necessity for tailored educational programs to fast-track the adoption of such technologies by POs. Through these programs, philanthropy can become more efficient and effective in handling time-consuming processes, such as application and disbursement. **Prity Khastgir and Shweta Shalini** then present more examples of *AI tools to effectively support philanthropic internal and external operations*. The authors first showcase diverse AI tools supporting POs' operational activities in *fundraising, outreach, and engagement*, then move to improved Human Resources (HR) methods and staff upskilling to enhance organizational growth and innovation. Next, they look at the role of *blockchain and Decentralized Autonomous Organizations (DAOs) for improved governance* and crowdsourced funding, before concluding with the convergent role of AI and philanthropy in advancing the Sustainable Development Goals (SDGs), enhancing social impact with transparency and innovation.

**Anita Budziszewska and Oto Potluka** then take an international relations perspective to analyze the implications of leveraging AI and other modern technologies to support the development of civil society organizations (CSOs). From this vantage point, the authors discuss how AI and other modern technologies can positively contribute to these organizations' improvement, noting that even in highly developed countries, most CSOs are not sufficiently leveraging the potential of this technology. The authors recommend implementing *digital and technological policies* and regulations to support AI adoption by CSOs, noting that a major concern needing to be addressed is the global digital divide, i.e., the existing disparities in access to AI, data, and other digital technologies seen between countries with different levels of digitalization. The in-depth analysis of AI's role in *fundraising strategies* for different types of POs continues with **Luca Barzanti, Lia Benvenuti, and Enrico Gaudenzi**'s contribution. They present how AI and Information Technology (IT) can be used to develop decision support systems (DSS), a mathematically sound tool for simulating and designing more effective fundraising campaigns. Similarly, **Marc Schipper, Manouchehr Shamsrizi, and Adalbert Pakura** discuss the role of AI-enhanced virtual environments (or metaverses) that POs can leverage to offer individuals immersive, meaningful, and interactive experiences. For example, AI can generate realistic characters and environments, allowing participants to interact with the real environments of philanthropic causes, thereby stimulating empathy and connection to the cause. Such tools stand to enhance *fundraising, volunteering, and recruiting*. **Marta Herrero and Shauna Concannon** follow a similar theme, detailing how AI can surpass traditional digital methods for fundraising by deepening understanding of donors and personalizing communications. It introduces a resilience framework for nonprofits to effectively adopt AI, focusing on *AI-driven storytelling to foster donor loyalty and long-term financial support*.

## Part II: Philanthropies' regional AI adoption, readiness, and applications

At this stage of this book, the reader will have acquired a strong sense of how AI can be used to support philanthropic decision-making and complement their operations. The next chapters offer more pragmatic perspectives on the current status quo of AI usage in the nonprofit sector by showcasing how different global regions engage with and adopt AI, to then present several applied case studies.

We start with **Patricia Snell Herzog**'s contribution, which analyzes 349 POs to identify three different styles used by them to integrate technology: the first encapsulates organizations whose missions are *tech-centered* – develop said tech; the second are *technology-perpetuating* – promote

access to said tech; while the last are *technology-implementing* organizations – using said tech for social good initiatives. The nuances of these different styles are robustly supported by 14 case studies of POs belonging to each category. In the next chapter, **Sevda Killicalp, Jack O’Neill, and Daniel Spiers** provide an overview of *European POs’ engagement with data science and AI*. By contrasting the speed and tendency of POs, compared to for-profit entities, the authors note that the former faces challenges not only related to organizational resources, skills, or technology infrastructure, as the latter do as well, but also to ethical, environmental, and reputational considerations. These considerations are suggested to be the main reasons why philanthropic organizations are reluctant to embrace data science and AI; however, they need to be urgently addressed to reverse this delayed-adoption tendency. This trend reversal would ensure that POs can play an active role in promoting ethical, inclusive, and responsible AI. In central Europe, **Aline Kratz and Hubert Halopé** examine the *opportunities and risks of digital transformation in Swiss foundations* by drawing on insights from the private sector. The authors assess AI risks in the context of charitable foundations, considering measures to safeguard against them by leveraging industry best practices.

Away from the European continent, the AI adoption lag seen here mirrors trends of the Asian philanthropic ecosystem, as analyzed by **Kithmina Hewage** in the chapter that follows. This contribution discusses three key *insights on the Asian philanthropic sector* based on existing evidence: first, the foundational readiness of the Asian social sector is severely lacking due to infrastructural and financial constraints; second, a lack of skill and expertise is driving many of the impediments to improving the sector’s operational readiness; and third, the social sector in most Asian economies is vulnerable to adopting AI tools without adequate precautions, opening itself up to significant exploitation and fraud risks. In a similar geographic region, **Bertram Lang** describes how *AI found fertile ground within Chinese philanthropic organizations*, considering these have found themselves in the midst of a digital transformation, radically enhancing their reliance on internet-based approaches. The chapter stresses how as funding from foreign donors to China has plummeted, small-scale online donations channeled through the social media platforms of China’s largest internet and communication technology (ICT) firms have turned into a critical source of revenue for many grassroots NGOs – as exemplified by the Ten-Cent charity case. From a broader perspective, the chapter concludes by discussing that Chinese AI-enabled philanthropy has global implications, inviting further reflections about the growing role of digital platform power – and crucially, its relationship with state power – in the nonprofit sector.

Moving on from geographical analysis of AI adoption in different philanthropic ecosystems and influenced by diverse political climates, we turn our attention to applied case studies. **Nikola Milinkovic and Marko Gajak** present a case study on Catalyst Balkans – a *Western Balkan organization focused on using AI and LLMs to collect philanthropic data* and create the Giving Balkans database. This serves as a first-of-its-kind database to track and monitor philanthropic activities across the many multicultural and multilingual Balkan countries. Despite linguistic variance challenges, LLMs seem promising tools for automatically analyzing large multilingual sources and classifying philanthropic activities. The organization aims to minimize human roles in repetitive data collection and categorization while equally providing greater efficiency and accuracy in data handling. Nevertheless, challenges persist, such as false positives and data veracity concerns, which continue to require human surveillance. Another case of the transformative power of AI for philanthropic strategies and operations is illustrated in **Milos Maricic’s** case study on *Altruist League* – a Geneva-based philanthropic advisory firm. He discusses AI’s application in *matching donors who focus on systemic change with grassroots organizations and movements around the world*. This case study revealed that Altruist League’s approach effectively developed

an AI system for donor-partner matching, achieving over 90% accuracy. This performance was recently complemented by the development of an LLM trained on its data, allowing automated summarization and donor advice generation. The study concludes by offering an overview of the League's technological tools to operate at the intersection of AI, philanthropy, and systemic change. The potential of AI for boosting efficiency is of particular appeal to smaller organizations, such as family philanthropic organizations – precisely the object of analysis of the contribution by **Malgorzata Smulowitz and Peter Vogel**. Focusing on the Spandows family's business and philanthropic endeavors, this chapter offers several *example use cases alongside key recommendations on how to effectively blend AI and business acumen to drive social and environmental impact*, both in a for-profit and nonprofit approach. **Monica Lea and Lucia Gomez's** chapter offers the last case study of Part 2 of this *Handbook*. It examines the key success ingredients, as well as the ethical implications of an emerging form of philanthropy: *Digital Stunt Philanthropy (DSP)*. They do so by analyzing Mr Beast's YouTube channels and their significant philanthropic fundraising activity, which is done through engaging and entertaining YouTube videos. The authors note several pros of DSP while proposing that DSP offers innovative avenues for engaging private philanthropic partnerships with corporate sponsors and nonprofit organizations. Insights derived from this chapter can inform nonprofit organizations and content creators about the benefits and challenges of merging entertainment with charitable activities in the digital space.

### **Part III: Philanthropy for AI development and regulation**

The *Handbook* now transitions to its third main area of research – analyzing philanthropy's role in guiding, informing, and shaping AI developments – with discussions on advocacy efforts, regulation and policy-making, and governance recommendations. This part begins by discussing the role of philanthropy in developing AI, focusing on another emerging philanthropic trend – *Data Philanthropy (DP)*. The author, **Rahul Jha**, explores the emerging field of DP while addressing three primary aspects of AI application from a PO's standpoint on this emerging trend. The discussion addresses challenges such as the “data invisible” issue, highlighting the importance of Data Philanthropy (DP) in bridging the global digital divide and enhancing the representation of marginalized communities in AI systems. It includes a case study on an international organization's role in DP illustrating the practical applications and challenges of DP and open data. **Vilas Dhar and Yolanda Botti Lodovico** then observe the global trend of power transformation across all sectors, fueled by AI, except for the social sector which remains left behind. In this context, civil society and the communities they serve require both a partner and champion to bridge technology with social progress: philanthropic organizations. With AI support, these have the potential to strengthen their mission, boost operations' impact, and, therefore, better address global challenges. This chapter provides a detailed account of *how philanthropy can drive digital transformation across the social sector* and presents recommendations for organizations to harness AI toward building an ethical future. The transformative power of collective collaboration is further explored by **Thomas Maillart, Lucia Gomez, Mohanty Sharada, Dipam Chakraborty, and Sneha Nanavati**. The authors explore how altruistic collective intelligence (CI) is advancing AI technologies. Focusing on empirical evidence from *AIcrowd, a platform that leverages community-based development, the study illustrates how a risky “trial-and-fail” strategy can drive AI innovation through peer production*. The chapter provides a concrete case for philanthropic organizations interested in experimenting with innovative methods, for example, leveraging altruistic CI to reshape the future of AI, making it more inclusive, innovative, and ethically grounded. **Joe Wheeler's** essay explores different philanthropic giving models, ranging from



straightforward cash grants to more complex contributions of assets like technology or knowledge transfer, with a focus on the strategic donation of AI capabilities to NGOs. It argues that *in-kind donations, such as machine learning training and specialized software, are often more beneficial than cash*, particularly when such AI resources meet specific needs that NGOs cannot easily fulfill on their own and are managed effectively by donors without compromising their core business objectives. This chapter draws on management literature to identify key conditions in which asset donation is more valuable than cash.

Philanthropy's role in the development of AI systems is equally complemented by its role in supporting policy discussions for democratic and ethical regulations for AI. This regulation-centered section begins with **David Harris and Anamitra Deb's** chapter, who argue that *AI's rapid developments have vastly outpaced the ability of regulators in most of the world to implement rules that govern its use*. Without AI, its benefits are likely to flow to the few, while the many risks it poses and harms it has already wrought will be borne by society, and disproportionately so to already vulnerable communities. The authors thus call for the philanthropic sector to lead advocacy efforts for creating robust legal frameworks – to regulate AI's usage and development – and, while these take effect, provide a set of ethical principles and best practices ensuring that AI is not misused. This call for regulation is followed by **Diana Acosta Navas'** interrogation of *how democratic principles are stressed in different ways by both philanthropic foundations and foundation AI models*. Through this analogy, the author shows how AI and POs operate outside democratic institutions, with substantial societal impact and minimal accountability, leading to considerable power concentration that is unresponsive to the individuals whose lives are impacted by them. It then proposes to leverage the Deliberative Alignment method developed within AI also to democratize decision-making regarding foundation models. In line with the EU-AI Act<sup>3</sup> risk-based regulatory approach, **Simeon Schiff and Daniel Campos** emphasize *philanthropy's role in mitigating extreme risks linked to AI*. In particular, the chapter provides a detailed overview of the *AI extreme risk mitigation philanthropic sector* (AIERMPS). This segment emerged in the early 2000s to tackle issues related to artificial general intelligence and existential risk. The authors provide a historical review and landscape analysis, including a description of the ideologies and culture of the sector, noting how from its origin to now, the core approaches, stakeholders, and culture have all substantially evolved.

#### Part IV: Ethics, AI, and philanthropy

Departing from development and regulation, the last part of this *Handbook* deals with ethics discussing some of the ethical principles that can and should inform AI development and usage, in particular by and for philanthropy. It begins with the contribution of **Laura Tocmacov** – Founder of ImpactIA, a nonprofit organization dedicated to advancing the *Montreal Declaration's principles for ethical and inclusive AI development*. This chapter outlines ten principles, substantiating the discussion with concrete examples of ImpactIA's projects for advancing them. In the chapter that follows, **Rhodri Davies** dives into the understudied impact of AI on individual philanthropic decisions, *exploring how AI tools could potentially influence when, where, and how people choose to give*. It examines the dual nature of philanthropy, both as a systemic mechanism that reallocates resources across society and as a reflection of individual voluntary actions for public good, highlighting the profound implications of AI on both levels. Additionally, the chapter assesses the roles of various platforms and organizations in using AI to shape donor decisions, *raising important ethical questions and suggesting necessary policy or practice changes* to address them. Next, **Ursula Jasper, Siddhartha Jha, and Stefan Germann** offer insights on how a foundation develops a strategy to address one crucial question for this *Handbook*: *can*

societies adjust and keep pace with the speed of AI? The authors argue that *philanthropies have a significant contribution to the complex ethical, economic, political, and legal questions posed by the AI revolution* while attempting to guarantee core values such as *fairness, non-discrimination and non-stigmatization, benefit-sharing, participation, privacy protection, safety, informational self-determination, and autonomy*. They concretely identify five roles for philanthropies to play in shaping the ethical and inclusive AI revolution: from funding and sponsoring research to catalyzing equitable innovation through public good and facilitating a broad and transparent public dialogue on AI and digital futures. **Elizabeth Searing and Donald Searing** address the fundamental question of *how and what to teach AI such that it learns to process information and produce outputs (i.e., “think”) in an ethical and philanthropic way*. This essay proposes a learning journey to train ethical and philanthropic AI by first looking at how humans have been taught, successfully or unsuccessfully, to behave in such a way, in particular, analyzing the behaviors, principles, and tendencies characterizing the philanthropic ecosystem. As a result of this comparative analysis, the authors then provide practical advice on how to train a philanthropic AI. This advice is then complemented by a more normative *philosophical discourse* by **Ezekiel Kwetchi Takam**. Here, he discusses that *inclusion and non-proprietary/open-source approaches should be championed by philanthropy* as it strives to direct the development of AI and even more sophisticated forms of it, such as *Artificial General Intelligence (AGI)*. As a result of endorsing and advocating for these principles, philanthropy can empower marginalized groups that might otherwise be excluded from the benefits offered by AI/AGI. Part 4 ends with **Charles Sellen and Joost Mönks**, who list several ethical perspectives that philanthropy could adopt in their quest to use and develop AI, for example, when using AI technologies to increase performance, achieve the Sustainable Development Goals, or set out to design Ethical and Inclusive AI (EIAI) frameworks.

We have completed the journey through all four *Handbook* parts, covering a wide range of topics, from AI tools to enhance the operations and impact of the philanthropic sector to philanthropy's role in advocating for and developing ethical and inclusive AI. By combining academic knowledge, practical and theoretical perspectives, and numerous insights from evidence-rich use cases, the editors believe that this *Handbook* will serve as a reference to any philanthropist, professional, or scholar, who seeks to understand and leverage AI's capabilities while ensuring that it is used in an ethical, inclusive, and responsible way. Looking forward, this *Handbook* concludes with a thorough analysis by **Ravit Dotan** on *how philanthropy and AI can elevate each other in a responsible way* and how philanthropy can shape AI in different roles: as grantmakers, users, developers, buyers, investors, and social justice advocates. This analysis relies partly on ideas and perspectives debated among participants and panelists of the *first Artificial Intelligence and Philanthropy academic conference*<sup>4</sup> organized by the Geneva Center for Philanthropy, **Henry Peter, Mara de Monte, and Margaux Casagrande**, and the research team on AI and Philanthropy of the University of Geneva: **Camilla Della Giovampaola, Hubert Halopé, Lucia Gomez, Nisa Thomas, Maria Cristiana Tudor, and Giuseppe Ugazio**. A summary of the themes, original ideas, and next steps that emerged from this conference is provided in the very last chapter of this book.

## Notes

1 <https://sdgs.un.org/goals>

2 <https://www.unige.ch/philanthropie/en>

3 <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

4 <https://www.unige.ch/artificial-intelligence-philanthropy/international-conference>

### References

- Floridi, L. (2020). AI and its new winter: From myths to realities. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3830584>
- Gill, K. S. (2017). Uncommon voices of AI. *AI & SOCIETY*, 32(4), 475–482. <https://doi.org/10.1007/s00146-017-0755-y>
- Salamon, L. M. (Ed.) (2014). *New frontiers of philanthropy: A guide to the new tools and actors reshaping global philanthropy and social investing*. Oxford University Press: New York.
- Sheikh, S. (2021). *Understanding the role of artificial intelligence and its future social impact*. IGI Global. <https://doi.org/10.4018/978-1-7998-4607-9>

## **PART I**

# AI applications in the philanthropic sector



# Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# 1

## THE RELATIONSHIP BETWEEN AI AND PHILANTHROPY

From historical roots to modern convergence

*Camilla Della Giovampaola and Maria Cristiana Tudor*

### 1 Introduction

Recent advancements in artificial intelligence (AI) technology have prompted sectors to critically assess how they can best adapt to an increasingly AI-operated world. The philanthropic sector is no exception, with philanthropists, practitioners, and academics questioning not only the implications of AI on the future of philanthropy but also the responsibility philanthropic organizations (POs) hold – as promoters of social good – in supporting the development of ethical and inclusive AI (EIAI) systems (Bernholz et al., 2010; Arrillaga-Andreessen, 2015; Chu & Wang, 2019; Madianou, 2021).

Philanthropies' two-way relationship with AI, as both users and developers, is not new but rather one of the latest expressions of the interlocking of the two fields (Henriksen & Richey, 2022). However, POs' engagement with AI varies greatly across the sector. On the one hand, philanthropies linked to tech giants are leading the integration of AI into philanthropy and largely shaping the AI for Social Good (AI4SG) movement, an example being the Schmidt Futures. On the other hand, the majority of more traditional POs are largely lagging behind in their adoption of AI-powered tools and overall digitalization (Google, 2019; Herzog et al., 2021). This raises questions about the role of philanthropy in an increasingly digitized civil society (Taddeo, 2016; Bernholz & Reich, 2017; Bernholz, 2021).

While most attention is currently focused on the organizational, legal, and ethical structures that need to be in place for a proper integration and development of AI within the philanthropic sector (Taddeo, 2016; Floridi et al., 2020; Kanter & Fine, 2020; Herzog et al., 2021), in this chapter we take a step back: we contextualize philanthropy's relationship with technology over time and highlight the forces driving modern applications of AI within the philanthropic sector. This exercise aims to provide a more informed approach to discussions on the future trajectory of AI and philanthropy. In more detail, with the goal of providing an overview narrative of the intersection of technology and philanthropy, this chapter adopts a historical lens and provides a trajectory over time from the mid-1950s to today. First, we outline philanthropy's early role in supporting AI research and development, and the subsequent rise of tech philanthropy and the AI for Social Good (AI4SG) movement. Second, we explore the opportunities and challenges that derive from philanthropies' adoption of technological tools, the latest of which are AI-powered tools. Finally, we dive into the AI4SG movement, mapping key stakeholders and prominent AI4SG initiatives and trends.

## **2 The evolution of philanthropic support in AI research**

### **2.1 Funding the future – philanthropy’s early role in AI**

The first recorded use of the term “Artificial Intelligence” (AI) is found in a 1956 grant application submitted by mathematician John McCarthy to the Rockefeller Foundation (Manning, 2020). Striding into uncharted territory, McCarthy needed to find a term to describe a new concept in computer science, where machines could perform tasks typically requiring human intelligence (Manning, 2020; Shubinski, 2022). McCarthy was seeking financial support from the Rockefeller Foundation to hold a two-month conference titled the “Dartmouth Summer Research Project on Artificial Intelligence.” Granting the mathematician only half the amount he requested, namely \$7,500, the conference took place in 1956 at Dartmouth College and is now widely recognized as the birthplace of modern AI (Rockefeller Philanthropy Advisors, 2019; Shubinski, 2022). Although this new field of computer science was understood by only a handful of researchers at this time, its significance was clear. Bell Laboratories, the International Business Machines Corporation (IBM), and RAND all supported the conference to enable their key researchers to attend (Shubinski, 2022).

Philanthropy’s early endorsement of the development of AI technologies at Dartmouth College in 1956 set the ground for the sector’s ongoing commitment, with contributions to AI, Machine Learning, and Data Science technology (AIMS) philanthropy reaching \$2.6 billion in 2021 (Herzog et al., 2021). However, the use of philanthropic resources to support the research and development of unexplored technological fields was not a new phenomenon (Zinsmeister, 2016). On the contrary, the two fields have been experiencing a dynamic relationship for centuries; on the one hand, philanthropy contributes to technological progress, and on the other hand, technology allows philanthropy to achieve greater results and operate more effectively (as discussed in Section 3). The intersection of AI and philanthropy is one of the many manifestations of the intertwining of the fields of technology and philanthropy (Henriksen & Richey, 2022).

Philanthropy’s long-standing commitment in supporting pioneering technological research and development is well documented in the literature (Bernholz et al., 2010; Michelson, 2020). From the Renaissance, where wealthy patrons supported inventors and scientists, to the Industrial Revolution, during which philanthropists played a pivotal role in the development of transportation, communication, and infrastructure, and up to the philanthropists of the 21st century who are spearheading disease eradication and technological revolutions, examples are plentiful (Zinsmeister, 2016; Michelson, 2020). Nonetheless, philanthropists’ belief in the need to tackle societal problems through scientific research and technology took an important turn at the end of the 19th century. Led by the work of the Rockefeller philanthropies, the rise of “science philanthropy” was a direct response to what the American magnate philanthropists of the time saw as the inability of charities to address the root causes responsible for creating and perpetuating human suffering (Bremner, 1994; Sealander, 2003; Bishop & Green, 2008).

Science philanthropy commonly refers to the giving of charitable funds for scientific or technological research (Falk & Michelson, 2021). Although this philanthropic field of action has evolved and adapted over time, it retains one crucial feature, namely, its high-risk tolerance (Bennett et al., 2016; Falk & Michelson, 2021). Unlike government agencies, which are often bound to tight budgets and lengthy bureaucratic practices, POs enjoy greater operational freedom. Fields of research that are difficult to fund with taxpayers’ money can be spearheaded by philanthropists who, as private individuals, are far more flexible and agile with their resources. A recent example is philanthropies’ fast response to the Covid-19 pandemic, with POs committing more than \$10 billion

globally in just over six months from the start of the pandemic in late 2019. Led by American and Chinese philanthropies, whose giving amounted to, respectively, more than USD 6 billion and USD 1 billion, POs' contribution by May 2020 represented 38% of all the total Covid-19 relief funds (Church, 2020; Council on Foundations, 2020; Watson, 2022). Moreover, as many POs are set up to help solve long-standing societal problems, they can invest in technologies that do not provide immediate or guaranteed results but that are expected to have a positive effect on society in the long term. This characteristic also distinguishes POs from the for-profit sector's short-term return approach.

Despite POs' long-standing role as funders of technology, it remains difficult to measure the extent to which POs fund science and technological innovations and the full impact of these investments on society. While education, health, and economic and community development remain the largest recipients of national and transnational philanthropic funds, the field of technology and science is rarely mentioned on its own in contemporary reports on philanthropic giving trends (Johnson, 2018; Rockefeller Philanthropy Advisors, 2022; Indiana University Lilly Family School of Philanthropy, 2023).<sup>1,2</sup> Quite exceptionally, the Rockefeller Philanthropy Advisors' report "Global Trends and Strategic Time Horizons in Philanthropy 2022" lists science and technology as the tenth focus area for that year. This appears to indicate that, although they exist as an individual focus area, technology and science are primarily funded by POs as a means to address global challenges such as education and health rather than for the sake of developing innovations in the field. This would align with the role of POs as promoters of social good rather than profit-seeking institutions.

## ***2.2 The rise of tech philanthropists in the 21st century***

Tech leaders began to play an increasingly prominent role in the field of philanthropy in the late 1990s and early 2000s following the dot-com boom. As the internet generated enormous wealth and boosted the software and computer industry, tech companies and their founders decided to devote part of this wealth to philanthropy (Bishop & Green, 2008; The Economist, 2023). New foundations and funds were established, pioneered by the Bill and Melinda Gates Foundation in 2000, with the ambitious missions to address some of the world's most challenging issues (Bennett et al., 2016). Tech moguls brought with them their own ideas on charitable giving, distancing themselves from the grandfathers of modern philanthropy like American industrialists Henry Ford, John D. Rockefeller, and Andrew Carnegie. These industrialists had created and operated foundations that were designed to outlive them, employing numerous advisors to provide funds over many years to achieve a goal. Tech philanthropists, instead, wanted to operate differently, prioritizing data, speed, and impact (Bishop & Green, 2008; The Economist, 2023). By framing their donations as an investment in humanity rather than charitable giving, tech founders bring not only their vast resources to the table but also their own culture and methodology (Bishop & Green, 2008; Bennett et al., 2016; Torres & Zinsmeister, 2018).

Today, tech actors are leading philanthropic players, contributing billions of dollars annually to philanthropic causes around the world (Torres & Zinsmeister, 2018). According to the Chronicle of Philanthropy, of the \$33.4 billion given away by America's top 50 donors in 2021, about three-quarters of those donations came from people who have made their fortunes in the tech industry (The Economist, 2023). In India, the consultancy Bain & Company reports that tech titans hold about 8% of the total wealth of the country's super-rich yet their donations account for 35% of charitable giving (The Economist, 2023). In 2022, tech magnate Bill Gates made what is considered the biggest philanthropic contribution of the year with a \$5 billion donation to the Bill and Melinda Gates Foundation (Di Mento, 2022).



In addition to their large and growing financial contributions, tech moguls' involvement in philanthropy continues to evolve as they craft new ways of practicing philanthropy. One such example is what is commonly referred to as "trust-based philanthropy." Popular among philanthropists of the caliber of Mackenzie Scott and Jack Dorsey, it involves moving away from philanthropy's traditional "hands-on," reporting-centered approach and giving trustees the freedom to decide how best to use the money (Kulish, 2021; The Economist, 2023). This frees grantees from time-consuming bureaucratic activities such as reporting requirements and grant applications, thus allowing them to focus on their work. For understaffed nonprofits operating with limited resources, this alternative way of giving is particularly valuable.

The recent rise of AI technologies, besides further enriching the tech industry, is also shaping tech moguls' engagement with philanthropy. Drifting away from the conventional practice of writing huge checks in support of philanthropic causes, tech philanthropies are now leveraging their own corporate expertise and technological resources to advance social good (Shi et al., 2020; Henriksen & Richey, 2022). The latest expression of the interlocking of the fields of philanthropy, humanitarianism, and technology, this AI-rooted philanthropic approach is commonly known as **AI for Social Good (AI4SG)** (Henriksen & Richey, 2022). While acknowledging that there is still a limited understanding of what exactly constitutes AI "for the social good" (Floridi et al., 2020; Shi et al., 2020), for the sake of clarity, this chapter adopts the following definition of AI4SG developed by Floridi et al. (2020): "the design, development, and deployment of AI systems in ways that (i) prevent, mitigate or resolve problems adversely affecting human life and/or the wellbeing of the natural world, and/or (ii) enable socially preferable and/or environmentally sustainable developments."

### **3 Technology's impact on philanthropy**

#### ***3.1 Tech-driven change in philanthropy – a double-edged sword***

With a legacy of supporting the advancement of scientific and technological developments, as illustrated in Section 2, the philanthropic sector itself is shaped by these developments. Recent technological advancements have been both positively and negatively disrupting traditional philanthropic practices.

On the one hand, POs have been benefiting from technological innovations on multiple fronts, from the creation of new avenues for donor engagement and fundraising, to the facilitation of impact measurement and reporting activities (Bernholz & Skloot, 2010). The emergence of online giving platforms, combined with the rise of digital communication and social media, is amplifying the reach and effectiveness of philanthropic efforts, allowing for information and resources to travel at an unprecedented speed. By providing greater access to information and lower barriers to entry, digital giving is contributing to the democratization of philanthropy and the forming of "networked philanthropy" (Bernholz et al., 2010; Arrillaga-Andreessen, 2015). Not only can people give directly to the causes they care about, but innovative giving mechanisms such as crowdfunding allow small donors to come together and pool their resources for greater impact, while forming networks dedicated to finding solutions to complex social problems. From peer-to-peer fundraising platforms such as GoFundMe and JustGiving, to the global philanthropic collaborative Co-Impact, which runs million-dollar funds and is financed by some of the world's most resourceful philanthropic actors, technology has been key to pushing down barriers to philanthropic collaboration (Co-Impact, 2023). Moreover, by narrowing the gap between giver and receiver, these technologies have the potential to empower both actors, giving donors more control and

information over their contributions and providing receivers a medium through which they can independently voice their demands (Arrillaga-Andreessen, 2015).

On the other hand, the adoption of technological tools by philanthropies continues to present a number of challenges, risks, and ethical considerations, as discussed in the literature (Taddeo, 2016, 2017; Bernholz & Reich, 2017; Floridi et al., 2018, 2020; Kanter & Fine, 2020; O'Brien, 2022). A first challenge is the availability of data within the philanthropic sector. Databases of POs' activities and strategies are often unavailable, incomplete, inaccurate, or contain irrelevant data (known as "data deserts") (Tudor et al., 2024). These shortcomings can severely hinder POs' ability to leverage the power of AI and, in worst-case scenarios, heighten bad practices that can dangerously magnify and reinforce preexisting inequalities and bias (Kanter & Fine, 2020; O'Brien, 2022). Second, even when data is available, storage and handling practices may not be aligned with the work of POs. Most AI software is designed to extract the maximum profit from digital data, which often entails the collection, long-term holding, and handling of digital data. Such practices, particularly in vulnerable humanitarian settings, can be dangerous and lead to discrimination and polarization (Tudor et al., 2024). In addition to these more tangible risks, several ethical concerns surround the incorporation of AI technologies in POs, as explored in detail by Floridi et al. (2018, 2020), Taddeo (2016, 2017), and Bernholz and Reich (2017). Overall, it appears that despite the numerous ethical frameworks and principles for AI that have been suggested, there remains a sense of disillusionment about their effectiveness, with POs questioning whether these frameworks adequately address the specific requirements of their sector (Coppi et al., 2021). This may well combine with a fear of alienation; being a largely human-centric sector, the delegation of tasks from humans to machines may be perceived as unnatural and inadequate by many philanthropic professionals, who may view it as a dilution of their efforts (Tudor et al., 2024).

External risks, such as cyber-attacks and data breaches, also exacerbate nonprofits' mistrust in technological tools and affect the sector's digitalization. Prominent examples include the cyber-attack conducted against the International Committee of the Red Cross (ICRC) and the data breach against the NPO Broward Health of California, both in 2022 (CBS Miami, 2022; Duguin, 2022; ICRC, 2022). In the United Kingdom (UK), the government's Cyber Security Breaches Survey reported in the winter of 2022–2023 that 24% of UK charities had been victims of cyber breaches and/or attacks (United Kingdom Government Department of Science, Innovation & Technology, 2023). However, the UK government data also comes with the recognition that the charity sector "still has a long way to go" when it comes to preventing and responding to such attacks (United Kingdom Government Department of Science, Innovation & Technology, 2023). In the Asia-Pacific region (APAC), the numbers are even higher. Infochange's recent APAC NGO Digital Capacity report shows that one in six of the surveyed nonprofits had been the victim of a cybersecurity incident in the past year, with the number rising to one in three in Indonesia (Infochange, 2023). These examples highlight that the relationship between NPOs and technological tools must be one of understanding, not just adoption; nonprofits must invest in building the necessary infrastructure to ensure the safe and effective use of these tools. This, undoubtedly, requires an investment of resources on the part of nonprofits that may not always be readily available.

Overall, the recent fast pace of AI development is opening up a myriad of new opportunities for all sectors, including the philanthropic sector. At the same time, however, this acceleration of digitalization has strained the ability of some actors to rapidly build the infrastructure needed to successfully adopt and benefit from AI-powered tools. In this race to adapt, the nonprofit sector has been lagging behind, with the sector continuing to have one of the lowest rates of AI usages (Google, 2019; Herzog et al., 2021). A recent survey investigating Swiss POs' current and potential use of AI tools appears to support this trend, indicating that, with a few exceptions,

the majority of Swiss philanthropies do not use AI tools or do so minimally, with less than 15% of POs reporting the use of any form of AI (Della Giovampaola et al., 2023). Moreover, the survey reveals a mismatch between POs' areas of current AI use and areas of desired AI support. Another study on Swiss POs also found an overall low level of digital presence, with only 30% of POs mapped across Switzerland having live websites (Tudor et al., 2024). In the UK, the Charity Digital Skills annual reports (2017–2023)<sup>3</sup> outline how the country's nonprofit sector continues to have a digital skills gap characterized by a lack of resources and unclear digital strategies, even after the Covid-19 pandemic and lockdown, which forced the sector to largely go remote (Charity Digital Skills Report, n.d.). At the European level, a 2023 survey led by Philea on data science, AI, and data philanthropy in foundations across Europe showcases how, despite the diverse spectrum of data maturity levels among foundations, the internal use of AI and data science remains widely infrequent, with only a handful of exceptions (Candela et al., 2024). The survey report identifies a lack of expertise and know-how as the primary reason for foundations' lack of engagement with AI. Noting that these considerations are informed by the European context, which limits their generalizability, the lack of data and reports on the digitalization of philanthropies in other regions could be taken as an indication that the sector is also lagging behind elsewhere.

In contrast, tech philanthropies are leading philanthropic actors, especially in the AI for Social Good (AI4SG) space. While traditional philanthropic organizations struggle to adopt AI and digital technologies, tech-focused philanthropies, particularly those specializing in AI4SG, are at the forefront of this movement. These tech philanthropies are not only more adept at using AI, but they are also driving innovation in this space. The contrast, then, is that while the broader philanthropic sector is lagging in AI adoption and struggling with digital transformation, a specific subset of the sector – tech philanthropies, especially those focused on AI4SG – are not only adapting but leading in the use of AI for philanthropic purposes. This creates a divide within the sector, where the capabilities and impact of different types of philanthropic organizations vary significantly based on their engagement with and adoption of AI technologies. This reinforces a somewhat paradoxical relationship between technology and philanthropy. While non-tech-led philanthropies continue to fund technological advancements, with AI serving as a prime contemporary example, they remain cautious about the widespread adoption of technological tools, including those they themselves fund.

In sum, while technological innovations offer unprecedented opportunities to democratize giving, enhance donor engagement, and foster collaborative impact, they also pose significant challenges, such as data privacy concerns, cybersecurity threats, and the potential to exacerbate inequalities. This way, the tech-driven change in philanthropy comes with both opportunities and challenges, representing a double-edged sword.

### ***3.2 Data philanthropy – an example of opportunities and challenges***

The recent phenomenon of “Data Philanthropy” exemplifies the opportunities and challenges that derive from integrating AI in philanthropy and how these can impact POs' digitalization. AI tools require data to operate, even when they are used to achieve social good. The digitalization of POs and the integration of AI technologies has created a demand for data, on the part of POs. While the philanthropic arms of tech companies can draw from their parent companies' data storages, this is somewhat unnatural for the rest of the philanthropic sector. First, the sector suffers from so-called “data deserts” due to a lack of good practices for uniformly collecting, filtering, and storing complete and accurate data (Kanter & Fine, 2020). Moreover, uneven data availability entails that issue areas where data is more abundant, such as health and climate change, receive far more attention,

as opposed to peace and justice, an issue area more complex to capture with data (Google, 2019). Second, due to the sensitive nature of the information, philanthropies often gather data with the principle of “collect little and destroy as soon as possible” (Bernholz & Reich, 2017). The corporate sector, on the contrary, is an important collector of data, particularly given the great value data holds in today’s digital civil society (Lev Aretz, 2019). Thus, data itself has now become a philanthropic resource, potentially on par with the more traditional financial and human resources donated to philanthropic causes.

Data philanthropy,<sup>4</sup> the donation of data from private companies and individuals for socially beneficial purposes, and data-raising, the effort to get people to give their data for a cause, are gaining traction (Taddeo, 2016, 2017; Lev Aretz, 2019; Bernholz, 2021). Data philanthropy, in particular, is becoming increasingly popular following the pioneering 2015 Ncell-Flowminder collaboration that used mobile data to track the displacement of individuals after the Nepal earthquake (Lev Aretz, 2019). Today, private sector companies such as Pfizer, Genentech, and Reddit are donating data to organizations, including the UN. The practice of data philanthropy offers the opportunity to harness the value of data for the social good, unlocking the many benefits that are derived from the sharing of information, especially in emergency settings (Taddeo, 2016). It also allows for the harnessing of an abundant resource. In today’s digital civil society, data is constantly being generated, whether actively through the use of devices or passively, such as passing through controlled spaces (Bernholz, 2021). While this constant tracking has many drawbacks, it also provides significant access in times of need.

Nonetheless, data philanthropy differs from the donation of other resources such as financial or human resources. This is because while philanthropy has, traditionally, focused on voluntary giving of private resources, the ownership of which is largely clear and undisputed, digital data donated by private companies is contested property (Bernholz & Reich, 2017; Lev Aretz, 2019; Bernholz, 2021). The question as to who is the “true” owner remains: the person whose information is involved, the company that provides the software collecting the data, or the platform on which the data is collected? According to Taddeo (2017), data philanthropy is both morally ambiguous and desirable. It is morally ambiguous because, as currently practiced, it is in tension with individual rights, and desirable because of the positive change it can promote, such as speeding emergency responses and advancing scientific knowledge. This tension between individual rights and data philanthropy, Taddeo (2017) explains, is operational rather than structural, and can and should be resolved by putting in place the right ethical principles, protocols, and infrastructure. While the recent enactment of the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) are important steps toward giving individuals greater control over their data, the question of data ownership is far from resolved.

At the same time, regulations governing how AI technologies use this data remain largely inadequate. The fast pace of development of AI systems exacerbates this challenge, not only making it very difficult for policymakers and ethicists to keep up but also creating an imbalance of knowledge between the developers of AI systems and those charged with regulating them. The European Union’s (EU) AI Act, which was passed in 2023 and will come into force at the end of 2025, is the world’s first comprehensive AI law and an important step in the regulation of AI. Nevertheless, the regulation of AI and its data lags far behind the rapid developments seen in the AI space.

Overall, tech philanthropies’ access to and ability to manage this new form of philanthropic resource, needed to run AI systems, namely data, allows them to position themselves as leaders in integrating AI into philanthropy. Moreover, tech philanthropies have the necessary mindset to drive change due to their culture of innovation. On the contrary, non-tech philanthropies, which suffer – from the start – from a low level of digitalization and potential resource limitations due to their

nonprofit nature, are at a disadvantage when it comes to capitalizing on the benefits that AI can bring to philanthropic actions. This divergence between the tech and non-tech philanthropies, however, should not obfuscate the fact that the innovation promoted by tech giants, including that advanced under the label of “philanthropy,” does not always result in public social good. Traditional philanthropic entities, with their expertise, can and should act as important checks and balances on AI philanthropy. In other words, while tech philanthropies can help the sector benefit from AI technologies, NPOs can work to mitigate the potential harms of AI in the field of social good. As Floridi et al. (2018) outline, it is not only the misuse of AI that needs to be avoided but also its underuse. As Section 4 expands, the discrepancy between the nonprofit and for-profit AI4SG actors can be successfully bridged through cross-sector, multi-stakeholder partnerships, to ensure the successful deployment of AI.

## **4 AI for Social Good – stakeholders and modern applications**

Having navigated the historical developments of the intersection of these two fields, AI and philanthropy, we now turn our attention to the main movement dominating this landscape – AI for Social Good (AI4SG). This section will illustrate some of the key players in the field, as well as emerging trends and modern applications.

### **4.1 AI4SG stakeholders**

Leaders in the research, development, and implementation of AI systems, transnational tech companies are driving the integration of AI4SG through their philanthropic arms. Prominent examples include Google.org, Microsoft’s AI for Good, and IBM’s Open Source (see more stakeholder examples in Table 1.1), which are using their products, technological expertise, and financial resources to advance philanthropic endeavors. For example, in 2023 alone, Google.org invested \$1 million to train NGOs in AI and cybersecurity, while Microsoft spent \$60 million to empower NPOs and other organizations tackling the world’s most challenging health issues, in part by providing them with the necessary AI tools and expertise (Choudhary, 2023; Microsoft, 2023).

However, the dominant role of tech companies should not obfuscate the fact that the AI4SG ecosystem is also populated by a variety of other stakeholders. It is important to recognize who these stakeholders are, their contribution(s), and how they interact, as it is the way they interact that determines the why, what, and how AI tools are used to advance social good. Table 1.1 provides an overview of the main actors present in the space of AI4SG, categorized by role.

The use of AI to propel positive societal impact can stem from one of the following three stakeholder dynamics:

- 1 Tech companies purposely developing AI technologies for social good;
- 2 NPOs adopting AI tools designed for the for-profit market and using them to propel positive societal impact;
- 3 NPOs partnering with tech companies to create new AI tools (Bernholz & Reich, 2017; Kanter & Fine, 2020; Shi et al., 2020; Henriksen & Richey, 2022).

The three dynamics show how, even though actors in the AI4SG space may act alone or in partnership(s), tech companies are always present in one form or another. This presence can be direct, when they consciously develop an AI tool intended to bring societal good, or indirect, when their AI for-profit products are utilized for such purposes. In other words, as Fine and Kanter (2020) note, a clear trend when it comes to AI and giving is the need of NPOs to partner with tech

Table 1.1 Primary, secondary, and regulatory AI4SG stakeholders

	<i>Stakeholder</i>	<i>Examples</i>
Primary stakeholder: Developer, user, and/or deployer of AI4SG technologies	Technology company or philanthropic arm of tech companies	Microsoft Philanthropies Google.org Google AI for Social Good DeepMind Ethics & Society NVIDIA Foundation
	Philanthropic organizations (encompasses all nongovernmental organizations working to achieve positive societal impact)	Mastercard Impact Fund Bloomberg Philanthropies Robin Hood Foundation OpenAI Inc.
Secondary stakeholder: Financier or beneficiary of AI4SG technologies	Donor	The Rockefeller Foundation The Ford Foundation The Bill & Melinda Gates Foundation The Open Society Foundations
	Beneficiary	Local communities affected by climate change Patients benefiting from AI-driven healthcare initiatives Students in underprivileged areas receiving AI-enhanced education Small-scale farmers using AI for agricultural improvement Refugees and displaced people receiving aid through AI-enabled systems
	(Potentially) Technology provider	Amazon Web Services (AWS) IBM Watson

(Continued)

Table 1.1 (Continued)

	<i>Stakeholder</i>	<i>Examples</i>
Regulatory stakeholder: Provide the legal and/or ethical framework for the development and/or deployment of AI4SG technologies	Governments <sup>a</sup> or supranational organizations	European Union’s AI Act <sup>b</sup> European AI Alliance
	International organizations	United Nations’ International Telecommunication Union (ITU) AI for Good Initiative <sup>c</sup> World Economic Forum – Centre for the Fourth Industrial Revolution
	Research institutions	Data Science for Social Good – University of Chicago AI Now Institute – New York University Ada Lovelace Institute (an independent research institute, funded by the Nuffield Foundation) Alan Turing Institute Oxford Internet Institute – University of Oxford Stanford Institute for Human-Centered Artificial Intelligence (Stanford HAI) The Berkman Klein Center for Internet & Society Swiss Data Science Center (SDSC) – a joint venture between EPFL and ETH Zurich Centre for Artificial Intelligence Policy (CAIP) – University of Zurich AI Ethics Lab – University of Basel Electronic Frontier Foundation (EFF)
	Advocacy groups	Access Now Future of Life Institute (FLI) Center for Humane Technology Algorithmic Justice League Partnership on AI (PAI)

<sup>a</sup> Governments are also an important element of these partnerships. Not only do they provide the legal framework within which these partnerships can operate, but both for-profits and NPOs often depend on governments for public data sources (Bernholz & Reich, 2017).

<sup>b</sup> <https://artificialintelligenceact.eu/the-act/>.

<sup>c</sup> <https://aiforgood.itu.int/>.

*Relationship between AI and philanthropy*

*Table 1.2* Examples of partnerships between NPOs and for-profit companies

<i>For-profit</i>	<i>NPO</i>	<i>Objective</i>
Maxar Technologies’ DigitalGlobe (a satellite imaging company)	USA for UNHCR (a nonprofit created to support UNHCR)	Provide satellite imaging to support with refugee assistance.
Microsoft	Operation Smile	Develop a facial modeling algorithm, which works with Microsoft Pix, to improve facial surgeries.
Salesforce.org	Philanthropy Cloud (philanthropic arm of Salesforce.org)	An employee engagement database product for corporations to facilitate employee giving, volunteering, and other social impact activities.

Based on Kanter and Fine (2020).

companies. Table 1.2 lists some examples. On paper, the partnering of NPOs with tech companies seems ideal. NPOs bring sector expertise and access to the problem(s) being addressed while tech companies provide the necessary resources and technical know-how, as otherwise very few organizations have both the social and technical expertise to successfully design and implement AI for good projects (Gosselink & Bromberg, 2019). Lacking either social or technical expertise, the risk of unintended consequences upon deployment majorly increases.

Increasing interconnectedness with and dependence on the private sector presents both opportunities and obstacles for the nonprofit sector. On the one hand, these partnerships can greatly benefit NPOs and their work. In the case of POs, the use of AI can improve operational efficiency, donor engagement, grantmaking, monitoring and evaluation, and communication. It allows organizations to be more transparent, thus building trust with their public (Chu & Wang, 2019; Kanter & Fine, 2020). AI technologies can also uncover synergistic partnerships among philanthropic actors, thus enhancing collaboration and maximizing pooling of resources for greater social impact (Tudor et al., 2024). In addition, once successfully adopted, these technologies can allow POs to cut down costs and operate on “new” budgets.

On the other hand, these ties can hinder the independence of the nonprofit sector. In particular, proximity to the for-profit sector leads to the clash of two very different cultures about the relationship between profit generation and social change. As Henriksen and Richey (2022) note in their research on Google’s tech philanthropy, profitability is highlighted as a key element in the use of AI4SG, with profit generation seen as positive for the advancement of social change. But Google is not the only example. This form of “for-profit philanthropy,” which combines making money with doing good, is particularly popular among tech philanthropists. The Chan Zuckerberg Initiative (CZI) caused quite a stir in 2015 when it registered as a for-profit limited liability company (LLC), openly blurring the lines between philanthropy and investment. Others, such as Peter Thiel and John Doerr, have also set up mechanisms designed to generate returns on their philanthropic investments. Open AI is another prominent example. Founded in 2015 by Elon Musk, Peter Thiel, and Sam Altman, among others, the research organization now consists of two entities: a nonprofit research segment, OpenAI Inc., and a for-profit subsidiary, OpenAI Global LLC, which was established at a later date to enable the commercialization of its AI technologies and applications.



Although this new approach of for-profit philanthropy can free philanthropic actors from some of the constraints of the nonprofit status and open new avenues for continuous reinvestment, Henriksen and Richey (2022) note how the message behind AI4SG problematically “frames controversial and profitable data practices as having public value, [...] obscuring the power relations and politics of digital capitalism.” Notably, this juxtaposition of different values is possible, in part, because AI4SG remains a vague concept, as there is still limited understanding of what exactly constitutes AI “for the social good” (Floridi et al., 2020; Shi et al., 2020). The lack of a clear definition can benefit AI4SG, allowing it to grow and innovate beyond definitional boundaries. Floridi et al. (2020) outline how “context-specific design and deployment could prevent such value misalignment and deliver successful AI4SG projects on a more consistent basis.”

#### 4.2 AI4SG modern applications

Alongside the diverse landscape of stakeholders of the AI4SG movement, equally important are the practical applications of this technology in philanthropic efforts. This section illustrates some of the varied ways in which AI is being used to address social challenges and enhance philanthropic initiatives. It is worth noting that under the umbrella of AI for philanthropic purposes, we distinguish between two types:

- 1 AI adopted by POs for an *internal* purpose, i.e., adopting AI technologies as part of the organization’s normal operations to improve operational efficiency, such as AI-powered donor matching; and
- 2 AI used by POs or tech philanthropies<sup>5</sup> for an *external* purpose, i.e., adopting AI technologies to enhance their social impact, for AI-based satellite imagery analysis to better mitigate crisis response. The latter case is closely related to the broader concept of AI4SG, as previously defined in Floridi et al. (2020).

Concerning type (1) above, as discussed earlier, traditional POs remain either hesitant or under-resourced to adopt AI tools for internal purposes, lagging behind other sectors in their level of digitalization and AI adoption. Traditional POs’ unfamiliarity with AI systems also leads to a limited deployment of AI technology for external operations (type 2), which can hinder their ability to achieve philanthropic impact. In contrast, tech philanthropies are deploying extensive AI solutions for both internal (type 1) and external (AI4SG, type 2) purposes. This comes as no surprise, considering that “philanthropy is just a drop in the bucket compared to the goliath-sized tech platforms, the goliath-sized AI companies, the goliath-sized regulators and policymakers that can actually take a crack at this” (Dervishi, 2023). In addition, insights from the AI Index (Stanford University, 2023) reveal a shifting landscape in which the tech industry has rapidly outpaced academia in developing state-of-the-art AI and machine learning algorithms since 2014, reiterating its clear leadership in the AI space.

Most major tech companies have initiated AI4SG programs. For example, Microsoft Philanthropies has launched five initiatives: *AI for Health*, *Earth*, *Accessibility*, *Humanitarian Action*, and *Cultural Heritage*; as well as a closely related program called *Data for Society*. Alphabet, Google’s parent company, has several programs, including *AI for Social Good*, *AI Impact Challenge*,<sup>6</sup> and *AI for Global Goals*.<sup>7</sup> Examples of implemented projects include preventing blindness by detecting diabetic retinopathy with AI,<sup>8</sup> forecasting river floods,<sup>9</sup> building greener cities,<sup>10</sup> and helping people with non-standard speech be better understood.<sup>11</sup>

From the side of traditional POs,<sup>12</sup> such as Novartis Foundation, several notable AI for Health have been developed. *AI4Leprosy* aims to accelerate leprosy detection through image analysis of skin lesions, while *AI4BetterHearts* is pooling cardiovascular health data from hospitals and primary care centers to improve heart health outcomes globally. Similarly, *AI4HealthyCities* set out to understand how heart health can be improved by modifying the underlying social, economic, or environmental health determinants. Lastly, they partnered with Tencent to develop an AI nurse for patients diagnosed with heart failure – used to anticipate disease progression and provide targeted interventions, while allowing medical practitioners to track patients remotely.

In such a vast and exorbitantly fast-paced field, understanding the landscape of possible AI4SG use cases is not a defined end goal, but rather an ever-shifting landscape of novel, emerging solutions. Overall, AI4SG projects have been implemented across most sectors, domains, and Sustainable Development Goals (SDG), with a McKinsey Global Institute (2018) report mapping over 160 non-exhaustive use cases across ten sectors. Examples range from improving cancer diagnostics, to enhancing blind people’s ability to better navigate their environment, to aiding disaster relief efforts by using AI to analyze satellite imagery. It is worth noting that the pace of AI has evolved exponentially since 2018, with 2023 alone seeing the unprecedented rise of generative AI and the popularization of Large Language Models (LLMs), meaning that the number of use cases today has increased drastically. For instance, LLMs are finding new applications in fields previously thought to be the exclusive domain of human labor, such as mental health care (Ji et al., 2023; Xu et al., 2023). Today’s use cases are virtually limitless to any social or environmental issue, provided the right data can be sourced and fed to an appropriate AI model (see Figure 1.1). Other notable examples include increasing accessibility for vulnerable populations,<sup>13</sup> supporting crisis response interventions,<sup>14</sup> human rights,<sup>15</sup> climate change,<sup>16</sup> charitable giving (Kanter & Fine, 2020), civic engagement,<sup>17</sup> and predicting poverty using satellite imagery (Jean et al., 2016).

In summary, while tech philanthropies are pioneering the AI4SG space, traditional philanthropic organizations are increasingly recognizing the imperative to adapt and integrate these powerful tools. However, as the sector evolves, it must also navigate the complexities of cybersecurity, data privacy, and ethical use to ensure that technological advancements effectively serve its mission to foster social good. To ensure that it fulfills this mission, it can be guided by principles to “become good at AI for good” (Kshirsagar et al., 2021). Key among these principles are:

Healthcare	Environment	Crisis Response	Climate	Education
<ul style="list-style-type: none"> <li>• Drug discovery</li> <li>• Cancer screening</li> <li>• Automated diagnoses</li> <li>• Epidemic modelling</li> </ul>	<ul style="list-style-type: none"> <li>• Wildlife conservation</li> <li>• Predicting plant disease</li> <li>• Preventing overfishing</li> <li>• Predicting wildfires</li> </ul>	<ul style="list-style-type: none"> <li>• Satellite data to assess damage on a large area (floods, earthquakes)</li> <li>• Find missing people</li> </ul>	<ul style="list-style-type: none"> <li>• Predicting extreme precipitation</li> <li>• Modeling carbon sequestration</li> <li>• Energy efficiency</li> </ul>	<ul style="list-style-type: none"> <li>• Enhance classroom learning</li> <li>• Maximise student achievement</li> <li>• Identifying student distress</li> </ul>

Examples of data sources include: satellite imagery, weather data, medical imaging, social media, text, images.

Figure 1.1 Examples of AI for Social Good (adapted from Tudor et al., 2024).

- 1 **Educational Alignment** – continually educate POs about AI’s potential and limitations, tailoring expectations to fit achievable and workflow-enhancing goals;
- 2 **Dynamic Scoping** – engage in ongoing dialogue with POs to develop solutions that remain practical and responsive to evolving needs;
- 3 **Data Integrity and Security** – ensure comprehensive understanding and management of datasets, their metadata, and associated privacy concerns to build reliable models. Several AI4SG domains suffer from subjective labeling or insufficient datasets;
- 4 **Inclusive Expertise** – integrate POs’ domain expertise into model development to enhance feature selection and engineering, model choice, and model regularization;
- 5 **Ethical and Practical Deployment** – prioritize project constraints and domain-specific metrics in model development and deployment, to create solutions that are both ethical and practical;
- 6 **Human-Centric AI Design** – maintain a “human-in-the-loop” approach to AI projects that actively engages POs in the modeling process for better aligned outcomes;
- 7 **Long-Term Commitment** – recognize the need for sustained engineering resources for maintaining and updating deployed models, focusing on efficiency and practical impact rather than just machine learning metrics (Kshirsagar et al., 2021).

In the absence of these guiding principles, AI4SG is not without risks, unclear ethical standards, or even unintended consequences. Some of these risks arise from the tech industry’s culture of moving fast and iterating solutions on the go (Tomašev et al., 2020), while not paying enough attention to long-term outcomes or sector-specific challenges. This culture is likely to bleed into tech philanthropies’ approach to AI4SG projects, which currently dominates this space and therefore requires greater attention to ethical principles. Long-term commitment, dynamic scoping, and deep partnerships between the nonprofit and for-profit sectors remain paramount.

## 5 Conclusion

The relationship between technology and philanthropy is neither new nor static, with the interlocking of the fields of AI and philanthropy being one of its most recent expressions (Henriksen & Richey, 2022). This chapter provides an overview of the intersection of AI and philanthropy, from the early support of POs to AI development, to the modern application of AI in philanthropic work, with the aim of understanding its current status and charting its future trajectory.

We begin by exploring the early role of philanthropy in supporting AI research and development, and then focus on the rise of tech philanthropy and the AI4SG movement. This history shows that, prior to the rise of tech philanthropy and the AI4SG movement in the early 21st century, the relationship between the two was primarily one of philanthropic funding. In other words, philanthropes would fund technological innovation, but would rarely be users of those technologies themselves. Fast forward to today, and apart from a few leading foundations, philanthropic funding in the field of science and technology is primarily directed at fostering innovation as a means to address global challenges such as education and health, rather than for the sake of developing innovation in the field. Tech philanthropists, with their resources and expertise, are disrupting this relationship, positioning themselves as both funders and users of technological tools like AI. This gap in technological knowledge and resources within the philanthropic sector has created a wide discrepancy in the sector’s digitalization levels.

In discussing the opportunities and challenges presented by philanthropies’ adoption of technological tools, the latest of which are AI-powered tools, we outline the multiple elements that contribute to POs’ varying AI adoption rates. This also helps explain, in part, why the majority of the

nonprofit sector is lagging behind when it comes to digitalization. In sum, we conclude that while technological innovations offer unprecedented opportunities to democratize giving, enhance donor engagement, and foster collaborative impact, they also pose significant challenges such as data privacy concerns, cybersecurity threats, and the potential to exacerbate inequalities. Overcoming these challenges is important to enable an equitable and inclusive digitalization of the sector.

Finally, we dive into the AI4SG movement, mapping key stakeholders and prominent AI4SG initiatives and trends. Again, the AI4SG movement is not static, but rather an ever-shifting landscape of novel, emerging solutions. These solutions are widely applicable across most sectors, domains, and SDGs, given appropriate and ethically sourced data that can be fed into appropriate AI models. However, this wide applicability speaks to the versatility and utility of such kinds of tools in addressing social and environmental issues, without forgetting the principles of “becoming good at AI for Good” (Kshirsagar et al., 2021) for successful implementation and deployment.

### Notes

- 1 In total, 150 respondents from 30 countries completed the survey used to inform the report (Rockefeller Philanthropy Advisors, 2022).
- 2 The Global Philanthropy Tracker (GPT) presents data on four flows – philanthropic outflows, official development assistance (ODA), remittances, and private capital investment – for 47 countries. The data gathered represent the year 2020 or the most recent year with available data.
- 3 Established in 2017, the Charity Digital Skills Report is the annual barometer of UK charities’ digital skills, attitudes, behaviors, and needs.
- 4 The term was reportedly coined by World Economic Forum CTO Brian Behlendorf during a spontaneous conversation at the 2011 World Economic Forum (Lev Aretz, 2019).
- 5 We distinguish between traditional philanthropic organizations (POs) and the philanthropic arm of tech companies (i.e., tech philanthropies) because the latter have access to massive tech capital and deep know-how tech expertise that most traditional POs often lack.
- 6 The Google AI Impact Challenge is an open call to nonprofits, social enterprises, and research institutions worldwide to submit their ideas for using AI to address social and environmental challenges. The program aims to support projects that address issues in the areas of health, economic opportunity and empowerment, environmental protection and conservation, education, misinformation, and crisis and emergency response.
- 7 Google’s AI for Global Goals initiative is a program that aims to accelerate progress on the United Nations’ Sustainable Development Goals (SDGs) by supporting organizations that use artificial intelligence (AI) to address social and environmental challenges.
- 8 Google’s Project ARDA.
- 9 Google’s Flood Forecasting Project.
- 10 Google’s Project Green Light.
- 11 Google’s Project Relate.
- 12 It is worth noting that it is predominantly large-scale POs that have been implementing and deploying AI tools, whereas most of the nonprofit sector, made up of small actors, is lagging behind.
- 13 E.g., Google’s Project Along and Project Relate.
- 14 E.g., Facebook Disaster Maps.
- 15 UN Universal Human Rights Index.
- 16 E.g., Google’s Project Contrails.
- 17 E.g., Salesforce’s Philanthropy Cloud.

### References

- Arrillaga-Andreessen, L. (2015). Disruption for Good. *Stanford Social Innovation Review*, 13(2), 34–39. <https://doi.org/10.48558/X8FD-ZC88>
- Bennett, N., Resney, R., Carter, A., & Woods, W. (2016, February 10). How Tech Entrepreneurs Are Disrupting Philanthropy. *Boston Consulting Group Global*. <https://www.bcg.com/publications/2016/innovation-strategy-how-tech-entrepreneurs-are-disrupting-philanthropy>

- Bernholz, L. (2021). *Philanthropy and Digital Civil Society: Blueprint 2022*. Stanford PACS. <https://pacs-center.stanford.edu/publication/philanthropy-and-digital-civil-society-blueprint-2022/>
- Bernholz, L., & Reich, R. (2017). *Nonprofit Data Governance*. Stanford PACS. <https://pacscenter.stanford.edu/publication/nonprofit-data-governance/>
- Bernholz, L., Skloot, E., & Varela, B. (2010). *Disrupting Philanthropy: Technology and the Future of the Social Sector*. Center for Strategic Philanthropy and Civil Society at Duke University Sanford School of Public Policy: Belgium.
- Bishop, M., & Green, M. (2008). *Philanthro-capitalism: How the Rich Can Save the World* (1st U.S. ed). Bloomsbury Press: New York.
- Bremner, R. H. (1994). *Giving: Charity and Philanthropy in History* (1st, 1st pbk. ed.). Routledge: New York. <https://doi.org/10.4324/9780203790724>
- Candela, F., Kilicalp, S., & Spiers, D. (2024). *Data Science, AI and Data Philanthropy in Foundations: On the Path to Maturity*. Philea and Fondazione Compagnia di San Paolo. <https://philea.issueelab.org/resource/data-science-ai-and-data-philanthropy-in-foundations-on-the-path-to-maturity.html>
- CBS Miami (2022, January 3). Broward Health Suffered Data Breach That Exposed Personal Info of Patients, Employees. *CBS Miami*. <https://www.cbsnews.com/miami/news/broward-health-suffered-data-breach/>
- Charity Digital Skills Report (n.d.). Digital Skills Report for the Charity Sector—Introduction. *Charity Digital Skills Report*. Retrieved 31 August 2023, from <https://charitydigitalskills.co.uk/the-charity-digital-skills-report-introduction/>
- Choudhary, L. (2023, August 24). Google.org Invests \$1m to Train NGOs in Asia on AI, Cybersecurity. *Tech in Asia*. <https://www.techinasia.com/google-org-invests-us-1m-to-train-ngos-in-asia-on-ai-cybersecurity#:~:text=2%20min%20read-,Google.org%20invests%20%241m%20to%20train,in%20Asia%20on%20AI%2C%20cybersecurity&text=Google.org%2C%20the%20tech%20giant's,data%20analytics%2C%20and%20impact%20reporting.>
- Chu, P., & Wang, O. Y. (2019). *Philanthropy in China*. Asian Venture Philanthropy Network [https://avpn.asia/wp-content/uploads/dlm\\_uploads/2019/01/Philanthropy-in-China\\_Web-Version-1.pdf](https://avpn.asia/wp-content/uploads/dlm_uploads/2019/01/Philanthropy-in-China_Web-Version-1.pdf)
- Church, A. (2020, May 7). Philanthropy's Response to COVID-19 Now More Than \$10 Billion Worldwide. *Candid Blog*. <https://blog.candid.org/post/philanthropys-response-to-covid-19-now-more-than-10-billion-worldwide/>
- Co-Impact (2023). *A Global Collaborative for Systems Change*. Co-Impact. <https://co-impact.org/>
- Coppi, G., Moreno Jimenez, R., & Kyriazi, S. (2021). Explicability of Humanitarian AI: A Matter of Principles. *Journal of International Humanitarian Action*, 6(1), 19. <https://doi.org/10.1186/s41018-021-00096-6>
- Council on Foundations (2020, July 15). *Policy Brief: Foundation Payout and the COVID-19 Crisis*. Council on Foundations. <https://cof.org/content/policy-brief-foundation-payout-and-covid-19-crisis>
- Della Giovampaola, C., Tudor, M. C., Gomez, L., & Ugazio, G. (2023, June 14). Current and Potential AI Use in Swiss Philanthropic Organizations—Survey Results. *SwissFoundations*. <https://www.swiss-foundations.ch/fr/actualites/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/>
- Dervishi, K. (2023, August 11). Foundations Seek to Advance AI for Good—And Also Protect the World from Its Threats. *AP News*. <https://apnews.com/article/ethical-ai-foundations-philanthropy-6021ffd4ca62c7b7064af0e524878307>
- Di Mento, M. (2022, December 30). Bill Gates Made 2022's Biggest Charitable Donation. *Los Angeles Times*. <https://www.latimes.com/business/story/2022-12-30/bill-gates-made-2022s-biggest-charitable-donation-5-billion>
- Duguin, S. (2022, February 22). Cyberattacks: A Real Threat to NGOs and Nonprofits. *CyberPeace Institute*. <https://reliefweb.int/report/world/cyberattacks-real-threat-ngos-and-nonprofits>
- Falk, A., & Michelson, E. S. (2021). A Vision for the Future of Science Philanthropy. *Issues in Science and Technology*. <https://issues.org/future-science-philanthropy-sloan-michelson-falk/>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Floridi, L., Cows, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 26(3), 1771–1796. <https://doi.org/10.1007/s11948-020-00213-5>

- Google. (2019). *Accelerating Social Good with Artificial Intelligence: Insights from the Google AI Impact Challenge*. [https://services.google.com/fh/files/misc/accelerating\\_social\\_good\\_with\\_artificial\\_intelligence\\_google\\_ai\\_impact\\_challenge.pdf](https://services.google.com/fh/files/misc/accelerating_social_good_with_artificial_intelligence_google_ai_impact_challenge.pdf)
- Gosselink, B. H., & Bromberg, C. (2019, September 12). *2,602 Uses of AI for Social Good, and What We Learned from Them*. Google. <https://blog.google/outreach-initiatives/google-org/2602-uses-ai-social-good-and-what-we-learned-them/>
- Henriksen, S. E., & Richey, L. A. (2022). Google's Tech Philanthropy: Capitalism and Humanitarianism in the Digital Age. *Public Anthropologist*, 4(1), 21–50. <https://doi.org/10.1163/25891715-bja10030>
- Herzog, P. S., Naik, H. R., & Khan, H. A. (2021). *AIMS Philanthropy Project: Studying AI, Machine Learning & Data Science Technology for Good* [Technical Report]. Indiana University Lilly Family School of Philanthropy and Indiana University School of Informatics and Computing, IUPUI, Indianapolis, IN. <https://scholarworks.iupui.edu/handle/1805/25177>
- ICRC. (2022). *Cyber Attack on ICRC: What We Know* (Europe and Central Asia/Switzerland). <https://www.icrc.org/en/document/cyber-attack-icrc-what-we-know>
- Indiana University Lilly Family School of Philanthropy (2023). *Global Philanthropy Tracker 2023*. <https://globalindices.iupui.edu/tracker/index.html>
- Infoxchange (2023). *APAC NGO Digital Capability Report*. Infoxchange. [https://digitaltransformation.ngo/sites/default/files/IX\\_APACReport23\\_FA2-Screen.pdf](https://digitaltransformation.ngo/sites/default/files/IX_APACReport23_FA2-Screen.pdf)
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science*, 353(6301), 790–794. <https://doi.org/10.1126/science.aaf7894>
- Ji, S., Zhang, T., Yang, K., Ananiadou, S., & Cambria, E. (2023). *Rethinking Large Language Models in Mental Health Applications* (arXiv: 2311.11267). arXiv. <https://doi.org/10.48550/arXiv.2311.11267>
- Johnson, P. (2018). *Global Philanthropy Report – Perspectives on the Global Foundation Sector*. John F. Kennedy School of Government, Harvard University. [https://cpl.hks.harvard.edu/files/cpl/files/global\\_philanthropy\\_report\\_final\\_april\\_2018.pdf](https://cpl.hks.harvard.edu/files/cpl/files/global_philanthropy_report_final_april_2018.pdf)
- Kanter, B., & Fine, A. (2020). *AI4Giving: Unlocking Generosity with Artificial Intelligence: The Future of Giving*. <https://drive.google.com/file/d/1sQFYZsyuQDXIZSLNdUbuU7ICWiF0DIRY4/view>
- Kshirsagar, M., Yang, S., Robinson, C., Gholami, S., Klyuzhin, I., Mukherjee, S., Nasir, M., Ortiz, A., Oviedo, F., Tanner, D., Trivedi, A., Xu, Y., Zhong, M., Dilkina, B., Dodhia, R., & Ferres, J. M. L. (2021, April 1). *Becoming Good at AI for Good*. AAAI/ACM Conference on AI, Ethics, and Society (AIES'21). <https://www.microsoft.com/en-us/research/publication/becoming-good-at-ai-for-good/>
- Kulish, N. (2021, December 20). Giving Billions Fast, MacKenzie Scott Upends Philanthropy. *The New York Times*. <https://www.nytimes.com/2020/12/20/business/mackenzie-scott-philanthropy.html>
- Lev Aretz, Y. (2019). Data Philanthropy. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3320798>
- Madianou, M. (2021). Nonhuman Humanitarianism: When 'AI for Good' Can Be Harmful. *Information, Communication & Society*, 24(6), 850–868. <https://doi.org/10.1080/1369118X.2021.1909100>
- Manning, C. (2020). *Artificial Intelligence Definitions*. Stanford University Human-Centered Artificial Intelligence. <https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf>
- McKinsey Global Institute. (2018). *Applying AI for Social Good | McKinsey*. <https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good>
- Michelson, E. S. (2020). *Philanthropy and the Future of Science and Technology*. Routledge: New York. <https://doi.org/10.4324/9780429444111>
- Microsoft (2023). AI for Health Empowers Researchers Who Are Using Artificial Intelligence to Advance the Health of People and Address Global Health Challenges Like COVID-19. <https://www.microsoft.com/en-us/ai/ai-for-health>
- O'Brien, C. (2022). *Big Data and A.I. for the SDGs: Private Corporation Involvement in SDG Data-Driven Development, Policy and Decision-Making*. United Nations Department of Economic and Social Affairs Sustainable Development. <https://sdgs.un.org/documents/big-data-and-ai-sdgs-private-corporation-involvement-sdg-data-driven-development-policy>
- Rockefeller Philanthropy Advisors (2019). *Philanthropy and the SDGs: Practical Tools for Alignment*. <https://www.rockpa.org/project/sdg/>
- Rockefeller Philanthropy Advisors (2022). *Global Trends and Strategic Time Horizons in Philanthropy 2022*. <https://www.rockpa.org/wp-content/uploads/2022/07/Time-Horizons-2022-1.pdf>
- Sealander, J. (2003). *Curing Evils at Their Source: The Arrival of Scientific Giving*. In L. Friedman & M. D. McGarvie (Eds.), *Charity, Philanthropy and Civility in American History*, 217–240. Cambridge University Press: Cambridge.

- Shi, Z. R., Wang, C., & Fang, F. (2020). *Artificial Intelligence for Social Good: A Survey* (arXiv: 2001.01818). arXiv. <https://doi.org/10.48550/arXiv.2001.01818>
- Shubinski, B. (2022, January 6). ‘A Roomful of Brains’: Early Advances in Computer Science and Artificial Intelligence. *Rockefeller Archive Center*. <https://resource.rockarch.org/story/a-roomful-of-brains-early-advances-in-computer-science-and-artificial-intelligence/>
- Stanford University, 2023. (2023). *AI Index Report 2023 – Artificial Intelligence Index*. <https://aiindex.stanford.edu/report/>
- Taddeo, M. (2016). Data Philanthropy and the Design of the Infraethics for Information Societies. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160113. <https://doi.org/10.1098/rsta.2016.0113>
- Taddeo, M. (2017). Data Philanthropy and Individual Rights. *Minds and Machines*, 27(1), 1–5. <https://doi.org/10.1007/s11023-017-9429-2>
- The Economist (2023, February 9). How a Tide of Tech Money Is Transforming Charity. *The Economist*. <https://www.economist.com/international/2023/02/09/how-a-tide-of-tech-money-is-transforming-charity>
- Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C. M., Ezer, D., Haert, F. C. van der, Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., Wever, W. de, ..., & Clopath, C. (2020). AI for Social Good: Unlocking the Opportunity for Positive Impact. *Nature Communications*, 11(1), Article 1. <https://doi.org/10.1038/s41467-020-15871-z>
- Torres, J., & Zinsmeister, K. (2018). The Calculating Philanthropy of Silicon Valley. *Philanthropy Roundtable*. <https://www.philanthropyroundtable.org/magazine/the-calculating-philanthropy-of-silicon-valley/>
- Tudor, M. C., Gomez, L., Della Giovampaola, C., Ugazio, G., & Halope, H. (2024). Leveraging AI to Map SDG Coverage and Uncover Partnerships in Swiss Philanthropy. In *Artificial Intelligence for Sustainability—Innovations in Business and Financial Services*. Palgrave Macmillan: Switzerland. [https://doi.org/10.1007/978-3-031-49979-1\\_9](https://doi.org/10.1007/978-3-031-49979-1_9)
- United Kingdom Government Department of Science, Innovation & Technology (2023, April 19). *Cyber Security Breaches Survey 2023*. GOV.UK. <https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2023/cyber-security-breaches-survey-2023>
- Watson, S. (2022, April 5). Philanthropy Must Never Forget What the Pandemic Taught Us About How to Support Public Health. *The Chronicle of Philanthropy*. <https://www.philanthropy.com/article/philanthropy-must-never-forget-what-the-pandemic-taught-us-about-how-to-support-public-health>
- Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2023). *Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data* (arXiv: 2307.14385). arXiv. <http://arxiv.org/abs/2307.14385>
- Zinsmeister, K. (2016). The Power of Science Philanthropy. *Philanthropy Roundtable*. <https://www.philanthropyroundtable.org/magazine/the-power-of-science-philanthropy/>

# 2

## MEASURING MEANINGFUL CHANGE

### AI-enhanced impact measurement in philanthropy

*Simone Bartalucci, Antonia Muhr, Sina Sauer and Volker Then*

#### 1 Introduction

Measuring and transparently reporting the impact of organizations and projects is becoming increasingly important (Then et al., 2017). This is observable in both profit-oriented companies and the non-profit sector (Fruchterman, 2016; Lapucci, 2021; Münscher & Schober, 2015; Then et al., 2017).

Impacts are defined as changes in stakeholders or target groups as a result of specific activities or interventions and include a range of both positive and negative outcomes. These changes span intentional as well as unintentional consequences. Impact measurement aims to measure the changes in the target population achieved by the activities or interventions (Hehenberger & Buckland, 2023; Then et al., 2017).

Several factors are contributing to the rising enthusiasm for impact measurement: Stakeholders are seeking to validate the effectiveness of their contributions and thus driving the demand for transparency. Notably, investors and donors are increasingly inquisitive about the tangible impact of their societal contributions (Hehenberger & Buckland, 2023; Lapucci, 2021; Schober & Then, 2015; Then et al., 2017). This trend aligns with the broader societal movement toward greater transparency and accessibility of information (Then et al., 2017). The growing prominence of evidence-based policy further underscores this shift, with public funding increasingly emphasizing evidence and impact over input requirements (Hehenberger & Buckland, 2023; Then et al., 2017). In addition, with growing public concern and demand for organizational transparency, there is a growing recognition within organizations of the benefits associated with a systematic approach to impact measurement: Insights from impact measurement allow organizations to adjust internal strategies and enhance organizational learning. It can improve understanding of the environments in which the organizations operate and explore the cause-and-effect relationships between their actions. This improved understanding has the potential to enhance an organization's ability to make informed decisions and allocate resources effectively, thereby improving its services and impact (Hehenberger & Buckland, 2023; Lall, 2019; Lapucci, 2021; Münscher & Schober, 2015; Ní Ógáin et al., 2013; Then et al., 2017).

These benefits of impact measurement are also being discussed in the philanthropic sector. It has been argued that impact measurement and its outcomes play a pivotal role in fostering a



deeper understanding of the context in which philanthropic organizations operate and a more precise definition of the intended beneficiary's needs. This includes unraveling correlations between causes and effects and appreciating the interconnected nature of the issues being addressed. Such increased understanding, it is argued, allows for a more efficient approach to tackling the core and root causes of problems, thereby enhancing the likelihood of achieving positive impact (Bixler et al., 2018; Juech, 2021; Verhulst, 2021). This discussion goes hand in hand with a discernible shift in the sector toward a greater emphasis on strategic planning, outcomes, and impact (measurement) (Anheier & Leat, 2007; Then & Kehl, 2022). Notably, concepts such as evidence-based philanthropy and strategic philanthropy are gaining popularity, guiding the strategic allocation of investments according to proven effectiveness (Brest, 2012; Greenhalgh & Montgomery, 2020). Additionally, the growing and increasingly pressing challenges of the 21st century, coupled with the rather limited resources of philanthropic organizations, make it more urgent than ever to use these resources as effectively and purposefully as possible (Anheier & Leat, 2007; Anheier et al., 2017; Verhulst, 2021).

Although more and more organizations, including those in the philanthropic sector, are joining this movement and are increasingly using data to improve their strategies and evaluate their effectiveness, they face several obstacles: The first challenge for these organizations is to define what data is needed and to access and collect the relevant data. Organizations can often use internal metrics and data to determine the resources invested and actions taken. However, it can be challenging to trace the impact of these activities on the target group, particularly when assessing long-term and sustainable impacts (Fruchterman, 2016). Once data collection is established, organizations need a strategy for analyzing and using the data to inform decision-making and strategy development. Successfully navigating these challenges requires not only the financial resources and technical capabilities to obtain data but also a comprehensive plan that outlines the appropriate use and analysis of the data (Bixler et al., 2018; Fruchterman, 2016). Currently, many philanthropic and social organizations do not have the necessary knowledge and skills to fully utilize available data and lack the required financial resources to conduct impact measurement (Hehenberger & Buckland, 2023; Kassatly, 2018; Verhulst, 2021).

The growing interest in impact measurement has led to a greater professionalization and a wider range of methodological approaches for measuring the social and ecological impact of organizations or interventions (Schober & Then, 2015). This range of methodological approaches can challenge organizations when it comes to selecting the appropriate methods (Kah & Akenroye, 2020). This, in turn, makes it difficult to establish consistent standards and guidelines for measuring impact (OECD, 2021). Nevertheless, several initiatives have attempted to forge common rules and standards: The Impact Management Project (IMP), a coalition of more than 2,000 professionals worldwide, stands out in its efforts to achieve universal agreement on how to measure, manage, and disclose sustainability impacts. Recent developments, such as the UN SDG Impact and UNDP SDG Impact Standards for Enterprises, represent significant progress toward coordinated language and integration across different sectors in support of SDG targets and indicators. However, a standard framework for measuring impact across its multiple social and ecological dimensions is still lacking (Hehenberger & Buckland, 2023). In this discussion, it has been argued that the diversity of (social) organizations, their interventions and goals, and the complexity of social and environmental impacts make the creation of a universal framework difficult (GECES, 2014; Hehenberger & Buckland, 2023; Kah & Akenroye, 2020).

However, as this chapter outlines, even if developing a standardized framework is challenging, it is possible to define a generic process for creating an impact measurement model. We describe a four-step procedure for impact modeling that is applicable to organizations and interventions in

different settings that share the goal of measuring the societal impact created. We will show that digital technologies and artificial intelligence (AI) can be used to identify and select relevant indicators for a given intervention area. We will discuss two options supported by a case study from the field of children’s palliative care. These options demonstrate how digital and AI processes can enhance the indicator selection process within impact modeling, thereby simplifying the process of impact measurement. This could serve as a roadmap for organizations seeking to measure their impact and increase their effectiveness. For philanthropic organizations in particular, impact measurement can be a valuable tool to ensure transparency and demonstrate the effectiveness of their work to key stakeholders such as donors or investors. Additionally, impact measurement results can help philanthropic organizations understand the contexts and issues they aim to address, leading to the creation of appropriate solutions and informed funding decisions.

## 2 Four steps of impact modeling

In general, an impact model is a (mostly graphical) representation of the activities and actions undertaken by an organization or project, and the impacts they generate (Then et al., 2017). The impact modeling process can be categorized into four steps as shown in Figure 2.1, which are further elaborated in this chapter and are based on the deliberations of the Expert Group on Social Economy and Social Enterprises (GECES) on social impact measurement (GECES, 2014). The process begins with the establishment of the normative framework that defines the goals driving the actions. This first step lays the foundation for subsequent steps. In the second step, the identified impact goals are translated into impact chains that outline the resources (inputs) and activities (outputs) needed to achieve the desired effects (outcomes/impacts). The third step moves from theoretical impact modeling to empirical measurement. This includes defining indicators as metrics to measure the realization of intended impacts. The fourth step concludes the process by calculating the overall impact. In general, this modeling exercise must also strike an appropriate balance between representing the complexity of the real world of interventions and reducing it for the sake of analysis in order to identify attributable impacts.

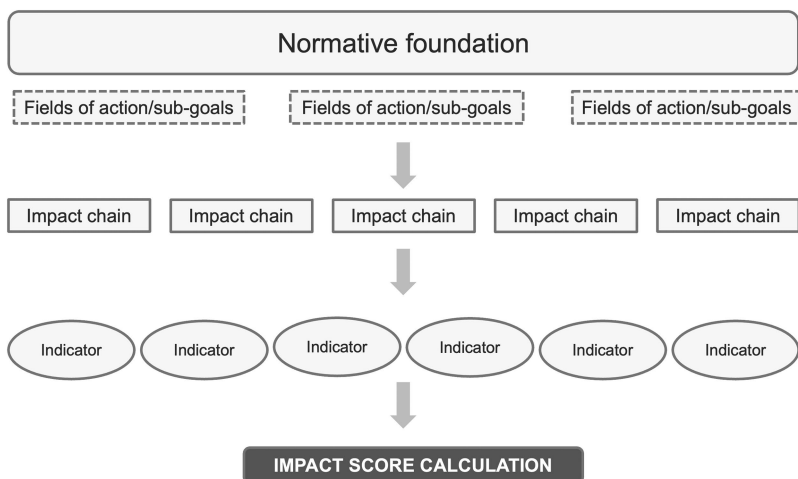


Figure 2.1 Basic impact modeling steps (own work).

### 2.1. Defining the normative basis

The first step in impact modeling requires to address fundamental questions that underlie the entire process. This is the essential task of establishing a normative foundation that serves as a critical touchstone to answer the central questions: What are the primary goals that guide our actions? What outcomes are we working toward? The normative framework serves as a reference point for objectively assessing whether the impact being generated is in line with the organization's goals, or whether it is not contributing to the desired goals or even having a negative effect. This is an essential step because impact data may represent inherently value-free changes, and the normative basis provides the lens through which these observed changes are evaluated as positive or negative (Grünhaus & Rauscher, 2021).

There are various methods for defining this normative basis. These include, for example:

- *Stakeholder-specific approach*: Objectives can be defined based on stakeholder groups affected, where a central goal might be to enhance the quality of life for one specific stakeholder group;
- *Organizational goals*: Normative goals can be derived from the overarching goals of the organization, potentially reflecting the legacy of the founder or the mission of the organization;
- *Established frameworks*: Recognized frameworks such as the Sustainable Development Goals of the United Nations provide an alternative approach. However, these frameworks can often be abstract in nature, which may require translating the goals into measurable sub-goals appropriate for the intervention.

The first step of impact modeling is completed when a realistic set of different (sub-)objectives has been developed. This set serves as a value compass to be able to distinguish which effect is rated as positive or negative in the empirical impact measurement.

### 2.2 Identification of impact chains

The second step is to define impact chains, which show how the organization's activities can contribute to achieving the goals. These cause-and-effect relationships show what resources and activities are needed to achieve the desired effects. Identifying the cause-and-effect relationships can be done by conducting a stakeholder analysis. The goal is to identify the relevant groups for whom an impact is to be achieved (Rauscher et al., 2015). Sub-goals can be formulated for the stakeholder group, specifying the changes required at this level to achieve the intended overarching goals (the normative foundation) in the medium to long term. Once the desired changes have been established for the target group, it is possible to analyze backward what activities contribute to these changes and what resources are required to carry out these activities. In terms of modeling language, the formulation of such impact chains must be understood as hypothetical subject to empirical testing.

The IOOI framework shown in Figure 2.2 can be used to illustrate this relationship, where input is the necessary resources used to perform the specified activity (output). These resources can be economic, such as capital, or social, such as time or labor. The category of outputs encompasses the actions or services that are implemented or delivered and result in a change or impact on the target group at the outcome level. Moreover, the IOOI model distinguishes between outcome and impact, by estimating the deadweight to account for changes that would have occurred even in the absence of inputs and outputs. This quantity is subtracted from the result to reveal the genuine, incremental value added by the intervention (Rauscher et al., 2015; Then et al., 2017).

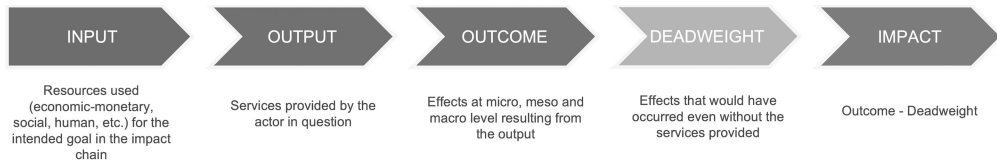


Figure 2.2 Impact chain structure according to the IOOI model (own work, based on Then et al., 2017. Social Return on Investment Analysis: Measuring the Impact of Social Investment, Springer International Publishing, reproduced with permission from SNCSC).

However, measuring environmental and social impacts within the complex social reality presents challenges in determining the deadweight. The complexity stems from the difficulty of determining what would have happened if the activity had not taken place. Several methodologies exist for estimating deadweight. One such approach is to conduct surveys that allow for a comparative analysis of conditions before and after an intervention. This retrospective study provides valuable insights into the changes attributable to the intervention. Moreover, empirical data can be used to calculate estimates of deadweight, allowing for a quantitative analysis of the counterfactual scenario (Grünhaus & Rauscher, 2021; Rauscher et al., 2015; Then et al., 2017).

### 2.3 Development of proxies (indicators)

Once the potential effects of the intervention have been theoretically modeled, it is necessary to examine the extent to which these effects are manifested in the target group (Rauscher et al., 2015). For this purpose, various indicators are identified as proxies to ascertain the degree to which the desired effect has been realized. Often, the effects cannot be measured directly as they involve intricate societal issues, such as an improvement in the quality of life. To make the theoretical model practically operational, a set of indicators is selected or defined that represents the most consistent approach to the social phenomenon of interest (Diaz-Bone, 2022).

There are different ways to select appropriate indicators. In this chapter, we present two experimental digital approaches that simplify and improve the process of indicator determination. These techniques speed up the selection and incorporate modern technology to bring a contemporary perspective to indicator selection.

- 1 *Matching prototype and data indicators*: The first approach involves a two-step process to identify indicators for the impact model. The first step is to identify theoretical indicators, based on the impact chains, called prototype indicators. These prototype indicators describe how the impact should be measured using realistically implementable indicators to assess whether (and how) a particular impact has been achieved. To formulate prototype indicators, relevant studies, literature, existing frameworks, and their respective indicators are reviewed, and suitable aspects are extracted. Once the indicators have been selected, a description of each indicator needs to be formulated. In a second step, a natural language processing (NLP) algorithm is used to search for data indicators whose descriptions match those of the prototype indicators. The data indicator with the highest degree of semantic similarity with the prototype indicator is selected as the data indicator for the impact model. The advantage of this approach is that the impact model can be quickly and continuously adapted to the set of available empirical indicators, which can be extensive and rapidly changing. The prototype indicators need to be defined only once and can then be periodically checked against sets of available data indicators.

2 *Identifying commonly used items*: The second approach involves analyzing existing impact measurement tools through similarity analysis to identify commonly used items. Based on this analysis, an NLP algorithm is incorporated into a tool to extract the most relevant indicators from existing methods and combine them into a consolidated tool. This method is particularly suitable for projects with multiple established tools in the field, where the aim is to identify significant dimensions and indicators. It is particularly useful in the initial stages of research when a thorough exploration of the existing field knowledge is required. The advantage of this approach lies in its ability to capitalize on extensive existing knowledge, synthesizing information while preserving the core consensus dimensions and minimizing noise.

Section 3 of the chapter presents the methodology of both approaches in detail.

#### **2.4 Empirical testing of the model and determination of impact**

Once the indicators have been identified, the theoretical impact model should be tested empirically to ensure its feasibility and validity. Data must be collected accordingly, with the choice of method depending on the selected indicators and data availability. Several methods are available for interpreting the data collected, such as the Social Return on Investment (SROI) analysis or the calculation of an impact score. The first approach compares the resources used with the resulting impact. Both values are monetized, allowing for the calculation of the societal return based on the difference between the resources utilized and the impact generated (Then et al., 2017). This methodology provides a valuable tool for assessing the relationship between resource use and resulting impact. However, monetary valuation may not be applicable or beneficial for all impact measurements, especially when impact is assessed in varying units. Alternatively, the second normalization approach addresses these challenges by calculating an impact score for each project, intervention, or organization. This method aggregates individual indicator scores, allowing flexibility in assigning weights to different indicators as needed. This approach is particularly useful for comparing the impact of different projects, interventions, or organizations without monetizing the measured impact.

### **3 AI support for the indicator selection**

Having outlined the generic impact modeling process, it is now possible to examine some of its practical implications and explore how artificial intelligence (AI) tools can facilitate its implementation. Before examining specific application cases, it is worthwhile to consider the design choices of the impact measurement model as they may affect the indicator selection process. Based on experience in impact modeling, it is recommended to pay attention to two macroelements that may motivate the choice of automation tools.

The first aspect to consider is the scope of the model. The broader the scope, the more important it is to examine a variety of impact areas in detail. Once the model is tuned to evaluate the impact of a specific, predefined intervention, its scope automatically narrows, resulting in a more workable array of indicators. Certain impact modeling requirements may necessitate the creation of a model with greater universality, where the focus of impact conceptualization is at the macro level. This results in a significant expansion of the range of impact areas.

The second factor to consider when designing the impact measurement model is the accessibility of existing indicator frameworks. It should be noted that the use of such indicators can be beneficial for two main reasons. The first reason is that the inclusion of validated indicators, as determined by industry professionals or previous research, can enhance the credibility of the model and, consequently, its acceptance and validation by stakeholders. The second reason is

that it streamlines data collection and facilitates comparison. Incorporating indicators that have already been measured or can be measured without additional effort into the impact measurement model significantly reduces the cost of implementing the model. However, it is often the case that these indicators were originally intended for purposes other than social impact measurement. Examples could include performance indicators used by the organization to measure their impact, indicators collected by third parties for risk assessment, or indicators defined by external parties, such as public institutions, to monitor specific areas. Selecting the appropriate indicators for this design choice requires careful consideration, especially given the many pre-existing frameworks available. In certain sectors, the search process may uncover an array of frameworks with a large number of indicators. Dealing with such profusion is a demanding task for humans. As the case study shows, this is especially true in the specialized and control-intensive field of healthcare.

These conditions may facilitate the integration of artificial intelligence to automate impact modeling intricacies. In the following sections, we present two different applications of AI capabilities that have been empirically tested in actual cases of social impact measurement projects. Both approaches implemented the same artificial intelligence tool during the model development phase for indicator identification in impact measurement. It is worth noting that the approaches differ primarily in their methodological settings. The first approach adopts a deductive position, while the second uses the openness characteristic of inductive research.

The deductive approach involves identifying indicators by starting with a pre-existing conceptualization of the impact area. In this method, the AI tool selects indicators by associating them with the defined components of the impact area. Essentially, the tool moves from the general (impact area conceptualization) to the particular (data indicators). The inductive approach is a method in which the conceptualization of the impact area is derived from the AI tool itself through the recognition of patterns in existing indicators. In this case, it moves from the particular (data indicators) to the general (impact area conceptualization).

As anticipated, both approaches use the same tool, so before presenting them in more detail, it is appropriate to make a brief mention of the artificial intelligence tool that was used in the empirical testing.

The tool is based on an algorithm derived from NLP. Essentially, the tool can compare two items and generate a score based on their semantic similarity. If the two items are identical, the tool assigns a score of one; if the items are very different, the score approaches minus one. The tool produces a square matrix showing the scores for each pair of items and a list of pairs with the highest similarity scores. Users can easily filter the list of pairs as needed. The tool allows users to set a defined number of pairs with maximum scores or a minimum limit of acceptable similarity scores by adding a few lines of code.

This chapter does not explore the tool's technical specifications in depth. It is acknowledged that more advanced NLP-based tools may emerge over time, improving the accuracy of the results. The main goal of the discussion is to provide interesting use cases to stimulate interest and exploration; more powerful tools can only improve the outcome of these applications.

### ***3.1 First approach: matching prototype and data indicators***

The first approach is what has been called deductive and has been empirically tested in a project of Fondazione AIS.

As described above, the first step is to define prototype indicators (PIs), based on a review of scientific literature and official reports from institutions. The purpose of the PIs is to guide the search for the indicators that will be introduced to calculate the impact score. This methodology is useful when the project does not intend to use data from primary sources, that is, from

measurements made directly by the responsible organization, but rather to use data already collected by third parties. This approach reveals the need for an AI-based tool to support the evaluation of existing data points, under two conditions. First, an abundance of indicators: if the model has an extremely broad scope, both in terms of the object of measurement (number of impact dimensions that are covered) and in terms of the unit of measurement (number of agents whose activity is measured), this can lead to the compilation of a list containing hundreds of PIs. The second condition concerns the indicators that already exist in the sector. It may be that the availability of external indicators is very heterogeneous and very numerous. Given the different purposes and methodologies that led to the development of these frameworks, it is clear that only a minority fraction of the indicators meets the requirements for measuring impact described above. However, the number may be so large that a selection based solely on human work is impractical within reasonable time and cost constraints.

Hence, the application of the similarity tool to the list of indicators can prove to be instrumental. This application facilitates the identification of a substantial portion of indicators and plays a key role in advancing the model's development. The remarkable result is the successful identification of a comprehensive set of potential indicators to be associated with PIs. While the final decision rests with human judgment, the existence of a collection of highly similar indicators for the defined measurement dimensions serves as a valuable resource, sparing individuals from the tedious task of reviewing all indicators across diverse frameworks. This not only significantly reduces development time but also results in a more efficient use of human resources. An additional advantage of this methodology is the rapid and continuous adaptability of the impact model to the collection of new accessible indicators. PIs are defined on a one-time basis, after which they can be subjected to automatic verification whenever the set of available indicators is modified.

Figure 2.3 shows a visual representation of the logical process delineated, highlighting the specific points at which the NLP-based similarity tool is employed.

### 3.2 Second approach: identification of commonly used items

The second approach is based on inductive logic and was tested in the context of the project presented as a case study in the fourth section of this chapter.

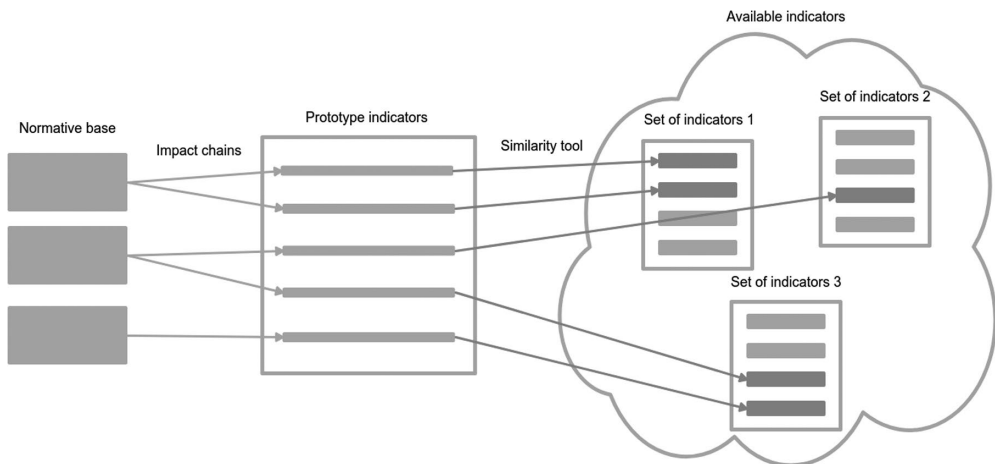


Figure 2.3 Logical scheme of the first approach (own work).

This approach follows the same general methodological framework outlined above. After identifying the normative foundation, impact chains are traced to define the effects of the individual activities to be measured. Although this conceptual work makes it possible to establish logical sequences describing impact generation in each area, the choice of indicators may require further research.

It may happen that an impact dimension, as in the case of quality of life, is characterized by intangibility and high dependence on the state and perception of the target stakeholders. In this setting, it can be difficult to identify proxy indicators, and an in-depth analysis of the scientific literature in the relevant sector may be necessary.

At this point, the second condition described earlier may manifest itself: the presence of a significant number of existing frameworks in the sector. In contrast to the previous approach, the similarity tool is not designed to map existing indicators to model indicators. Instead, it is used to support the logical development of the model by defining the most recurrent impact sub-dimensions in existing frameworks.

Despite the possible evidence in the literature to the application of each framework to the relevant sector, the frameworks may differ in terms of specificity, object of measurement, and conception of the impact area. As discussed above, impact frameworks may have been developed in very different contexts, and it is possible that none of them has been specifically designed to measure social impact.

Given the high degree of heterogeneity, a prioritization strategy may be necessary to exclude certain frameworks and prioritize those that best fit the objectives of the model. This strategy can be implemented by defining a set of criteria and classifying frameworks based on the scores associated with these criteria. It is then possible to identify the top-performing frameworks for each stakeholder category and compare them through the similarity tool.

In a relatively short time, the tool can provide insight into the most recurring impact sub-dimensions for each of the main stakeholders. Figure 2.4 illustrates the logical process of this approach, highlighting the role of the NLP-based similarity tool.

While this approach has proven useful in developing the logic model, it has certain limitations and should be used with caution.

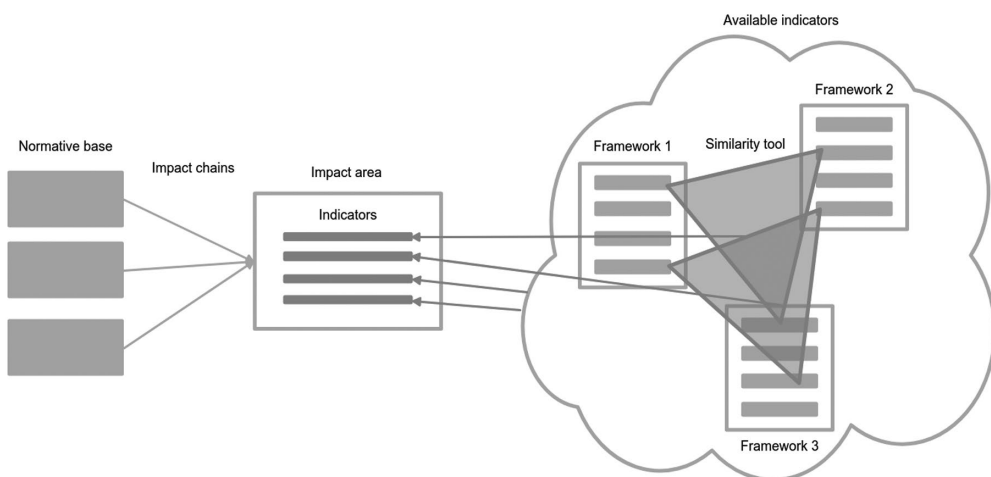


Figure 2.4 Logical scheme of the second approach (own work).



Although this process can broaden the understanding of an impact area by identifying the most recurrent dimensions, it should be noted that dimensions that are not recurrent or more innovative in the frameworks should not be dismissed as irrelevant. As seen in the first approach, the similarity tool only identifies areas of convergence and fails to provide potential insights into areas of divergence. However, it remains methodologically valid to assert that the identified elements are dimensions that are recognized as material in stakeholder impact assessment.

Compared to the previous approach, the output of this process is more sensitive to the quality of the input frameworks. Therefore, a robust pre-screening process is essential to ensure that only appropriate frameworks are selected.

It is important to emphasize that this approach is best suited for the exploratory stages of research. If a satisfactory knowledge base of the impact area already exists, the first approach is preferable due to its more rigorous methodology.

Lastly, it is recommended to validate the results with experts in the impact area. This external validation ensures the reliability and credibility of the outcomes and adds an essential layer of rigor to the research process.

#### **4 Case study: impact modeling in children's palliative care**

Having presented two possible applications of artificial intelligence for impact measurement, it is possible to move further down into practice by expanding on the details of one of these cases to illustrate the practical relevance and usefulness of the AI-supported processes. As a corresponding case study, this section presents elements of the children's palliative care (CPC) Impact Modeling Project carried out by Fondazione AIS.

The primary goal of the project was to create a model to assess the impact of CPC through the lens of private philanthropic investors. The results were supposed to be in line with universal characteristics applicable to different fields of endeavor, allowing for a comparative analysis of investments against alternative philanthropic uses of funds. The model assesses the extent to which a generic CPC investment has achieved its intended results, considering multi-stakeholder impacts on society. The model, which serves as a versatile tool for the entire CPC sector, is also intended to be concise in its handling of complexity, using a minimal set of indicators to measure impact while ensuring validity and reliability.

The first step was to establish a normative base, driven by a clear goal: to improve the quality of life (and quality of death) of children and their caregivers. From this overarching goal, three sub-goals were defined: (1) improving quality of life, (2) improving healthcare functioning, and (3) promoting social solidarity and community inclusion.

The next step was to identify impact chains. For the sake of brevity, we refrain from a detailed elaboration of the impact chains. We simply provide a sample set of impact chains (see Table 2.1) and a diagram that provides an overview of the different logical paths and sequence of effects (see Figure 2.5).

The next phase was to identify indicators that could measure the identified impact areas. The use of existing frameworks and indicators was preferred, based on the advantages discussed in the previous section. This decision was motivated by the aim to establish legitimacy within the CPC professional sector and integrate the data gathering phase with tools already in use for various purposes in the field. This approach received strong support from the expert field consultants involved in the conceptual development of the model.

The case is a demonstration of the application of the second of the approaches discussed. In fact, the AI tool was used to identify the appropriate indicators measuring variations in a specific impact area: quality of life (QoL) for children in need of CPC and their families as an outcome of

Table 2.1 Example of impact chains of a generic CPC investment (own work)

<i>Strategy</i>	<i>Sub-strategy</i>	<i>Input</i>	<i>Output</i>	<i>Outcome</i>	<i>Main beneficiary</i>
Develop CPC service provision units	Services	Capital	Implementing new care services that enhance accessibility	Improved access to CPC services (outreach)	Children and families
Develop CPC service provision units	Infrastructure	Capital	Investing in the expansion of CPC infrastructure	Improved access to CPC services (outreach)	Children and families
Create a supportive environment for CPC	Standard, Training, and Education	Capital	Supporting the professional education and specialization of current and future healthcare practitioners	Improved professionalization of the sector (increased health system functioning)	CPC and healthcare professionals
Create a supportive environment for CPC	Shifting Public Conversation	Capital	Initiating and supporting societal discourse about death	Improved recognition of the value of death	Society

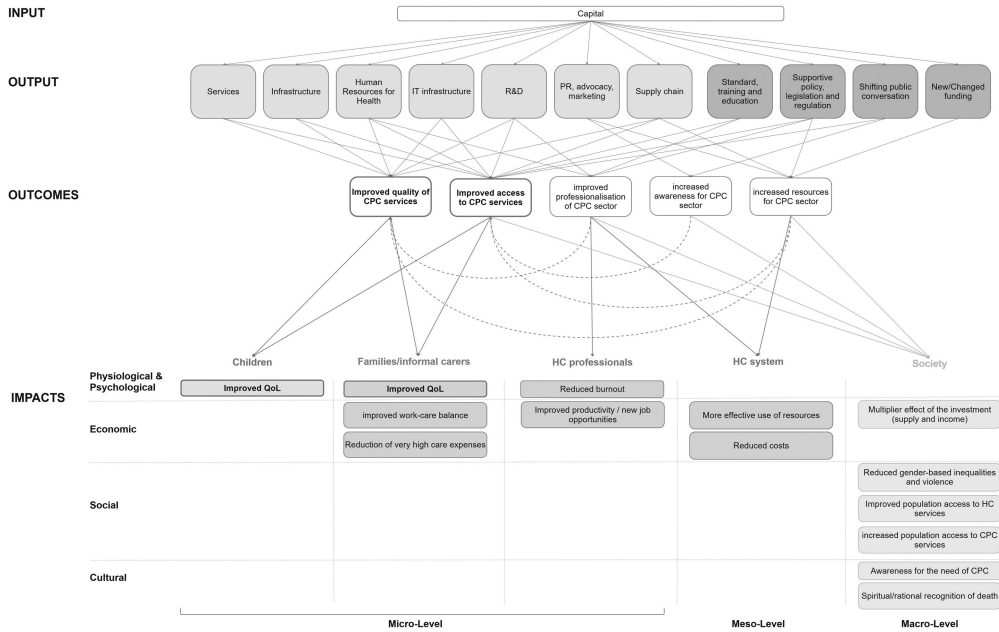


Figure 2.5 Synthetic impact chain flows and correlations according to the IOOI model (own work).

the intervention. Since QoL serves as one of the primary normative goals for this philanthropic investment, its measurement is a central component of the impact model.

Recognizing that QoL is a highly intangible impact area, influenced by multiple factors and conceptualized differently across scientific disciplines, a comprehensive analysis of existing knowledge was undertaken. The aim of this analysis was to identify the sub-dimensions that make up QoL for each category of individual stakeholders, in line with the normative base established earlier.

The process began with a collection of frameworks discussed in the literature. These frameworks were reviewed according to a set of criteria in order to prioritize the use of those that were more in line with the conceptual model. The similarity tool was then applied to the best frameworks to identify the most relevant sub-dimensions within the impact area. Figure 2.6 illustrates the process followed and the results of each phase.

After an analysis of the available literature, 33 scientific papers and official reports on CPC were considered relevant for this purpose. Within these studies, a total of 32 frameworks<sup>1</sup> emerged, each of which included a wide range of indicators, ranging from 5 to 50 per framework. This extensive compilation of frameworks provided valuable insights for the development of the model, leading to two critical observations: (1) the lack of a single (or a few) framework with a majority consensus within the scientific and professional communities, and (2) the urgent need for the assistance of automation tools to comprehensively analyze the various indicators included in the frameworks. Both findings supported the adoption of an AI-based approach to identify recurring patterns among the multitude of identified indicators.

To streamline the focus on the frameworks that contained information relevant to the prior conceptual development, an attempt was made to extract the essential details of these frameworks. Drawing from experience in impact measurement and drawing on the knowledge of expert advisors, a set of 16 criteria was developed (see Table 2.2).

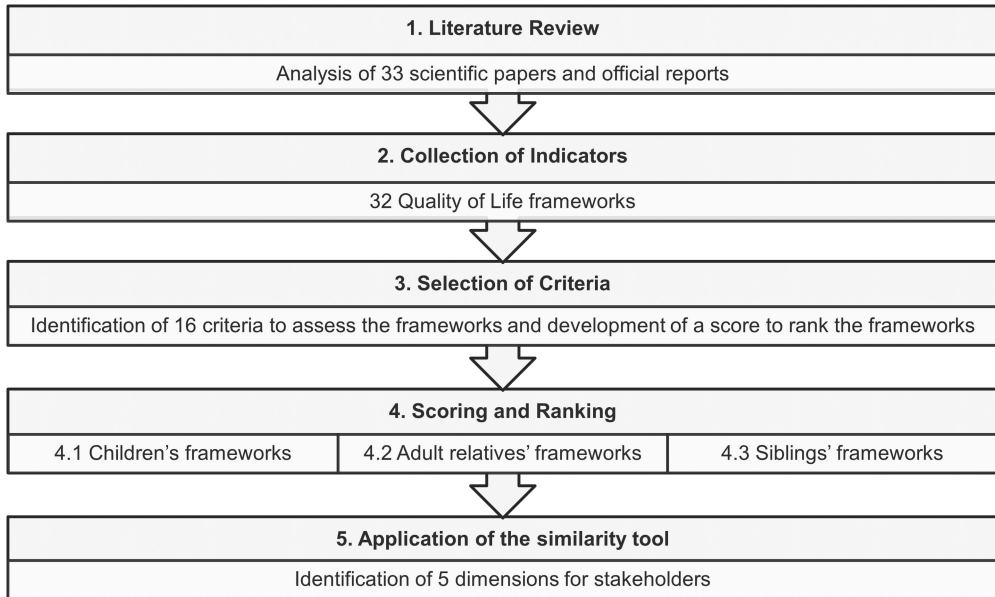


Figure 2.6 Process flow for identifying and selecting indicators to measure the quality of life of CPC stakeholders (own work).

Of the identified criteria, nine (highlighted in Table 2.2) proved to be instrumental in guiding the selection of the most appropriate framework. By systematically gathering information for each attribute of each framework based on the identified criteria, scores were assigned to each framework based on these nine criteria. This scoring system allowed for differentiation among the frameworks, tailored to the specific needs of each stakeholder category. The outcome of this phase was three distinct rankings of frameworks: one for children in need, another for adult relatives, and a third for siblings. Table 2.3 shows the three rankings.

Even with the narrow focus, the number of indicators (several hundred in total) was still significant enough to require the use of an AI-based approach. The similarity tool was applied using the second approach described in the previous section, as shown in Figure 2.7.

The main purpose of this process was to identify a list of the most recurring items for each target stakeholder. Specifically, the top seven frameworks in each of the three previously constructed rankings were considered. The process began with the aggregation of all questionnaires from different frameworks into a single CSV file. This consolidated file was then fed into the similarity tool, which produced a square matrix. The matrix, with the items in both rows and columns, represented the similarity between each pair of items.

A vector was then constructed that encapsulated the average values of each row in the matrix, providing a concise representation of the data.

The final step was to extract the five highest average values from the vector and identify the corresponding items. This allowed the most significant items to be identified based on their mean values.

In essence, this process facilitated a thorough analysis of the framework sets and provided valuable insight into the similarities between items across different frameworks. Table 2.4 shows the results of the process.<sup>2</sup>

Table 2.2 Criteria used to analyze CPC stakeholder quality of life indicators (own work)

<i>ID</i>	<i>Indicator criteria</i>	<i>Explanation</i>
1	Output vs outcome	Object of the measure captured by the indicator considering the impact chain model.
2	Stakeholder specificity	Identifying those indicators which are specifically intended for measuring QoL of pediatric population or their relatives.
3	Palliative care specificity	Distinguishing between general QoL measures and indicators which are specifically developed to assess palliative care outcomes.
4	Contextual dimension	Indicator's ability to capture values to assess the outcome in the different impact dimensions involved (physical, psychological, social, economic, ecological).
5	Respondent	Stakeholder that is involved in the data collection process.
6	Timeframe of the data collection	Period during which data is collected in relation to the palliative care process.
7	Target stakeholder	Stakeholder whose impact is measured by the indicator.
8	Monetizability	Distinguishing indicators to which it is possible to directly assign a monetary value from those to which it is not possible or require further processing.
9	Quantitative vs qualitative	Distinguishing between indicators that assess quantitative information and those that assess qualitative information.
10	Temporal dimension	Indicator's ability to capture long-term effects.
11	Level of complexity	Length (number of items) of the indicator as a proxy for its complexity. More items mean more time-consuming and more complex indicators to collect and interpret.
12	Granularity	Frequency with which the indicator is measured.
13	License provider	Type of organization that manages the diffusion of the indicator.
14	Empirically tested	Distinguishing between indicators that have undergone validation in real-life conditions and those still in preliminary stages.
15	Derivation	Indicators whose calculation is derived from other measures (e.g., QALY is based on other QoL measures like EQ-5D).
16	Structural dimension	Structural dimension influenced by the impact (e.g., physical, economical, social).

Table 2.3 Ranking of CPC’s quality of life measurement frameworks based on analysis criteria scores by stakeholder (own work)

<i>Children’s frameworks ranking</i>		<i>Adult relatives’ frameworks ranking</i>		<i>Siblings’ frameworks ranking</i>		
<i>Name</i>	<i>Score</i>	<i>Name</i>	<i>Score</i>	<i>Name</i>	<i>Score</i>	
1	MQOL-E	4.83	QOLLI-F	5.33	EQ-5D-Y	4.33
2	EQ-5D-Y	4.50	CQLI	4.75	16D	4.00
3	CHU9D	4.17	CES	3.67	17D	4.00
4	16D	4.08	ICECAP-A	3.50	CHU9D	4.00
5	17D	4.08	EQ-5D	3.25	PedsQL 4.0	3.67
6	ICECAP-SCM	4.00	ZBI	3.25	ICECAP-CPM	3.00
7	PedsQL 4.0	3.75	ICECAP-CPM	3.00	WHOQOL-BREF	3.00
8	CHQ-PF28	3.25	WHOQOL-BREF	2.75	SF-6D	2.67
9	KIDSCREEN-27	3.25	CES	2.75	HUI 2	2.67
10	PICU-QODD-20	3.00	HUI 2	2.58	HUI 3	2.67
11	WHOQOL-BREF	3.00	HUI 3	2.58	QWB	2.67
12	HUI 2	2.83	SF-6D	2.50	QWB-SA	2.67
13	HUI 3	2.83	QWB	2.50		
14	SF-6D	2.75	QWB-SA	2.50		
15	QWB	2.75	PICU-QODD-20	2.00		
16	QWB-SA	2.75	PaPEQu	1.00		
17	GDI-P	2.75				
18	CES	2.50				
19	PaPEQu	2.25				

The identification of these dimensions not only facilitated a comprehensive understanding of the frameworks but also enriched the insights. This process was instrumental in advancing the research methodology and was used in two different ways:

- 1 *Questionnaire development:* The identified dimensions were used to construct questionnaires for target stakeholders that encompassed items measuring each relevant sub-dimension within the impact area;
- 2 *Framework comparison:* Additionally, a comparative analysis was conducted with existing frameworks, assessing the average similarity between these frameworks and our identified list. This comparative evaluation was again performed using the similarity tool, which provides a quantitative measure of alignment.<sup>3</sup>

Both sets of results served as background material for in-depth discussions during a workshop with experts in the field. During the workshop, the results were not simply integrated into the existing model. On the contrary, their importance was to provide an outside perspective rooted in the synthesis of the work of many other experts. The findings were not treated as isolated additions to the model but were integrated through a collaborative process that enriched the model with diverse perspectives and expert insights. This iterative approach, incorporating both quantitative and qualitative assessments, strengthened the robustness of the research work.

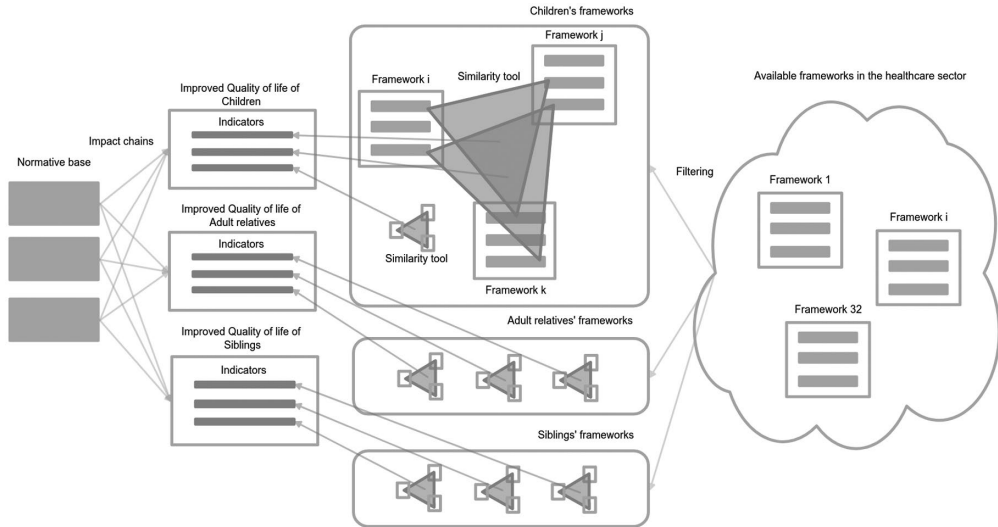


Figure 2.7 Logical scheme for utilizing a similarity tool in the CPC case study (own work).

Table 2.4 Dimensions of quality of life resulting from an analysis of existing frameworks using the similarity tool grouped by stakeholder (own work)

<i>Children</i>	<i>Families/informal carers</i>	<i>Siblings</i>
Physical wellbeing	Distress	Vitality
Control over life	Fulfillment	General health
Physical environment	Care-life balance	Safety
Melancholy	Social life	Daily activities
Vitality	Burden	Self-satisfaction

## 5 Conclusion

The systematic measurement of societal impact is becoming increasingly important for a wide range of organizations. This chapter focuses on the role of AI in the indicator selection process in impact modeling, highlighting two different AI-enabled approaches to indicator selection and how these tools can support the development of robust impact measurement models.

The first, a deductive approach, involves defining prototype indicators based on impact objectives and using the AI tool to efficiently identify similar indicators within extensive frameworks. Although it was successful in the empirical testing, this approach requires careful consideration of contextual factors, human validation, and potential limitations in identifying conceptual gaps. The second approach is inductive and uses AI to support the logical development of the model by identifying recurring impact sub-dimensions in existing frameworks. This technique is particularly suited to exploratory research phases, but it requires robust pre-screening of input frameworks and external expert validation. Another limitation is that the two approaches have only been applied and tested in two individual projects and would require further validation and empirical application to increase the robustness.

Despite its limitations, this chapter explores the practical implications of integrating AI into the indicator selection process in impact modeling. It has great potential in situations where there is an abundance of data sources that are unstructured, inconsistent, or incomplete. This technology streamlines processes, enhances efficiency, and expands opportunities to integrate existing work. Additionally, it can be systematically valuable in cases where there are multiple competing measurement approaches for the same area of intervention, and their respective scopes need to be compared, as shown in the reported case study.

By presenting these two applications in the context of impact measurement, we aim to stimulate discussion about the use of AI in philanthropy. We encourage the further exploration of AI tools and techniques for the indicator selection process, and to uncover the potential of AI for the other steps in the impact measurement process, such as defining the normative basis or calculating impact.

## APPENDICES

### Appendix A: List of Quality of Life and Outcome Frameworks

*Table A* List of quality of life and outcome frameworks

ID	Name	Full name	Source
1	QALY	Quality-Adjusted Life Years	<a href="https://en.wikipedia.org/wiki/Quality-adjusted_life_year">https://en.wikipedia.org/wiki/Quality-adjusted_life_year</a>
2	EQ-5D	EuroQol Five-Dimension	<a href="https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-3l/">https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-3l/</a>
3	SF-6D	Short-Form Six-Dimension	<a href="https://www.qualitymetric.com/health-surveys/sf-6d-health-utility-2/">https://www.qualitymetric.com/health-surveys/sf-6d-health-utility-2/</a>
4	HUI 2	Health Utilities Index 2	<a href="http://www.healthutilities.com/hui2.htm">http://www.healthutilities.com/hui2.htm</a>
5	HUI 3	Health Utilities Index 3	<a href="http://www.healthutilities.com/hui3.htm">http://www.healthutilities.com/hui3.htm</a>
6	PedsQL 4.0	Paediatric Quality of Life Inventory	<a href="https://www.pedsqol.org/about_pedsqol.html">https://www.pedsqol.org/about_pedsqol.html</a>
7	QWB	Quality of Well-Being Scale	<a href="https://en.wikipedia.org/wiki/Quality_of_well-being_scale">https://en.wikipedia.org/wiki/Quality_of_well-being_scale</a>
8	QWB-SA	Quality of Well-Being Scale – Self-Administered	<a href="https://hoap.ucsd.edu/qwb-info/QWB-Manual.pdf">https://hoap.ucsd.edu/qwb-info/QWB-Manual.pdf</a>
9	16D	16D	<a href="http://www.15d-instrument.net/16d-and-17d/16d/">http://www.15d-instrument.net/16d-and-17d/16d/</a>
10	17D	17D	<a href="http://www.15d-instrument.net/16d-and-17d/17d/">http://www.15d-instrument.net/16d-and-17d/17d/</a>
11	EQ-5D-Y	EuroQol Five-Dimension Youth	<a href="https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-y-3l/">https://euroqol.org/information-and-support/euroqol-instruments/eq-5d-y-3l/</a>
12	CHU9D	Child Health Utility 9D	<a href="https://licensing.sheffield.ac.uk/product/CHU-9D">https://licensing.sheffield.ac.uk/product/CHU-9D</a>
13	PaLY	Palliative Care Yardstick	<a href="https://www.sciencedirect.com/science/article/pii/S0885392410010614">https://www.sciencedirect.com/science/article/pii/S0885392410010614</a>
14	VIP	Valuation Index Palliative	<a href="https://www.sciencedirect.com/science/article/pii/S0885392410010614">https://www.sciencedirect.com/science/article/pii/S0885392410010614</a>
15	ICECAP-CYP	ICEpop CAPability measure for Children and Young People	<a href="https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-cyp/">https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-cyp/</a>

(Continued)



Table A (Continued)

ID	Name	Full name	Source
16	ICECAP-SCM	ICECAP Supportive Care Measure	<a href="https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-scm/">https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-scm/</a>
17	ICECAP-A	ICECAP Adults	<a href="https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-a/">https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-a/</a>
18	ICECAP-CPM	ICECAP Close Person Measure	<a href="https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-cpm/">https://www.bristol.ac.uk/population-health-sciences/projects/icecap/icecap-cpm/</a>
19	GDI-P	Good Death Inventory – Paediatrics	<a href="https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft79367-000">https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft79367-000</a>
20	PICU-QODD-20	Paediatric Intensive Care Unit – Quality of Dying and Death 20	<a href="https://pubmed.ncbi.nlm.nih.gov/24878067/">https://pubmed.ncbi.nlm.nih.gov/24878067/</a>
21	PaPEQu	Parental PELICAN Questionnaire	<a href="https://pubmed.ncbi.nlm.nih.gov/26265326/">https://pubmed.ncbi.nlm.nih.gov/26265326/</a>
22	QCPCI	Quality of Children’s Palliative Care Instrument	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6397460/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6397460/</a>
23	CHQ-PF28	Child Health Questionnaire – Parent Form 28	<a href="https://www.qualitymetric.com/health-surveys/child-health-questionnaire-chq/">https://www.qualitymetric.com/health-surveys/child-health-questionnaire-chq/</a>
24	KIDSCREEN-27	KIDSCREEN-27	<a href="https://www.kidscreen.org/english/questionnaires/kidscreen-52/">https://www.kidscreen.org/english/questionnaires/kidscreen-52/</a>
25	C-POS	Children’s Palliative care Outcome Scale	<a href="https://www.kcl.ac.uk/research/c-pos">https://www.kcl.ac.uk/research/c-pos</a>
26	QOLLI-F	Quality of Life in Life Threatening Illness – Family Carer Version	<a href="https://www.dgpalliativmedizin.de/images/QOLLI-F_v2_Family_Caregivers_english.pdf">https://www.dgpalliativmedizin.de/images/QOLLI-F_v2_Family_Caregivers_english.pdf</a>
27	CQLI	Caregiver QOL Index	<a href="https://eprovide.mapi-trust.org/instruments/caregiver-quality-of-life-questionnaire-physical-emotional">https://eprovide.mapi-trust.org/instruments/caregiver-quality-of-life-questionnaire-physical-emotional</a>
28	WHOQOL-BREF	World Health Organization Quality Of Life – Brief Version	<a href="https://www.who.int/tools/whoqol/whoqol-bref">https://www.who.int/tools/whoqol/whoqol-bref</a>
29	ZBI	Zarit Burden Inventory	<a href="https://eprovide.mapi-trust.org/instruments/zarit-burden-interview">https://eprovide.mapi-trust.org/instruments/zarit-burden-interview</a>
30	CES	Care Evaluation Scale	<a href="https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft24004-000">https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft24004-000</a>
31	MQOL-E	McGill Quality of Life Questionnaire – Expanded	<a href="https://pubmed.ncbi.nlm.nih.gov/31672131/">https://pubmed.ncbi.nlm.nih.gov/31672131/</a>
32	CES	Carer Experience scale	<a href="https://www.bristol.ac.uk/population-health-sciences/projects/icecap/ces/">https://www.bristol.ac.uk/population-health-sciences/projects/icecap/ces/</a>

## APPENDIX B

### Most Recurrent Dimensions

*Table B* Framework dimensions per stakeholder

<i>Stakeholder</i>	<i>Original framework</i>	<i>Item</i>	<i>Label</i>
Children	MQOL-E	Over the past two days (48 hours), I felt: physically terrible/physically well	Physical wellbeing
	MQOL-E	Over the past two days (48 hours), I felt that the amount of control I had over my life was: not a problem/a huge problem	Control over life
	MQOL-E	Over the past two days (48 hours), my physical surroundings met my needs: not at all/completely	Physical environment
	EQ-5D-Y 16D	FEELING WORRIED, SAD, OR UNHAPPY I feel healthy and energetic/I feel extremely weary, tired, or weak	Melancholy Vitality
Informal carers	QOLLTI-F	Over the past two days (48 hours) the condition of the family member/ friend I'm caring for was distressing to me: not often/always	Distress
	QOLLTI-F	Over the past two days (48 hours) being able to provide care or company for the family member/friend I'm caring for made me feel good: rarely or never/always	Fulfillment
	ZBI	Do you feel stressed between caring for your relative and trying to meet other responsibilities for your family or work?	Care-life balance
	ZBI	Do you feel that your social life has suffered because you are caring for your relative?	Social life
	ZBI	Overall, how burdened do you feel in caring for your relative?	Caregiver burden
Siblings	16D	I feel healthy and energetic/I feel extremely weary, tired, or weak	Vitality
	WHOQOL-BREF	How satisfied are you with your health?	General health
	WHOQOL-BREF	How safe do you feel in your daily life?	Safety
	WHOQOL-BREF	How satisfied are you with your ability to perform your daily living activities?	Daily activities
	WHOQOL-BREF	How satisfied are you with yourself?	Self-satisfaction

APPENDIX C

**Rankings of Frameworks Similarity to the List of Most Recurring Dimensions**

*Table C.1 Children’s framework similarity*

<i>Children</i>		
<i>Framework</i>	<i>Average score</i>	<i>Score standard deviation</i>
MQOL-E	0.42	0.15
ICECAP-SCM	0.33	0.15
EQ-5D-Y	0.32	0.18
16D	0.28	0.14
CHU9D	0.27	0.15
PEDSQL 4.0	0.24	0.14
17D	0.21	0.14

*Table C.2 Informal carers’ framework similarity*

<i>Informal Carers</i>		
<i>Framework</i>	<i>Average score</i>	<i>Score standard deviation</i>
ZBI	0.65	0.13
QOLLI-F	0.44	0.18
CES	0.41	0.17
ICACAP-CPM	0.37	0.16
CQLI	0.33	0.16
EQ-5D	0.30	0.19
ICECAP-A	0.30	0.22

*Table C.3 Siblings’ framework similarity*

<i>Siblings</i>		
<i>Framework</i>	<i>Average score</i>	<i>Score standard deviation</i>
WHOQOL-BREF	0.41	0.16
EQ5DY	0.31	0.17
ICECAP-CPM	0.29	0.17
CHU9D	0.26	0.16
16D	0.25	0.14
PEDSQL 4.0	0.22	0.14
17D	0.22	0.14

**Notes**

- 1 For more information, see Appendix A.
- 2 For more details on the items identified by the tool, see Appendix B.
- 3 The results of this analysis are provided in Appendix C.

## References

- Anheier, H., Förster, S., Mangold, J., & Striebing, C. (2017). *Stiftungen in Deutschland 1: Eine Verortung*. Springer Fachmedien. <https://doi.org/10.1007/978-3-658-13369-6>
- Anheier, H., & Leat, D. (2007). *Creative Philanthropy: Towards a New Philanthropy for the Twenty-First Century* (1. publ., transferred to digital print., p. VIII, 277 S.). Routledge.
- Bixler, R. P., Zappone, M., Li, L. R., & Atshan, S. (2018). Unpacking the Role of Data in Philanthropy: Prospects for an Integrated Framework. *The Foundation Review*, 10(2). <https://doi.org/10.9707/1944-5660.1415>
- Brest, P. (2012). A Decade of Outcome-Oriented Philanthropy. *Stanford Social Innovation Review*, 10(2), 42–47. <https://doi.org/10.48558/K9H3-7Z08>
- Diaz-Bone, R. (2022). Messen. In N. Baur & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 105–122). Springer Fachmedien Wiesbaden. [https://doi.org/10.1007/978-3-658-37985-8\\_6](https://doi.org/10.1007/978-3-658-37985-8_6)
- Fruchterman, J. (2016). Using Data for Action and for Impact. *Stanford Social Innovation Review*, 30–35. <https://doi.org/10.48558/24MM-J709>
- GECES. (2014). Proposed Approaches to Social Impact Measurement in European Commission Legislation and in Practice Relating to EuSEFs and the EaSI. <https://www.fi-compass.eu/publication/other-resources/proposed-approaches-social-impact-measurement-european-commission>
- Greenhalgh, C., & Montgomery, P. (2020). A Systematic Review of the Barriers to and Facilitators of the Use of Evidence by Philanthropists When Determining Which Charities (Including Health Charities or Programmes) to Fund. *Systematic Reviews*, 9(1), 199. <https://doi.org/10.1186/s13643-020-01448-w>
- Grünhaus, C., & Rauscher, O. (2021). Impact und Wirkungsanalyse in Nonprofit Organisationen, Unternehmen und Organisationen mit gesellschaftlichem Mehrwert. [https://www.wu.ac.at/fileadmin/wu/d/cc/npocompetence/12\\_Publikationen\\_NPO\\_SE/Gr%C3%BCnhaus\\_Rauscher\\_Impact\\_Wirkungsanalyse\\_gesellMehrwert\\_Apr2021.pdf](https://www.wu.ac.at/fileadmin/wu/d/cc/npocompetence/12_Publikationen_NPO_SE/Gr%C3%BCnhaus_Rauscher_Impact_Wirkungsanalyse_gesellMehrwert_Apr2021.pdf)
- Hehenberger, L., & Buckland, L. (2023). How Impact Measurement Fosters the Social Economy: From Measurement to Impact of Learning and Management for Impact. In G. Krlev, D. Wruk, G. Pasi, & M. Bernhard (Eds.), *Social Economy Science: Transforming the Economy and Making Society More Resilient* (pp. 138–166). Oxford University Press.
- Juech, C. (2021). Building the Field of Data for Good. In M. Lapucci & C. Cattuto (Eds.), *Data Science for Social Good: Philanthropy and Social Impact in a Complex World* (1st ed., pp. 41–54). Springer International Publishing. <https://doi.org/10.1007/978-3-030-78985-5>
- Kah, S., & Akenroye, T. (2020). Evaluation of Social Impact Measurement Tools and Techniques: A Systematic Review of the Literature. *Social Enterprise Journal*, 16(4), 381–402. <https://doi.org/10.1108/SEJ-05-2020-0027>
- Kassatly, A. (2018). How Philanthropy Infrastructure Can Promote Evidence-Based Giving. *Alliance Magazine*. <https://www.alliancemagazine.org/analysis/philanthropy-infrastructure-can-promote-evidence-based-giving/>
- Lall, S. A. (2019). From Legitimacy to Learning: How Impact Measurement Perceptions and Practices Evolve in Social Enterprise–Social Finance Organization Relationships. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 30(3), 562–577. <https://doi.org/10.1007/s11266-018-00081-5>
- Lapucci, M. (2021). Introduction. In M. Lapucci & C. Cattuto (Eds.), *Data Science for Social Good: Philanthropy and Social Impact in a Complex World* (1st ed., pp. 1–7). Springer International Publishing. <https://doi.org/10.1007/978-3-030-78985-5>
- Münscher, R., & Schober, C. (2015). Welches Interesse verfolgen Organisationen mit einer Wirkungsanalyse ihres sozialen Engagements? – Ein Wegweiser. In C. Schober & V. Then (Eds.), *Praxishandbuch Social Return on Investment* (pp. 23–40). Schäffer-Poeschel.
- Ní Ógáin, E., Lumley, T., & Pritchard, D. (2013). Making an Impact: Impact Measurement among Charities and Social Enterprises in the UK. *New Philanthropy Capital*. <https://www.thinknpc.org/wp-content/uploads/2018/07/Making-an-impact.pdf>
- OECD. (2021). *Social Impact Measurement for the Social and Solidarity Economy: OECD Global Action Promoting Social & Solidarity Economy Ecosystems* (2021/05); OECD Local Economic and Employment Development (LEED) Papers, Vol. 2021/05). OECD Publishing. <https://doi.org/10.1787/d20a57ac-en>
- Rauscher, O., Mildenerger, G., & Krlev, G. (2015). Wie werden Wirkungen identifiziert? Das Wirkungsmodell. In C. Schober & V. Then (Eds.), *Praxishandbuch Social Return on Investment* (pp. 41–58). Schäffer-Poeschel.

- Schober, C., & Then, V. (2015). Was ist eine SROI-Analyse? Wie verhält sie sich zu anderen Analyseformen? Warum sind Wirkungen zentral? Die Einleitung. In C. Schober & V. Then (Eds.), *Praxishandbuch Social Return on Investment* (pp. 1–22). Schäffer-Poeschel.
- Then, V., & Kehl, K. (2022). Philanthropy in Europe. In R. A. List, H. K. Anheier, & S. Toepler (Eds.), *International Encyclopedia of Civil Society* (pp. 1–8). Springer International Publishing. [https://doi.org/10.1007/978-3-319-99675-2\\_640-1](https://doi.org/10.1007/978-3-319-99675-2_640-1)
- Then, V., Schober, C., Rauscher, O., & Kehl, K. (2017). *Social Return on Investment Analysis*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-71401-1>
- Verhulst, S. G. (2021). The Value of Data and Data Collaboratives for Good: A Roadmap for Philanthropies to Facilitate Systems Change through Data. In M. Lapucci & C. Cattuto (Eds.), *Data Science for Social Good: Philanthropy and Social Impact in a Complex World* (1st ed., pp. 9–27). Springer International Publishing. <https://doi.org/10.1007/978-3-030-78985-5>

# 3

## PHIL4DEV

### A text-based machine learning model to compare philanthropic funding across the world

*Nelson Amaya, Harry de los Rios and Madeleine Lessard*

#### **1 A global standard to understand what goes to philanthropy for development**

Since 2017, the OECD has been collecting and publishing financial information from large philanthropic organizations that provide substantial support to activities in developing countries and integrating this information into Official Development Assistance (ODA) statistics. The scope of this work has since expanded, with more than 40 international philanthropic organizations regularly providing financial information to the OECD as of 2023 and philanthropic disclosure standards in emerging markets changing rapidly, making more information on philanthropic funding available. The OECD Centre on Philanthropy has added financial information for over 160 additional organizations from 32 different countries – making available the largest open data repository on philanthropic activities to date (OECD, 2021a, 2021b). This data is quickly becoming an international standard for understanding the contributions of philanthropic organizations to development.

The OECD assesses the nature of philanthropic support using the same classification system applied to all ODA operations in the OECD’s Creditor Reporting System (CRS). The CRS system includes classifications such as financial instruments, geographic scope, and thematic focus. Activities are thematically classified by *purpose codes* – a list of specific topics used to identify the sector to which an individual financial contribution belongs (OECD, 2022). The CRS purpose code classification system offers a comprehensive range of possible development activities, making it suitable for classifying development-related philanthropic activities.<sup>1</sup>

Currently, philanthropic activities are mapped to purpose codes based on a manual review of grant and project descriptions and input from each reporting organization under the guidance of the OECD. This process is key to effective and accurate classifications, as new data providers are unfamiliar with the OECD classification system and face a steep learning curve to use it correctly. However, the review process is onerous and prone to inconsistencies because it requires the verification of thousands of financial transactions at a time, which limits the amount of information that can be verified on a regular basis.

The OECD Centre on Philanthropy has developed a Natural Language Processing (NLP) model to scale, streamline, and harmonize the application of the OECD CRS classification system to philanthropic financial data. *PHIL4DEV* aims to facilitate comparisons of the purposes of

philanthropic funding in a widely applicable way so that any grant or project, described in any language at any point in time, can be understood in a broader context. NLP is effective in text classification because the algorithm can learn the patterns inherent in text through a series of human-coded classifications and then predict the classifications for new text, allowing thousands of grants and projects to be classified in a reliable, reproducible, and accurate manner.

Ultimately, *PHILADEV* aims to become an international standard to help communicate and understand the nature and scope of philanthropy worldwide.

## **2 Mapping philanthropic activities to the classifications in the OECD Creditor Reporting System**

The OECD CRS purpose code classification system is a standardized framework for classifying and categorizing international development assistance projects and activities. The classification system serves as a common language to facilitate the reporting and analysis of aid flows between donor countries and recipient countries. While this classification system was developed with ODA in mind, it has also been used to classify philanthropic giving, including in the Private Philanthropy for Development database (OECD, 2021a, 2021b, 2022).

The CRS classification system is hierarchical and consists of three levels: sector, subsector, and purpose code. At the highest level, the sector represents the broad area where development assistance is directed. The 24 CRS sectors include education, health, agriculture, infrastructure, governance, and environment. Each sector is further divided into subsectors, which are more specific categories within the broader sector. For example, the education sector includes subsectors by level of education: basic, secondary, and postsecondary education. Finally, purpose codes are the most detailed level of classification, representing the specific objectives or activities within the subsectors.<sup>2</sup> These purpose codes (of which there were 234 as of 2021) allow for precise categorization of aid activities, such as the construction of a health clinic (12230 Basic health infrastructure), the provision of textbooks for schools (11120 Education facilities and training), or a university research project on wind energy generation (23182 Energy research).

In cases where a project description does not provide enough detail for a specific objective code to be assigned, it can be reported under its relevant sector or subsector (Figure 3.1). The hierarchical nature of the codes allows for a certain degree of uncertainty at the purpose code level without compromising the accuracy of sector-level totals and comparisons. This is an important feature given the varying quality and level of detail of project descriptions – some organizations provide a high level of specificity on the sectors of their activities, while others provide only very general descriptions. The CRS purpose code system can accurately categorize more broadly described activities while also capturing further detail when available.

Purpose codes are mutually exclusive areas of development, with each grant or project assigned to a single code. This is important for broader analyses of sector allocation across donors, geographies, and over time. However, in practice, many grants support activities across multiple sectors and cannot be accurately described by a single purpose code. This issue is dealt with in two ways: First, if the project is related to two purpose codes within the same subsector or sector, it can be reported using the higher-level code. For example, a program that targets Basic life skills for youth (11230) and Primary education (11220) could be reported as Basic Education (11200), and one that targets both Basic Education and Secondary Education could be reported as Education, Level Unspecified (11100). The second option is to split the project into multiple project lines, each with a unique purpose code, and divide the total funding of the project among the new subprojects. For example, a grant of \$100,000 that supports both feeding lunches to schoolchildren and providing primary education would be split

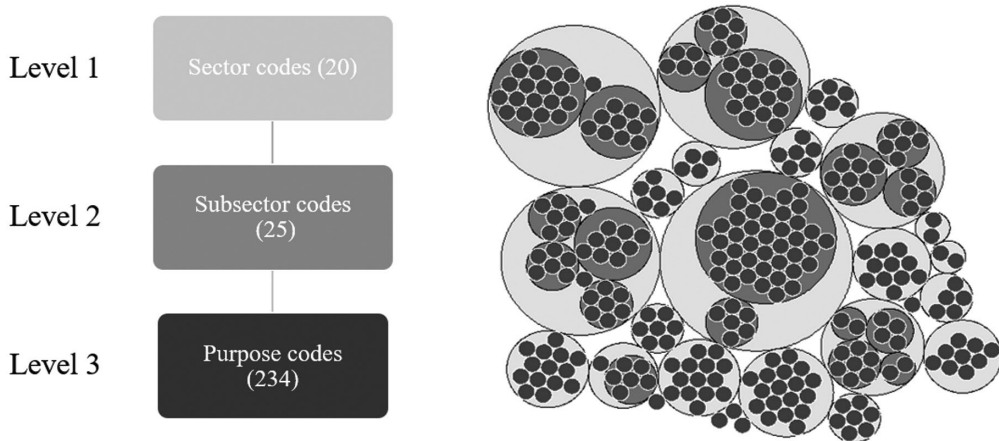


Figure 3.1 Overview of OECD CRS purpose code classification system as of 2021.

Source: Authors based on (OECD, 2022).

into two activities: one under the purpose code School feeding (11250) and one under the code for Primary Education (11220), the total funding for which should add up to \$100,000.<sup>3</sup>

Figure 3.1 summarizes the scope and scale of the OECD classification system. The 234 purpose codes are not evenly distributed across sectors – some sectors have more detailed areas, like Agriculture, Forestry, and Fishing, while others have fewer, such as Action Relating to Debt. While some sector codes contain several subsectors, each of which contains multiple purpose codes, other sectors have only two levels. A handful of sectors, such as Reconstruction Relief & Rehabilitation and Disaster Prevention & Preparedness, have only one purpose code, making them effectively single level.

CRS codes have several useful properties for matching a wide range of philanthropic funding related to economic, social, and environmental development and offer some distinct advantages as a general thematic classification system for philanthropy.

To begin with, it allows for thematic comparisons with ODA funding, helping to put philanthropic activities in a broader context. In addition, by focusing on finding a single best match to the objective of any financial operation, whether it is providing access to education, healthcare, social protection, renewable energy projects, or other possibilities included in the 234 different CRS code descriptions available, it helps overcome issues related to how grants and projects are named versus what they actually do. Moreover, the breadth of the CRS allows for accurate classification of issues commonly supported by philanthropy, such as research, because it can distinguish the specific sector to which each research project seeks to contribute (e.g., health research is classified differently from biodiversity research). Finally, the classification system is adaptive, changing over time as needed to include additional relevant categories or add precision to existing ones; for example, a new code for Covid-19-related funding was added in 2020.

However, the classification system is not without important limitations when applied to philanthropy. First of all, philanthropic organizations often engage in several different types of activities through a single grant or project, and the treatment of these multisector activities can distort areas that are key to understanding the scope of philanthropy. For example, a project that develops a single infrastructure to provide both education and healthcare services to children might be classified



as Multisector, reducing the estimates of total spending in Education or Healthcare, even though it provides services in each of these areas. As mentioned above, this can be addressed by splitting the project into two activities and allocating a share of the funding to each, but it is often difficult to determine precisely what share of the funding goes to each service from philanthropic grant data alone. Finally, the CRS classification lacks granularity in a few areas that are prominent and common in philanthropy, such as cultural and arts projects, provision of services to the elderly and children, microfinance, and donations to specific government funds.

Given the relevance of the CRS classifications, the *PHIL4DEV* model was designed to automate the mapping of philanthropic activities to their best-matched purpose code, based solely on the descriptions of each activity. The model is described in detail in the next section.

### 3 Building PHIL4DEV

*PHIL4DEV* is a supervised machine learning model that maps the descriptions of philanthropic grants and projects to the OECD CRS classifications. Figure 3.2 presents the workflow followed in building this model. The first step in building *PHIL4DEV* was to describe existing data on philanthropic activities, which had already been classified by OECD staff as of 2019. This data helps the model learn underlying patterns to generalize the classification of philanthropic activities to OECD codes. The description of the existing data points to sample imbalance, as the frequency of classes is not evenly distributed. This can cause the model to give more weight to the more frequent categories, resulting in a high number of false-positive classifications to these categories and overall poor performance. Therefore, the second step is to address the sample imbalance by removing target classifications that are infrequent and not relevant to philanthropy, such as public finance activities.

The third step involves estimating and evaluating the model. The model was estimated in the statistical computing language and environment R (Version 4.1) using XGBoost (Chen & Guestrin, 2016; R Core Team, 2021). XGBoost is based on decision trees and optimized for regression and classification tasks.

#### 3.1 Data preparation

To estimate the relationships between philanthropic activities and OECD thematic classifications, the model uses a bag-of-words approach, which represents texts by the presence of words within the text of each activity. To prepare the vocabulary for the model, the data must be cleaned by removing URLs, punctuation, numbers, and other symbols. Data collected in Spanish, French, and Portuguese was translated into English using Google API Cloud. The text was then tokenized by splitting all descriptions into a list of words, trimming English stop words, and stemming the remaining words using the tm R package (Feinerer & Hornik, 2023). Finally, the data was represented as a document term-matrix.



Figure 3.2 PHIL4DEV workflow.

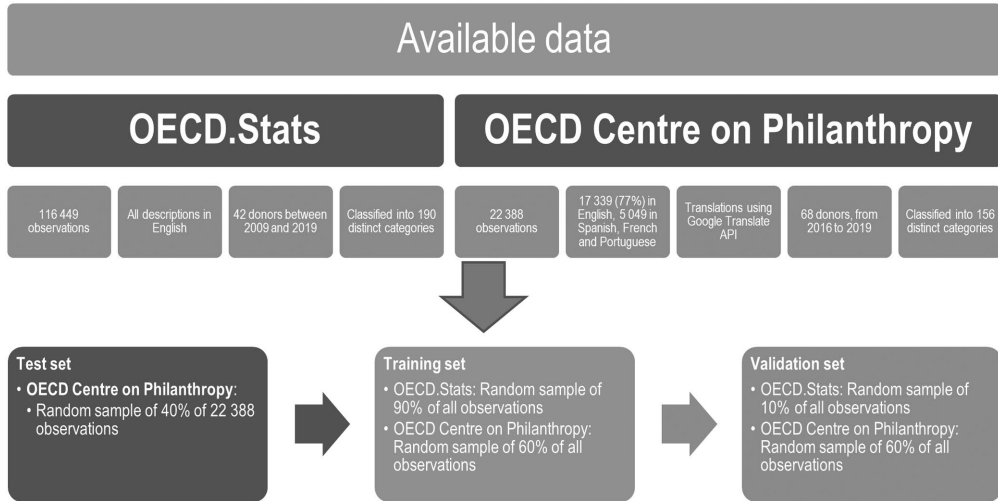


Figure 3.3 PHIL4DEV data workflow.

The process of splitting our dataset into testing (to check out-of-sample performance), training (for the model to identify the underlying patterns between text and classifications), and validation (to calculate performance metrics to compare model parameters) sets helps to prevent overfitting and to adjust the hyperparameters of the model. The process and the data in each set are summarized in Figure 3.3.

### 3.2 Addressing category imbalance

Two steps were taken to address the category imbalance in the sample. The first step was to reduce the number of examples in very frequent classes by randomly removing some examples (Krawczyk, 2016). This meant randomly reducing the examples provided to the training set for a few classes: 80% of the data was removed for activities classified as Infectious disease control, 50% of the data was removed for Family planning and Human rights, and 40% of the data was removed for Reproductive healthcare, Agricultural research, and STD control.

The second step was to add examples for underrepresented classes – those with less than 1,000 examples. This was done by selecting a random sample of classified data from these underrepresented categories and then imputing additional examples using a BERT model (Devlin et al., 2018; Ma, 2024).

### 3.3 Estimation and validation

The hyperparameters for PHIL4DEV were chosen using a random search approach (Bergstra & Bengio, 2012). The model was trained 500 times, each time with a parameter value chosen randomly from a wide range. Once the parameters that give the best model performance are identified, the model is trained. Finally, the model’s performance is evaluated on the test set, which consists of data that the model has not seen before.

To evaluate the model, the performance in predicting each of the categories must be weighed against the proportion that each of the classes is present in the data, given the high imbalance in the original dataset. To do this, we use the micro-average F1 score (Tharwat, 2021). Since the categories

that make up the sample are unbalanced, we use this metric to measure how well our model predicts each category, weighting it by the proportional occurrence of the category across the data. Weighting each class's F1 score by its proportion of samples in the entire dataset results in a micro-F1 score of 0.74. The weighted F1 score has significant variation across classification codes, with some classifications being assigned accurately most of the time, such as Anti-corruption organizations and institutions or Early childhood education, while others perform poorly, such as Basic sanitation or Agricultural policy and administrative management, because they are not common in the original data.

For these infrequent categories, additional human verification, collecting additional data from organizations working on these topics, and user feedback will improve the accuracy of classifications.

### 3.4 Prediction, transparency, and availability to the public

Making the model accessible to potential survey respondents and data providers is an important feature of *PHIL4DEV*. For this reason, the model has been adapted for public consultation and made available online through a Shiny dashboard (OECD, 2021a, 2021b). Users can enter plain text descriptions and find up to three most likely predictions, or a warning that the text is uninformative if no purpose code can be assigned with more than 10% probability (Figure 3.4). Each time the model is consulted, the full description of the predicted purpose codes is displayed, allowing the user to understand in more detail what the model is predicting for their text (Figure 3.5).

## 4 The future of PHIL4DEV

PHIL4DEV can become a more widely used tool, and an international standard, with a few improvements, which will be described in turn.

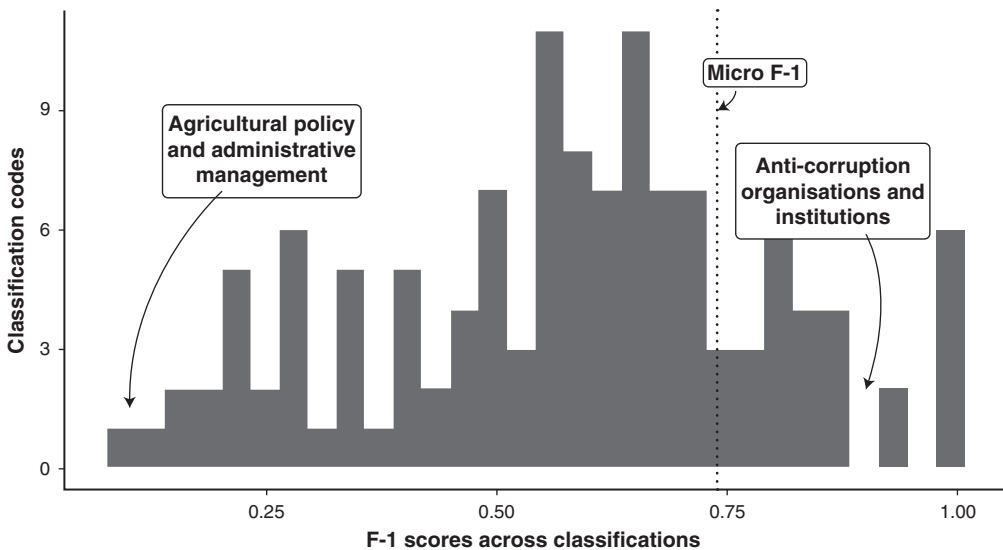
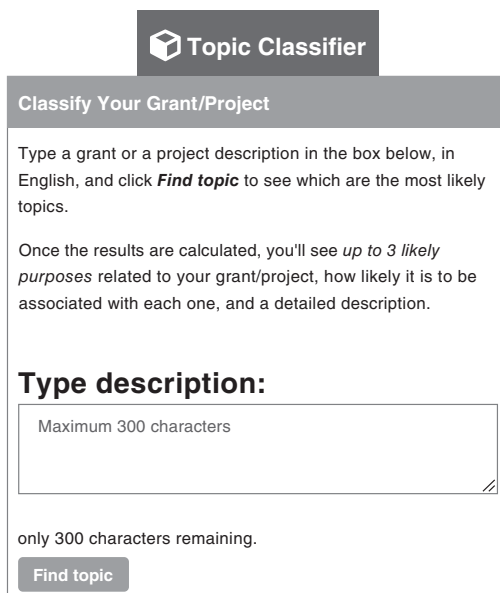


Figure 3.4 Histogram of PHIL4DEV performance across classifications.



**Topic Classifier**

**Classify Your Grant/Project**

Type a grant or a project description in the box below, in English, and click **Find topic** to see which are the most likely topics.

Once the results are calculated, you'll see *up to 3 likely purposes* related to your grant/project, how likely it is to be associated with each one, and a detailed description.

**Type description:**

Maximum 300 characters

only 300 characters remaining.

**Find topic**

Figure 3.5 Public accessibility to PHIL4DEV model using Shiny App.

Source: OECD (2021a, 2021b) <https://oecd-main.shinyapps.io/philanthropy4development/>.

#### ***4.1 Data limitations: multiple languages, text ambiguity, and category frequency***

PHIL4DEV has several limitations. First, the model is not language agnostic, as it is built to make predictions for English texts. This means that if a text is written in another language, it must first be translated into English, with all the errors that a language translation may entail.

Second, since PHIL4DEV uses a bag-of-words approach, it is unable to understand the different meanings of words depending on the context. Two sentences that share the same words but have opposite meanings will be predicted to belong to the same category.

Third, the input to the model, descriptions of philanthropic activities, can often be vague or ambiguous, making predicting classifications very difficult. For a prediction to be accurate, the model requires some degree of clarity in the text, but if the text is too short or does not include words related to the purpose of the activity, the accuracy of the model will be compromised by uninformative input.

Finally, sampling imbalance significantly hampers the ability of the model to generalize to some classifications. Ideally, we would like to have a similar number of examples for each category. We are forced to compensate for this imbalance in the data with sophisticated techniques that, while significantly improving the performance of the model, could introduce biases in the definition of each category.

#### ***4.2 Classification limitations: classification into multiple distinct topics, thematic uncertainty, and lack of granularity in important topics***

Because the CRS classification system was not designed to classify philanthropic activities, the target classification system itself is not without problems.

The one-to-one correspondence between activities and codes is both an advantage and a disadvantage of the classification system. In cases where a project, by its very nature, tackles multiple distinct sectors, and there is no information available on how to break it down into several activities to estimate how much funding corresponds to each one, the classification system will place the project under 43010 – Multisector aid, losing the information on which sectors the project addresses. To overcome this problem, a tagging system for projects can be added so that one project can have multiple purpose codes associated with it, but at the cost of losing the ability to aggregate funding within sectors.

Moreover, while the CRS classification system is well suited to dealing with uncertainty about a project within a sector, it does so better in some sectors than in others. For example, in education, a project that clearly targets an educational intervention but is unclear about the level of education (e.g., primary, secondary) would be classified as Unspecified Level but within the education sector. However, general support to a local Non-Governmental Organization (NGO) without a narrow thematic focus would be placed in the subsector *Government & Civil Society-general* because the classification system does not provide separate treatment for non-governmental organizations, which are the most common recipients of philanthropic funding.

The final issue is one of incompleteness: there are areas of philanthropy for which there is no appropriate classification, even though they are development-related and relevant to philanthropic donors. Philanthropic organizations often provide specific social services, such as supporting orphanages and running older people’s homes, all of which are classified under 16010 – Social Protection, but it is often valuable to distinguish precisely how many resources go to either service.

In conclusion, some small extensions to the classification system to allow for more granularity in activities that are relatively more relevant to philanthropy than to ODA, a parallel tagging system that classifies multisector projects, and a more comprehensive classification for NGOs would all improve the understanding of philanthropy.

### **4.3 Technological limitations: new NLP tools**

The NLP technologies used in the PHIL4DEV model favor interpretability, reproducibility of results, and ease of implementation but come at the expense of greater predictive accuracy and wider scope.

More sophisticated Deep Learning NLP tools, such as BERT models, and Large Language Models such as [GPT4] can address many of the weaknesses of PHIL4DEV, but at the cost of introducing new challenges in terms of interpretability of results (Chakraborty et al., 2017) and overall complexity (Hu et al., 2021). These tools can improve accuracy (Kamath et al., 2018), support multilingual documents (Manias et al., 2023), enable information extraction, and perform question-answering tasks through concrete queries (Brown et al., 2020).

Finally, in order for PHIL4DEV to serve an open tool that can be used by philanthropic organizations to align their philanthropic activities with standardized goals, it should be able to provide immediate responses to a large number of users and also receive feedback from them. State-of-the-art machine learning engineering practices would allow for faster response times for a larger flow of users and data.

In using newer technologies, we see two challenges. The first is the inheritance of biases from pre-trained models that have limited transparency. Since these models are trained on large amounts of manually labeled data, they contain errors and biases inherent in the population from which they were collected. Without detection and treatment (which are limited because these models are trained on non-open data), these errors and biases tend to be amplified in the predictions of these

models (Navigli et al., 2023). More effort should be put into detecting, measuring, and mitigating these biases. The second is the implementation of these technologies, which requires additional hardware and software resources, from training and testing the models to implementing the inference service (Thompson et al., 2020).

## **5 Conclusion: global philanthropy in need of modern tools**

*PHIL4DEV* is the first attempt at a globally applicable tool for comparing philanthropic activities from any context, in any language, based on mapping their descriptions to the OECD classifications for development assistance. The model served as a scaffold for the OECD (2021a, 2021b) global report *Private Philanthropy for Development – Data for Action* and presents a promising way to expand and consistently compare philanthropic activities from very different contexts – to map the philanthropic sector on a global scale.

How will *PHIL4DEV* improve in the future? What can it do and what new insights can it bring to global philanthropy for development?

First, the vocabulary of the model has been trained with 2019 data and should be updated so that it can be applied to new classifications, such as Covid-19. The model should be retrained with newly collected data, and more efforts need to be made to improve the model's precision in some thematic areas, such as civil society and multisectoral interventions. After updating the model with more recent data, estimates of philanthropic contributions to Covid-19 will be more accurate and reliable.

Second, *PHIL4DEV* could be significantly improved by using more advanced NLP techniques that can overcome imbalances and accurately predict classifications for rare cases. This improvement will come from using new techniques and collecting additional data from organizations working on topics that are not very common.

Beyond the accuracy of the model, there are additional opportunities that come from a deep understanding and accurate measurement of what philanthropy is funding, who is receiving it, and where it is going. On the one hand, classified philanthropic data can help reveal trends in the sector that are invisible at any other level, such as reallocations from one topic to another or from one geography to another. On the other hand, this data can help predict how funding could be allocated to new topics, grantees, and countries, which will help private donors with common interests find each other more easily, as information is one of the most prevalent barriers to collaboration in the sector (OECD, 2021a, 2021b, p. 52). With detailed multi-year data on philanthropy for development, the model allows for portfolio similarity analysis so that foundations with common interests and geographies can more easily find spaces to collaborate or avoid duplicating their efforts.

An updated version of *PHIL4DEV* will be developed in 2024 to help classify data for a new global survey that will attempt to integrate all of the above.

## **Notes**

- 1 Other classification systems for philanthropic activities have also used the OECD CRS classification. For foundations based in the United States, see Candid's Philanthropy Classification System. <https://taxonomy.candid.org/>.
- 2 The purpose code taxonomy describes the economic and social sector of the activity, rather than its ultimate objective. For example, an activity focused on improving sanitation would be classified in the Water Supply and Sanitation sector, even if its ultimate goal is to improve health. Similarly, an energy access project would fall under the energy sector, even if its goal is to increase productivity by providing energy services to businesses.

- 3 When a project is split between multiple purpose codes, donors are asked to indicate the percentage of funding that belongs to each code (e.g., in the example above, the donor could indicate that 80% of costs was for primary education and 20% toward school meals). However, in practice this level of information is not always available, and funding is instead allocated assuming equal shares for all relevant purpose codes.

## References

- Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(2). <https://dl.acm.org/doi/10.5555/2188385.2188395>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C.,..., & Amodei, D. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of Deep Learning Models: A Survey of Results. *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1–6. <https://doi.org/10.1109/UIC-ATC.2017.8397411>
- Chen, T., & Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (785–794)*. <https://doi.org/10.1145/2939672.2939785>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. <https://doi.org/10.48550/arXiv.1810.04805>
- Feinerer, I., & Hornik, K. (2023). *tm: Text Mining Package (0.7–11)*. <https://CRAN.R-project.org/package=tm>
- Hu, X., Chu, L., Pei, J., Liu, W., & Bian, J. (2021). Model Complexity of Deep Learning: A Survey. *Knowledge and Information Systems*, 63(10), 2585–2619. <https://doi.org/10.1007/s10115-021-01605-0>
- Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification. *Proceedings of the ACM Symposium on Document Engineering 2018*, 1–11. <https://doi.org/10.1145/3209280.3209526>
- Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence*, 5(4), 221–232. <https://doi.org/10.1007/s13748-016-0094-0>
- Ma, E. (2024). *NLP Augmentation [Jupyter Notebook]*. <https://github.com/makcedward/nlpaug> (Original work published 2019).
- Manias, G. M., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual Text Categorization and Sentiment Analysis: A Comparative Analysis of the Utilization of Multilingual Approaches for Classifying Twitter Data. *Neural Computing and Applications*, 35(29), 21415–21431. <https://doi.org/10.1007/s00521-023-08629-3>
- Navigli, R., Conia, S., & Ross, B. (2023). Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2), 1–21. <https://doi.org/10.1145/3597307>
- OECD (2021a). *Private Philanthropy for Development – Second Edition: Data for Action*. Paris: OECD Publishing. <https://doi.org/10.1787/cdf37f1e-en>
- OECD (2021b). *Private Philanthropy for Development: Data for Action Dashboard [dataset]*. <https://oecd-main.shinyapps.io/philanthropy4development/>
- OECD (2022). *Development Assistance Committee Creditor Reporting System Classifications*. <https://www.oecd.org/development/financing-sustainable-development/development-finance-standards/dacandcr-scodelists.htm>
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing (4.1.0)* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org>
- Tharwat, A. (2021). Classification Assessment Methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The Computational Limits of Deep Learning. *MIT Initiative on the Digital Economy Research*, 4, 1–4.

# 4

## FAST-TRACKING THE USE OF AI IN EVERYDAY PHILANTHROPY

*Stefan Schöbi*

### **1 Technology adaptation with AI and philanthropy**

The exchange dynamics in philanthropy, traditionally anchored in formalized written communication, hold inherent potential for enhancement through language models. This chapter does not delve into the merits of this communication style. Still, it acknowledges its established presence, paving the way for language models to revolutionize the efficiency and effectiveness of the grant application process. Through AI-enhanced grant allocation, the sector can ensure that funds are effectively distributed to maximize its impact.

The sheer potential does not mean technology is adopted heavily and quickly. There has been a lot of research on technology adoption, underlying that the path of technology adoption isn't linear nor is it uniform across individuals or institutions. Historically, society's integration and adoption of new technologies have followed patterns of skepticism, resistance, exploration, acceptance, and eventual dependence. Every transformative invention, from the printing press to the smartphone, experienced a journey from novelty to necessity.

In the context of innovation, the existence of "assimilation gaps" typically refers to the disparities between the introduction of new technologies or practices and their widespread adoption or integration into regular use. In other words: Having access to innovative tools is one thing – the ability to utilize them effectively is another.

This chapter focuses on the actual state of use of AI in the grantmaking process and strategies for fostering AI dissemination, analyzing concrete use cases of applied AI in the grantmaking process. Section 1 is descriptive and delves into AI's current adoption and challenges within Swiss philanthropy, backed by empirical data. Section 2 opens a prescriptive perspective presenting actionable strategies rooted in specific AI applications and – again – supported by empirical survey data to pave a learning journey for everyday philanthropy.

#### ***1.1 Common models for innovation diffusion***

In his foundational text *Diffusion of Innovations*, American sociologist Everett M. Rogers explains how, why, and at what rate new ideas and technology spread through society. Rogers had a prescriptive view of the dynamics of the diffusion of technology, which he understood as adoption,



primarily focusing on the individual. The book is still an important foundational work, not only because of its length (over 500 pages) but mainly because it strongly focuses on communicative aspects of diffusion – which have turned out to be key.

In the very early phases of diffusion, mass media plays a significant role; later, when many personal experiences are available, interpersonal communication becomes essential. So-called Change Agents also play a key role. A Change Agent mediates between a Resource System and a Client System – the Change Agent himself has extensive expertise in dealing with innovation and can pass it on (Rogers, 2003, p. 368). In the later phases, opinion leaders are crucial, namely people who are above average in their ability to influence others' opinions, especially regarding the speed with which innovation is adopted (Rogers, 2003, p. 300).

Everett M. Rogers, in his seminal work, proposed the Innovation Adoption Curve to represent the stages through which an innovation travels from its introduction to its widespread adoption. This is by far the most cited depiction of the work, the most frequently used illustration on the subject of innovation, which should be familiar to most readers. This curve classifies adopters into five main categories: Starting with the Innovators, comprising 2.5%, they are the trailblazers, often taking risks and having the financial means to interact with like-minded pioneers. They are closely followed by the Early Adopters, who represent 13.5% and, while quick to embrace innovations, are more thoughtful than the innovators and often hold influential roles within their communities. The Early Majority, making up 34%, are those who venture into new ideas before the general population but take longer to decide than the early adopters. They're followed by the Late Majority, another 34%, who, despite their skepticism, eventually come around to adopting the innovation, albeit after most have done so. Finally, the Laggards, accounting for 16%, are the most resistant, often preferring tradition over change (see Figure 4.1).

If we compare this book with *The Technology Fallacy* by Gerald C. Kane et al. (2019), we realize that digital maturity is primarily about people and the realization that effective digital transformation involves changes to organizational dynamics.

Kane et al. expand the concept of adoption, which, as Rogers correctly captured, focused on individuals to include adaptation, where innovation is picked up and implemented by organizations, mainly businesses. While businesses adapted more quickly to technology than individuals did 15

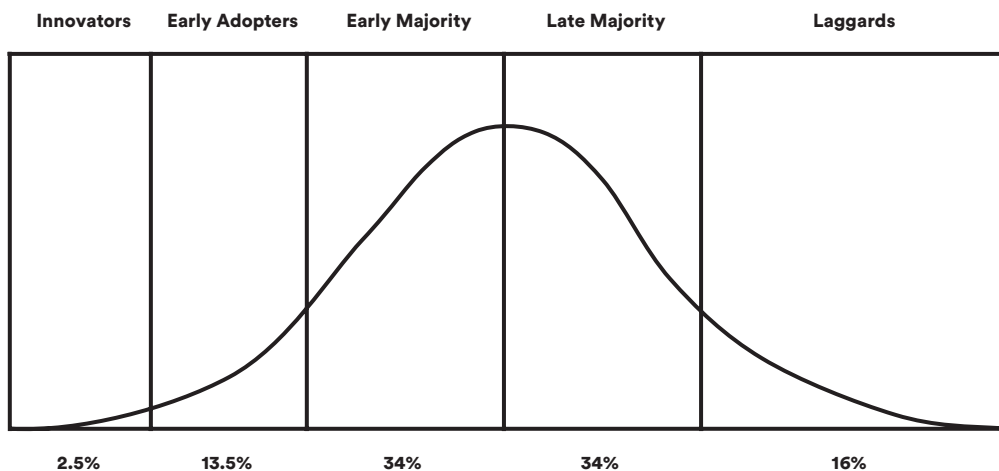


Figure 4.1 Innovation adoption curve by Everett Rogers, see Rogers (2003, p. 281).

years ago, this has changed. Interestingly, the authors say: “Adoption is not [anymore] the most critical digital disruption problem most managers face” (Kane et al., 2019, p. 31).

To substantiate adaptation, Kane et al. introduce the concept of “Digital Maturity” of an organization, which also strongly relates to the well-known concept of the “Growth Mindset” introduced by Carol Dweck (2006) (see Kane et al., 2019, p. 3ff. and figure p. 162). Regarding the diffusion of innovation, the focus has thus clearly shifted to the adaptation gap: “how the majority of individuals want (and expect) to use technology to engage with companies and how companies have adapted to support those interactions” (Kane et al., 2019, p. 33).

The past decades have shown how vital technology adaptation is for companies. This is precisely why Clayton Christensen (1997) discusses how big and successful companies lose their market leadership due to innovation assimilation gaps in *The Innovator’s Dilemma*. The book focuses on the pitfalls companies face when innovations disrupt their markets. He expounds on how these companies, though efficient in their operations, fail to adopt disruptive technologies early on because they fail to see the potential in new markets. Robert Sutton and Jeffrey Pfeffer’s well-known exploration into the dichotomy between knowledge and action underscores that action is needed (Pfeffer et al., 2000). Their principle, encapsulated in the first chapter’s title, “Knowing What to Do Is Not Enough,” highlights that companies often fail not due to a lack of knowledge but because of the inability to act on that knowledge.

### **1.2 The four stages in technology diffusion**

Combining the models of adoption and adaptation, we can create a simple phase model integrating individual adoption as well as organizational adaptation. Building on Rogers’ five adopter categories, we can detail the psychological journey an individual undergoes and combine it with the challenges impacting an organization’s ability to compete, based on Kane et al. (2019, p. 201), splitting the journey into four stages.<sup>1</sup>

During the initial phase of skepticism, potential adopters grapple with doubt and question the innovation’s efficacy. This uncertainty soon transforms into active resistance, as individuals perceive more disadvantages than advantages. However, as they gradually recognize the innovation’s benefits, they transition into exploration, experimenting with its potential. Ultimately, this journey culminates in the full acceptance, integration, and regular use of the innovation.

In *The Innovator’s Dilemma*, Clayton Christensen (1997) emphasizes the “resistance” phase. He suggests that many big corporations remain stuck in this phase due to organizational inertia, current customer demands, or sheer disbelief in the disruptive potential of the innovation. Expanding on the four-phase model, Christensen (1997) would likely add a phase of “Organizational Complacency” between Resistance and Exploration, where organizations (i.e., a significant part of the management) acknowledge the innovation but believe they can weather its effects or adopt it later with ease.

As stated, *The Technology Fallacy* by Gerald C. Kane et al. introduces the idea that the mere adoption of new technologies isn’t enough. What truly matters is understanding the change in organizational dynamics and the human element of digital transformation (Kane et al., 2019, chapter 6ff.). The primary argument is that digital maturity isn’t about the technology itself but how organizations adapt and change their cultures and processes to make the most of these tools (Kane et al., 2019, chapter 2).

Drawing from Kane’s insights, we can further expand on Rogers’ stages of innovation diffusion. *The Technology Fallacy* introduces an element of “Organizational Culture Shift” that should occur mainly alongside Exploration.<sup>2</sup> Before true acceptance can take place, organizations must

not only explore the technology but also adapt their organizational cultures to the demands and possibilities the innovation introduces. Moreover, Acceptance can be a stage of continuous adaptation, because: “Maturity is never complete” (Kane et al., 2019, p. 46). This aspect recognizes that in the digital era, innovations are continuous, and organizations must perpetually adapt, relearn, and refine their approaches to stay relevant.

Throughout our exploration of innovation diffusion, the contributions of Rogers, Christensen, and Kane have provided a framework for how individuals and organizations navigate the complexities of change. Rogers introduced us to the foundational trajectory of adoption, tracing the journey from initial skepticism to full acceptance. Christensen (1997), working on the pitfalls companies face, added depth to our understanding of the resistance phase, adding an important “Organizational Complacency” state – a state where organizations remain stuck and need a “push” to get into active exploration. Kane further refined this perspective by emphasizing the human element of digital transformation, suggesting that true acceptance requires not just the adoption of technology but also an organizational culture shift, both happening concentrated in the exploration phase, which in a certain sense becomes a permanent state.

In sum, the refined model of innovation diffusion begins with skepticism, characterized by doubt about the innovation’s value. It then progresses to resistance, where the perceived disadvantages often outweigh the benefits. However, as benefits become apparent and organizational dynamics shift to accommodate the new reality – especially when an additional kick occurs – the exploration phase ensues. This stage is crucial, as it involves experimenting with the innovation and adapting organizational cultures. The journey concludes with acceptance, where innovation becomes an integral part of regular operations, and, following Kane, must be understood as a kind of continuous exploration. The insights remind us that, while technology evolves, the success of its adoption hinges on human and organizational adaptation and readiness to change. In this way, the technology adoption and adaptation journey often mirror the human psychological response to change (see Figure 4.2).

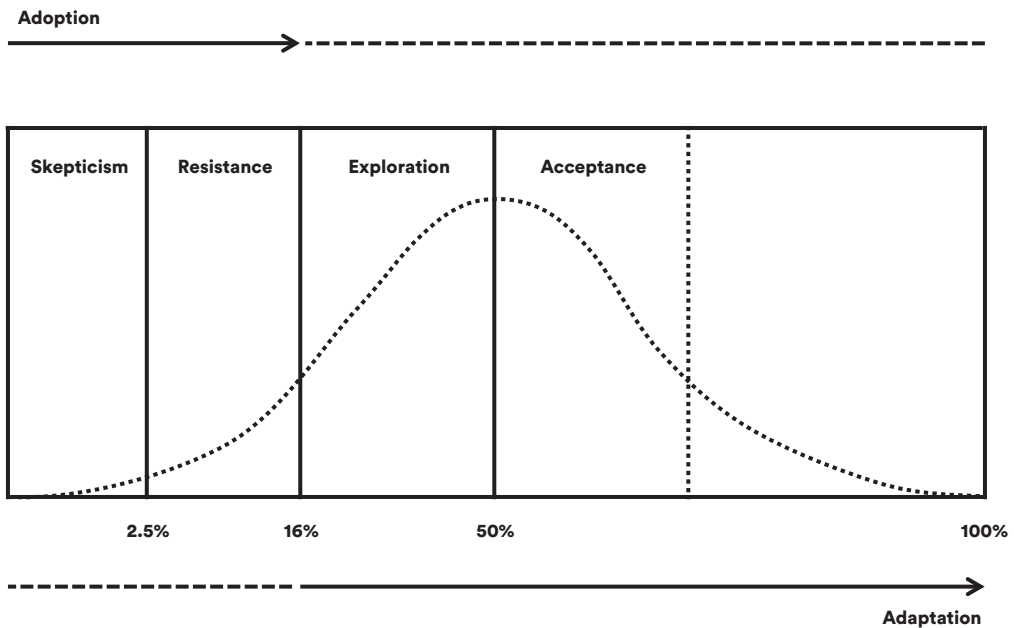


Figure 4.2 Four states of innovation diffusion, including confidence thresholds.

### 1.3 Defining the readiness of AI adaptation in Swiss philanthropy

But while the potential of AI is often touted, its real-world adaptation – particularly in niche sectors like philanthropy – remains an empirical question. To shed light on this, we embarked on a comprehensive survey, engaging both grantseekers and grantmakers in the Swiss philanthropic arena. With 89 representatives lending their insights, this endeavor represents one of the first thorough examinations of AI readiness within the sector to date. Our survey honed in on three key dimensions of AI within philanthropy. We first gauged its current use, capturing the present-day integration of AI technologies. We then assessed anticipated AI utilization, offering a glimpse into future aspirations. Finally, we identified challenges, spotlighting the primary inhibitors hindering AI adaptation in the sector. These insights, captured in the subsequent tables, provide a tangible snapshot of where Swiss philanthropy stands vis-à-vis AI adaptation (Tables 4.1–4.3).

Table 4.1 Current use of AI, n = 89

<i>Current use of AI</i>	<i>Funders (%)</i>	<i>Nonprofits/mixed (%)</i>	<i>All (%)</i>
We have no experience with AI.	88	71	78
We use chatbots or virtual assistants for communication.	0	7	4
We use AI for the analysis of donor data and fundraising.	0	2	1
We use AI for content creation, for example, for requests (text, image, motion picture).	3	16	11
We employ AI in recruitment/human resources management.	0	2	1
We use AI for automated application preliminary review.	0	2	1
We have developed our own AI applications.	3	0	1
Calculated usage index [first question inverted]	12	29	22

Table 4.2 Expected use of AI, n = 89

<i>Expected use of AI</i>	<i>Funders (%)</i>	<i>Nonprofits/mixed (%)</i>	<i>All (%)</i>
We expect AI to increase the efficiency of our operational processes.	6	27	19
We hope that AI will assist us in identifying funding projects.	12	11	11
We wish that AI supports us in application submissions.	9	33	24
We see AI as an opportunity to strengthen our donor relationships.	12	24	19
We hope for innovative solutions to societal challenges through AI.	6	31	21
We plan to use AI for data-based decision-making.	0	13	8
Calculated confidence index [mean of all]	7	23	17

Table 4.3 Inhibitors, n = 89

<i>Inhibitors</i>	<i>Funders (%)</i>	<i>Nonprofits/mixed (%)</i>	<i>All (%)</i>
<i>Lack of understanding/internal training needs in the area of AI.</i>	47	35	39
Data protection and security concerns when dealing with AI.	35	51	45
Concerns about ethical issues and responsible use of AI.	47	47	47
Lack of financial resources for the implementation of AI.	24	35	30
Difficulties integrating AI into existing processes and systems.	35	22	27
General skepticism about the benefits of AI for our foundation's goal.	32	38	36

Our survey paints a distinct image of AI readiness within Swiss philanthropy. Members of grantmaking organizations show a modest 7% confidence level toward AI adoption, while grant-seeking counterparts reflect a heightened confidence at 23%. This highlights the nuanced perspectives and challenges across the philanthropic landscape.

Before further interpreting these figures, one might wonder: Are these confidence levels high or low in a global context? For perspective, we turn to the Artificial Intelligence Index Report 2023, an independent initiative at the Stanford Institute for Human-Centered Artificial Intelligence (HAI), supported not only by big corporates like Google but also by the grantmaking foundation “Open Philanthropy” based on the doctrine of effective altruism. Their broader question to the general populace delved into whether “products and services using artificial intelligence have more benefits than drawbacks.” Their findings were intriguing: while China displayed a robust confidence at 78%, Germany registered 37%, the US stood at 35%, and France at 31% (p. 324).<sup>3</sup>

When juxtaposing our findings against Stanford's, it is evident that even the lower international confidence levels, like those of the US and France, outpace Switzerland's grantmaking institutions. However, this discrepancy becomes understandable when considering the context of each survey. Stanford's inquiry tapped into general sentiments from a broader population regarding AI's overarching utility. In contrast, our survey navigated the more intricate waters of a specialized professional group inquiring about a specific AI application within philanthropy. Naturally, such specificity and niche focus would yield more conservative confidence levels.

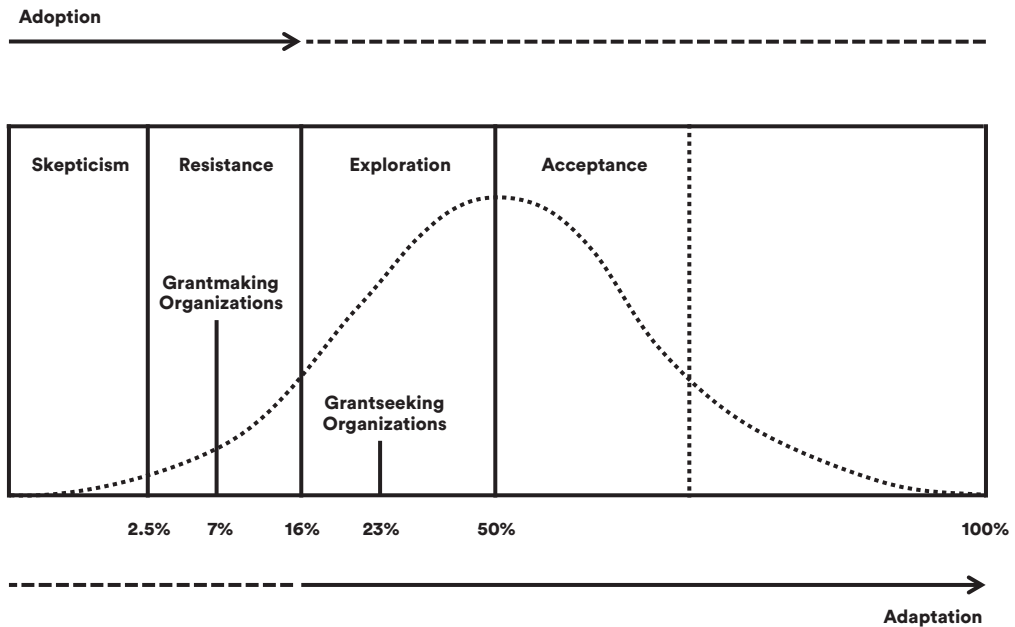
Now, mapping the intangible process of innovation adoption and adaptation into tangible metrics is a challenging task, as linking a conceptual model to quantified research findings poses various methodological hurdles. Straightforward methods like linear mapping or using the bell curve approach seem unhelpful at first glance. An empirical cluster analysis seems ideal in many respects, pinpointing natural groupings based on real-world feedback. However, such an approach necessitates extensive surveys and comprehensive studies, which we currently lack due to the dearth of such empirical research.

Consequently, our attention is drawn – again – to a widely recognized schema: the Rogers' innovation adoption curve. While not directly geared to our precise needs, it is a formidable tool

when guided by certain premises. Acknowledging the significant time lag between technological preparation and widespread use is essential. Thus, the midpoint of Rogers’ curve at 50% must signal a beginning phase of acceptance. Notably, the exploration stage does not demand a majority; the actions of innovators and early adopters alone can set the stage. Armed with these premises, Rogers’ curve offers a promising bridge between the conceptual and the quantifiable.

Building on Rogers’ curve, we can derive a refined mapping tailored to our needs. The skepticism phase represents the initial latency, encapsulating confidence levels from 0% to 2.5% – where, with regard to the model, it is only the innovators who are at work. Progressing into the resistance bracket, which spans to 16%, we see some initial but not yet comprehensive experiments. The exploration phase, stretching to 50%, witnesses an early majority amplifying the momentum, realizing the innovation’s potential. Surpassing the crucial 50% mark, we transition into the acceptance realm, marking the innovation’s broad and impactful resonance.

The data paints a clear picture: grantmaking organizations with a 7% confidence level are in the “Resistance” stage, indicating hesitations or barriers to AI adoption. Meanwhile, grantseeking organizations, at 23%, fall into an early “Exploration” stage, signaling an emerging interest in AI’s potential. Supporting this classification, a delve into inhibitors provides telling insights. A substantial 47% of grantmaking organizations pinpoint a “Lack of understanding/internal training needs in the area of AI” as a significant roadblock hindering them from entering the “Exploration” stage. In contrast, grantseeking entities display a higher concern (51%) about “Data protection and security” when engaging with AI, which is indicative of their exploratory phase where the practicalities of AI implementation come to the fore. These inhibitor findings thus resonate well with the earlier confidence levels – in sum, grantseekers are showing tentative steps toward AI, whereas grantmakers remain more reserved, highlighting nuanced distinct readiness levels within the Swiss philanthropic sector (see Figure 4.3).



*Figure 4.3* Empirical mapping of AI innovation adoption stages in Swiss philanthropy: grantseeking vs. grantmaking organizations.

## **2 Fast-tracking AI adaptation in philanthropy**

From a sociological standpoint, models like the four-phase one observe and describe phenomena without judgment. However, when viewed through an economic lens, especially considering insights from heavyweights like Christensen and Kane, the focus shifts to devising strategies to accelerate through the phases. This also reflects the general reception of Rogers' seminal work as moving from a descriptive to a proscriptive model (Rudd, 2016, par. 4).

To fasten the adoption of technology at the individual and institutional levels, it is crucial to understand and leverage certain strategies that address the barriers and motivators of adoption. Fast-tracking the process, especially the phase of exploration, can be a key to exploiting the full potential of innovations and maintaining a competitive edge in today's rapidly changing digital landscape.

Guided by this understanding, our approach will be methodical: initially, we will gauge strategies to expedite innovation adoption at large. Subsequently, we will identify a palette of use cases that encapsulate AI's potential. These cases will then be empirically evaluated, prompting another layer of our research with an additional survey, scrutinizing the desirability and feasibility of the cases with a focus on addressing the core concerns. Our endeavor will culminate in synthesizing these insights into an actionable guideline primed for smooth implementation.

### ***2.1 Strategies and influencing techniques for fast-tracking technology adoption***

In the journey of technology adoption, several overarching strategies are universally beneficial for accelerating innovation uptake, among showcasing advantages and its real-world efficacy by presenting success stories and their tangible outcome, which is probably the most obvious. To ease integration, ensuring the new technology's compatibility with existing systems can be instrumental: Making the technology compatible with existing systems and ensuring easy integration can remove significant barriers (Rogers, 2003, p. 15). Offering plug-and-play solutions, APIs, or interoperable features can be a game-changer. Furthermore, building trust through unwavering transparency about the technology's capabilities (and limitations), data privacy, and security features is essential.

Nevertheless, each of the four stages of adoption has its unique barriers and motivators. Addressing these specific concerns can significantly improve the speed of transitioning from one stage to the next. The primary barriers in the Skepticism phase include fear of the unknown and lack of perceived relevance or understanding of the technology, often indicated by vocalized doubts about its feasibility or utility. Addressing these requires clear communication about the technology's benefits and potential impact (this is why Rogers considers mass media to be beneficial in the early stages, see Rogers, 2003, p. 205). Demonstrations and real-life success stories can be powerful motivators, illustrating tangible benefits. Skepticism generally can be addressed by raising awareness and offering factual, easily digestible information about the technology (Fichman and Kemerer, 1999, p. 270).

Resistance, the second phase, often stems from perceived threats to job security or a steep learning curve. This requires understanding the root causes of the concerns and addressing them directly, often through transparent communication and possibly by tweaking the technology or its implementation strategy. Offering training sessions, mentorship programs, and assurance about job roles can motivate individuals to move beyond this stage.

The exploration phase holds a central role in the technology adoption journey. The critical barrier here is the uncertainty of implementation. Engagement in discussions, training sessions,

or hands-on workshops and showcasing early successes can be strong motivators, driving faster adoption. In essence, if a company succeeds in establishing a culture of experimenting, the exploration phase will be reached more quickly and will be more successfully shaped (which is a central conclusion of Kane et al., 2019, chapter 14).

Exploration should be encouraged by creating a supportive environment where potential users can learn, ask questions, and test the technology without judgment or excessive pressure. Strategies to optimize this phase include gamifying exploration or designing captivating onboarding processes. These measures can make the exploration journey informative and enjoyable, propelling users to delve deeper faster. When potential adopters see their peers benefiting from technology, they're more likely to explore and adopt it themselves. According to Rogers' four groups distinguished in the innovation adoption curve, partnering with innovators and early adopters as "Change Agents" can catalyze adoption rates, so to say, by translating "intent into action" (Rogers, 2003, p. 370).

The role of Change Agents in this process is indispensable and deserves special attention, as emphasized by Rogers, who dedicates an entire chapter to this topic (see Rogers, 2003, chapter 9). These agents act as critical interfaces between the resource and client systems, possessing in-depth expertise in the innovation at hand – in this case, artificial intelligence – while also having direct access to relevant application fields such as philanthropy. Their diverse responsibilities include developing a need for change, ensuring the exchange of information, identifying problems and hurdles in the client system, and ultimately facilitating the transition from intent to action (Rogers, 2003, p. 369). Change Agents mainly function as accelerators of technology adoption in heterogeneous groups, reaching beyond their group to integrate diverse perspectives and needs (Rogers, 2003, p. 305f.). This dynamic and adaptive approach is crucial for successful and accelerated technology adoption and represents a key factor in enabling organizations to meet the challenges of digital transformation effectively.

As said earlier, the ultimate goal is not just adoption but optimal utilization of the innovation. By optimally supporting the exploration phase, organizations can ensure that they're not just adopting a technology but truly integrating it into their operations, extracting maximum value; being swift in the exploration phase ensures that organizations stay ahead of the curve, continual learning being "the best response to digital disruption" (Kane et al., 2016, p. 240).

## ***2.2 Use cases for applied AI in the grantmaking process***

Commencing our exploration, we identify a palette of use cases that encapsulate AI's potential. Historically, the grantmaking process heavily relies on the art of the written word, making Large Language Models (LLMs) a tempting avenue to explore. The richness of textual narratives in grant applications and reports positions LLMs as a plausible tool. However, the waters are murky. The realms of philanthropy demand nuanced intuition, assimilation of deep-rooted cultural values, and a reservoir of background knowledge. Unfortunately, these intricacies and the dependence on non-formalized intuitions form a rather unwelcome terrain for LLMs.

While there's caution in the air, it is also ripe with opportunity. When dissecting the AI solution spectrum, we encounter two main avenues: (a) AI meant to enhance and work alongside humans, frequently termed as "augmented" or "assisted intelligence" (Walch, 2020), and (b) AI that works in isolation, intending to replace human roles, often known as "autonomous solutions" or "automated decision-making." The former holds promise for philanthropy, where the human touch is irreplaceable. Envision AI not as the decision-maker but as a trusted aide, augmenting human capabilities and helping philanthropists refine their choices.



Diving deeper into the core capabilities of LLMs, they fundamentally excel at “predicting the next word in a given sequence” (Benaich et al., 2023, Slide 6). Now, translating this strength to the philanthropic domain demands circumspection. Acknowledging that LLMs aren’t inherently crafted to perform grant-matching magic is crucial. Their utility in this domain is a hypothesis, and hypotheses need rigorous testing. Their use is comparable to the off-label use of a drug: While it might be effective in various scenarios, one should exercise extra caution due to associated risks and side effects. Evaluability becomes key. It is not merely about whether LLMs can aid in grantmaking but about the precision and accuracy of their predictions. Hence, any incorporation of AI into philanthropy should have an in-built mechanism to gauge its veracity. A feedback loop, ensuring regular sanity checks, is an indispensable component of this endeavor.

As we navigate deeper into AI’s potential in philanthropy, it becomes imperative to harness the revelations of industry reports, like the Google State of AI Report 2023, and mold them for our niche. The report’s emphasis on unlocking enterprise data dovetails seamlessly with philanthropy’s landscape. The “Retrieval-Augmented Generation” (RAG) model shines through as a beacon, wielding the power of LLMs merged with internal databases to sieve out pertinent content. The report’s acclaim for the “Reinforcement Learning from Human Feedback” (RLHF) approach resonates with the very ethos of philanthropy – where decision-making and precise matching are at the heart of operations. Though LLMs could spearhead the ideation of use cases, we’re equipped with a reservoir of insights from surveys and deliberations, curating a spectrum of use cases that, for our intent, are more illustrative than exhaustive.

The AI revolution in philanthropy sets the stage for numerous transformative interventions. AI’s capability to refine applications, promising greater chances of success, complements its knack for offering predictive analytics on application success probabilities even before they are submitted. As we proceed, AI plays a pivotal role in automating preliminary evaluations of applications and identifying funding needs. It can spotlight crucial areas of interest and potential gaps, all while ensuring a robust ethical foundation and fairness in projects. Moreover, AI could simplify the drafting of activity reports and streamline the creation and dispatch of progress reports. The crescendo is reached when AI independently reviews funded project reports or even supports measuring philanthropic initiatives’ tangible and intangible impacts.

Our spotlight predominantly illuminates the realms of matching but also selectively touches upon facets of need identification, reporting intricacies, and the pivotal aspect of impact measurement, so we come up with the following six use cases:

- AI optimizes applications and increases their chances of success;
- AI predicts the chances of success of an application before submission;
- AI facilitates automated preliminary reviews of applications;
- AI assesses the need for funding, suggests topics, and identifies gaps;
- AI independently evaluates reports from funded projects;
- AI assists in measuring the impact of projects.

### ***2.3 Understanding the root causes of Swiss actors’ concerns***

To unravel the use cases with the biggest potential to be tackled within the exploration phase, we turned to a model birthed at the Stanford-based consultancy IDEO during the early 2000s (Gerber, 2019). IDEO framework underscores the importance of achieving harmony between desirability, feasibility, and viability in innovation. Although the model was initially conceived for the broader business landscape, its core remains relevant and vital for philanthropy. With profitability being a

non-criterion for the philanthropic domain, the model succinctly converges to two pivotal dimensions: desirability and feasibility. Any potential AI use case in philanthropy, therefore, must cater to real needs while aligning seamlessly with the organization’s operational capabilities.

Embarking on the announced additional survey, we now presented our respondents with the curated selection of the six predominant AI use cases in philanthropy. Recognizing the expertise variance among our respondents, we tailored our inquiries to resonate more intuitively. Instead of traditional innovation jargon, we framed our questions around each case’s likelihood (feasibility) and allure (desirability). Crucially, we encouraged every participant to elucidate their choices, enabling us to glean insights into potential inhibitors and accelerators. For those ambivalent about desirability, we sought general perspectives to enrich our understanding (see Table 4.4).

It is important to remember that the secondary survey was dispatched exclusively to the vanguard of our respondents – those poised at the forefront of AI’s adoption in philanthropy. This filtering culminated in responses that bore an inherently progressive hue. Secondly, in a striking symmetry, the evaluations of both grantseeking and grantmaking organizations hovered in close proximity if we look at the mean of the six AI use cases, mirroring each other in terms of feasibility and desirability. However, desirability assessment trailed a more conservative path when juxtaposed against feasibility. This difference hints at a nuanced understanding among participants: just

*Table 4.4* Feasibility and desirability of the most common use cases for AI in everyday philanthropy, n = 20

<i>Feasibility and desirability of common use cases</i>	<i>Feasibility</i>			<i>Desirability</i>		
	<i>Funders (%)</i>	<i>Nonprofits (%)</i>	<i>All (%)</i>	<i>Funders (%)</i>	<i>Nonprofits (%)</i>	<i>All (%)</i>
In the near future, AI will optimize applications and increase their chances of success.	75	85	80	35	60	48
In the near future, AI will predict the chances of success of an application before submission.	60	60	60	55	35	45
In the near future, AI will facilitate automated preliminary reviews of applications.	75	90	83	60	65	63
In the near future, AI will assess the need for funding, suggest topics, and identify gaps.	45	55	50	50	50	50
In the near future, AI will independently evaluate reports from funded projects.	55	55	55	50	35	43
In the near future, AI will assist in measuring the impact of projects.	65	70	68	70	65	68
Calculated mean	63	69	66	53	52	53

because something can be technologically achieved does not mean it is unerringly coveted. To cite specifics, the zenith of feasibility was anchored by the potential of “automated preliminary reviews of applications,” enjoying an approval of 83%. Conversely, the gauge of AI’s potential in “assessing the need for funding, suggesting topics, and identifying gaps” was perceived more reservedly, marking the feasibility floor at 50%. This underscores the intricate and multifaceted nature of determining the full spectrum of requirements for effective philanthropic interventions. Regarding desirability, the prospect of “AI-fortified impact measurement of projects” basked in a robust 68% endorsement. In contrast, the allure dwindled to 43%–45% for scenarios positing “AI-driven evaluation of reports” or “predictive assessment of an application’s success probabilities.”

Venturing into a deeper dissection of the variances, specific use cases revealed palpable rifts between funders and nonprofits. The vision of “future applications being optimized by AI” resonated less with funders, who exhibited a lukewarm 35% desirability, starkly contrasting the more enthusiastic 60% from nonprofits. This trend may reflect nonprofits’ appetite for innovation and their pursuit of standing out in a competitive grantseeking landscape. Conversely, nonprofits manifested a restrained optimism regarding the desirability of “AI predicting an application’s chances of success,” with only 35% giving it the nod, as opposed to a more confident 55% from funders. Delving into the undercurrents of these stances, a dominant sentiment surfaces: nonprofits anticipate that algorithmic, data-driven evaluations might usher in a new era of impartiality, ensuring applications are assessed more equitably (see Table 4.5).

Our examination of the feedback on technology implementation in philanthropy brings forth some distinct perspectives from funders and nonprofits. For concerns, there’s a strong apprehension around AI leading to biases, both from its reliance on past data and potential intrinsic programming biases. Funders and nonprofits alike highlighted concerns about AI’s ability to capture the “human touch,” societal relevance, and the deeper, often silent impact aspects of projects. Another recurrent theme is the fear that AI might lead to homogenization in project selection, favoring specific projects due to built-in biases and potentially raising the bar so high that only projects explicitly tailored for AI evaluation might succeed.

On a brighter note, the chances or opportunities identified lean heavily on efficiency, optimization, and relevance. Both funders and nonprofits envision AI tools leading to more pertinent project matches, fewer misaligned grant applications, and a significant saving of time. Another crucial point of convergence is the idea that humans will be able to focus more on creative processes, leaving routine verifications and checks to AI, which can handle them with accuracy and speed.

#### ***2.4 Embarking on a learning journey for AI in everyday philanthropy***

Merging findings into a “guideline” for fast-tracking AI adoption, we face a vibrant panorama of the Swiss philanthropic landscape’s relationship with Artificial Intelligence. We discern a tangible hesitance from grantmaking organizations, while their grantseeking counterparts evince a budding curiosity about AI’s offerings. By consolidating these insights, we can sculpt a learning journey that pioneering organizations wanting to use AI in everyday philanthropy embark on.

*Create a nurturing experimentation platform* (based on this chapter’s Section 1.1; see also Kane et al., 2019): To facilitate adaptation within the philanthropic sector, it is essential to establish an environment conducive to experimentation and learning. This platform should allow individuals to take risks and innovate without fear of failure. Encouraging a homophilous tendency within this group can be beneficial, as it creates a comfortable and supportive atmosphere for members. By fostering such a space, philanthropic organizations can experiment with AI applications in a secure and encouraging environment, thereby accelerating the adoption and adaptation process.

*Table 4.5* Concerns and chances of technology implementation

<i>Concerns and chances of technology implementation</i>	<i>Funders</i>	<i>Nonprofits</i>
Concerns	<ul style="list-style-type: none"> <li>• AI-optimized language obfuscates real motivations leading to homogeneity.</li> <li>• Superficially good projects may not fit.</li> <li>• The personal, irrational factor.</li> <li>• AI decisions influenced by past decisions can lead to biases and stagnation.</li> <li>• AI's knowledge limits.</li> <li>• Context-specific knowledge gaps.</li> <li>• Biases in AI.</li> <li>• Need for bias-free reliable information.</li> <li>• Lack of learning from AI.</li> </ul>	<ul style="list-style-type: none"> <li>• AI may raise standards but risk generalization.</li> <li>• AI access becomes a criterion and may favor certain projects due to biases.</li> <li>• Uncertainty about AI's ability to replicate human touch in funding.</li> <li>• Access to AI solutions and its evaluation criteria.</li> <li>• Bias and discrimination.</li> <li>• AI's reliance on past data is unsuitable for innovation.</li> <li>• Concerns over transparency and bias-free AI.</li> <li>• AI can't replace human experience and critical thinking.</li> <li>• Challenge in capturing societal relevance and silent impacts.</li> </ul>
Statements	<ul style="list-style-type: none"> <li>• Selection and evaluation criteria will adapt with AI.</li> <li>• Uncertainty about AI's influence on decisions.</li> <li>• Machine bias may become an ethical issue.</li> <li>• Tendency to chase funds rather than pursue importance.</li> <li>• Risk to relationships between partners.</li> <li>• Report content and relevance written and read by AI.</li> </ul>	<ul style="list-style-type: none"> <li>• Equal AI usage does not necessarily improve individual success rates.</li> <li>• Uncertainty about AI's widespread adoption by 2030.</li> <li>• Decision depends on the current context of the donor.</li> <li>• The unpredictability of human decisions vs. AI's logic.</li> <li>• Humans' inconsistent behavior makes them difficult for AI to fully understand.</li> <li>• Emphasis on human involvement in solution crafting.</li> <li>• Need for holistic representation of a project's impact.</li> <li>• Expertise remains crucial for understanding context.</li> </ul>
Chances	<ul style="list-style-type: none"> <li>• Less "fishing in the dark."</li> <li>• Efficiency based on collected data.</li> <li>• More relevant project matches.</li> <li>• Better matching is desirable.</li> <li>• Fewer misaligned grant applications.</li> <li>• Only strategy-aligned proposals get through.</li> <li>• Proper language and structure simplify content verification.</li> <li>• Time-saving.</li> <li>• Emphasis on impact.</li> <li>• Data-driven decision-making becomes accessible even for smaller foundations.</li> <li>• Efficiency in assessing impact.</li> </ul>	<ul style="list-style-type: none"> <li>• AI can increase efficiency in draft creation and formulation.</li> <li>• Strengthening partnerships and optimizing resource allocation.</li> <li>• Humans can focus on creative processes.</li> <li>• AI-driven data analysis can maximize impact.</li> <li>• AI can enhance linguistic quality and ensure queries are addressed.</li> <li>• Preliminary evaluations support stakeholders.</li> <li>• AI can handle routine checks based on set criteria.</li> <li>• Efficiency gains with rapid AI-supported feedback.</li> </ul>

*Empower and expand change agent networks* (based on this chapters Section 1.1; see also Rogers, 2003, chapter 9): Establishing a multilayered approach to change advocacy is vital. Early adopters within organizations should be identified and empowered as primary Change Agents. These pioneers, equipped with AI experiences and insights, play a crucial role in initiating the ripple effect of knowledge dissemination. Envisioning this spread as a model of concentric circles, these primary agents can inspire and educate secondary agents within their immediate circle. These secondary agents, in turn, extend the reach to more distant circles, encompassing a diverse and wider audience. This tiered system of Change Agents ensures that knowledge and enthusiasm for AI in philanthropy permeate through various levels of the organization and community, creating a comprehensive and far-reaching impact on AI adoption and adaptation.

*Harness the readiness of nonprofits* (based on this chapter's Section 1.3): Given the data that delineates the more adventurous spirit of nonprofits, it is judicious to primarily focus on AI use cases that cater to their perspective. Being in the exploration phase, they bring forth a fertile ground for innovative experiments. However, their endeavors should not be isolated. Including funders in this exploration is vital. This dual participation ensures that AI solutions are molded, keeping both ends of the spectrum in mind.

*Foster assistive use of technology* (based on this chapter's Sections 1.3 and 2.1): Our findings strongly emphasize the pertinence of using AI as an assistive tool rather than a complete replacement. A prominent concern is the ethical intricacies and data protection involved with AI. To address this, the focus should be on developing assistive AI prototypes that boost the capabilities of human counterparts rather than overshadowing them and automating decision-making. When combined with continuous feedback loops, this approach guarantees that the technology remains transparent and malleable to philanthropic needs.

*Demonstrating AI's potential through prototypes* (based on this chapter's Section 2.1): In alignment with the strategies for fast-tracking, creating experimental AI prototypes that illustrate the tangible benefits in real-world use cases is paramount. These prototypes act as tangible evidence, quelling doubts and uncertainties, and enabling stakeholders to witness firsthand the efficacy of AI. Furthermore, they can serve as platforms for rigorous discussions, letting users critique, commend, and suggest improvements.

*Capitalizing on efficiency gains* (based on this chapter's Section 2.3): A recurring theme across our findings is the allure of efficiency. By steering AI toward automating preliminary tasks like application reviews, organizations can channel their human resources toward more intricate tasks. Emphasizing AI's role in bolstering efficiency, combined with ensuring humans still play a pivotal role in intuitive and dialogic tasks, balances the scales of automation and human touch.

*Addressing concerns head-on* (based on this chapter's Sections 1.3 and 2.3): As underscored by the discrepancies in the desirability evaluations between funders and nonprofits, it is evident that perceptions about AI vary widely. Open forums and discussion platforms must be created to address these differential views. The hesitance, predominantly stemming from the unknown, can be mitigated by transparent communication. Encouraging dialogue ensures that concerns, be it about biases, ethical quandaries, or the fear of reduced human touch, are tackled proactively.

In essence, the roadmap for integrating AI into Swiss philanthropy should be an orchestrated blend of innovation, collaboration, transparency, and continuous feedback. Leveraging AI's strengths while safeguarding against its pitfalls is not just an aspiration; with the right strategy, it is a tangible reality.

## Notes

- 1 It should be noted here that Rogers does not distinguish phases. He refers to the Stages of Change Model by James O. Prochaska et al. (1992) with the five phases: Knowledge, Persuasion, Decision, Implementation, and Confirmation. However, he does not divide the diffusion of innovations into phases.
- 2 We infer this particular significance, especially because Experimentation, that is, getting people to take risks, represents by far the most significant challenge; see Kane et al. (2019, p. 201).
- 3 On p. 326 the report shows that confidence levels correlate with education level (45%–59%).

## References

- AI Index Steering Committee (2023). *Artificial Intelligence Index Report 2023*. Stanford Institute for Human-Centered Artificial Intelligence (HAI). Retrieved from <https://aiindex.stanford.edu/report/>
- Benaich, N., et al. (2023). *State of AI Report 2023*. Retrieved from <https://www.stateof.ai/>
- Christensen, Clayton M. (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press, Boston, MA.
- Dweck, C. S. (2006). *Mindset: The New Psychology of Success*. Random House, New York.
- Fichman, Robert G., & Kemerer, Chris F. (1999). The Illusory Diffusion of Innovation: An Examination of Assimilation Gaps. *Information Systems Research*, 10(3), 255–275.
- Gerber, J. (2019, September 17). *How to Prototype a New Business*. IDEO U. Retrieved from <https://www.ideo.com/blogs/inspiration/how-to-prototype-a-new-business>
- Kane, Gerald C., Phillips, Anh Nguyen, Copulsky, Jonathan R., & Andrus, Garth R. (2019). *The Technology Fallacy: How People Are the Real Key to Digital Transformation*. MIT Press, Cambridge.
- Pfeffer, Jeffrey, & Sutton, Robert I. (2000). *The Knowing-Doing Gap: How Smart Companies Turn Knowledge into Action*. Harvard Business School Press, Boston, MA.
- Prochaska, James. O., DiClemente, Carlo C., & Norcross, John C. (1992). In Search of How People Change: Applications to Addictive Behaviors. *American Psychologist*, 47(9), 1102–1114.
- Rogers, Everett M. (2003). *Diffusion of Innovations* (5th ed.). Free Press, New York (Original Work Published 1962).
- Rudd, Rima E., & Simonds, Vanessa W. (2016, August). Diffusion of Innovations. *Encyclopedia Britannica*. Retrieved from <https://www.britannica.com/topic/diffusion-of-innovations>
- Walch, Kathleen (2020, January). Is There a Difference between Assisted Intelligence Vs. Augmented Intelligence? *Forbes*. Retrieved from <https://www.forbes.com/sites/cognitiveworld/2020/01/12/is-there-a-difference-between-assisted-intelligence-vs-augmented-intelligence>

# 5

## APPLYING DIVERSE AI TOOLS TO TRANSFORM PHILANTHROPIC OPERATIONS

Insights from the for-profit sector

*Prity Khastgir and Shweta Shalini*

### 1 Introduction

For philanthropic organizations (hereafter, POs) to fulfill their missions and generate positive change within their communities, they must effectively attract and retain donors and demonstrate resource efficiency through better internal operations. Donor retention and improved internal operation synergize also to increase their social impact. However, navigating the intricacies of donor engagement presents a significant challenge, characterized by evolving donor preferences (Ferris, 2021), shifting societal trends, and resource constraints. Moreover, the landscape of fundraising is undergoing a vital transformation, driven by shifts in donor behavior (Philanthropy, 2022). Donors increasingly utilize digital tools and platforms to contribute to charitable causes, reshaping how they engage with POs, exhibiting a heightened expectation for personalized interactions and a desire for transparency regarding the impact of their contributions. In response to these complexities, this work delves into the intersection of fundraising, donor engagement, and AI technology within the philanthropic sector.

Then, the work explores the integration of AI in human resources, to innovate the recruitment processes, upskill the workforce, and enhance operational efficiency. Furthermore, the transformative potential of Decentralized Autonomous Organizations (DAOs) and blockchain technology (Santana & Albareda, 2022) in philanthropy with the integration of self-executing smart contracts is explored, highlighting their role in enhancing transparency, efficiency, and inclusivity in resource allocation (Ahmed et al., 2024), thus reducing the need for intermediaries and minimizing vulnerabilities to corruption or financial mismanagement. The chapter finally looks at the intersection of AI-enhanced philanthropy to increase social impact by advancing the Sustainable Development Goals (SDGs) through technological innovation.

Through an overview of available technologies, this chapter highlights the many uses of AI to enhance POs' internal and external operations, to ultimately increase their impact. It underscores the significant impact of AI in reshaping operational landscapes and advancing social good initiatives with transparency and efficiency, covering fundraising, engagement, Human Resources (HR), governance, and social impact via the Sustainable Development Goals (SDGs).

## **2 The many possibilities of AI for philanthropic operations**

Central to philanthropy, donor and volunteer data can yield substantial insights. By prominently deploying Machine Learning (ML) algorithms that analyze past donation patterns and volunteer activities to estimate novel predictive models, the POs can proactively reach out to potential donors (Alkamoua, 2023) and volunteers, tailoring their outreach efforts to match individual interests and preferences. This facilitates POs to execute micro-campaigns with a greater return rate and begin modeling multiple primary factors that help gain more participants. AI can help identify beneficiaries who need support based on demographics, socioeconomic status, and past interactions with the organization. Overall, streamlining operations for non-profit organizations leads to a notable increase in fundraising efficiency.

Additionally, by using predictive analytics and segmentation processes, POs can allocate resources more effectively. Segmentation (Hsu et al., 2021) allows organizations to focus on individuals with a high likelihood of becoming committed donors. Through the continuous refinement of contributor lists by algorithms, the guesswork traditionally associated with segmentation is eliminated. Integrating AI technology with organizational databases facilitates the selection of contributors based on propensity scores (Rosenbaum & Rubin, 1983), enabling the creation of tailored messages for targeted audiences. This targeted approach enhances conversion rates and donations by engaging genuine contributors. Not to mention, propensity scores can be calibrated to assess donor affinity and capacity and significantly streamline the identification of suitable contributors. By examining historical data from past fundraising campaigns and evaluating information about potential donors and their interests, AI algorithms can predict the potential success of future initiatives. This analytical insight enables charities to customize their fundraising strategies, ensuring optimal resource allocation and maximizing the effectiveness of their efforts.

AI also presents a diverse strategy to combat scams in philanthropy by offering sophisticated tools for detecting fraudulent behavior, validating organizational legitimacy, and improving transparency. Despite existing challenges, the integration of AI analytics with philanthropic values can successfully protect donors, strengthen trust, and ultimately enhance the ability of philanthropic endeavors to make a positive difference. AI algorithms can swiftly identify anomalous activities (Chalapathy & Chawla, 2019), such as substantial contributions from unfamiliar sources or an unusual surge in transaction frequency. This proactive identification empowers POs to thwart potentially damaging fraudulent donations. By analyzing historical data and recognizing patterns indicative of fraudulent behavior (Olubusola Odeyemi et al., 2024), AI algorithms can effectively flag anomalies in donation behavior, transaction patterns, or organizational attributes, alerting POs to potential charity scams.<sup>1</sup> In addition, it can verify the authenticity of both donors and organizations through biometric authentication and document verification. POs can prevent unauthorized entities from posing as legitimate charities, thus safeguarding against potential fraud and misuse of funds.

Next, AI-driven sentiment analysis provides POs with valuable insights into the reputation and public perception of charitable entities. By analyzing online discussions, reviews, and social media interactions, AI algorithms can gauge the sentiment surrounding specific charities and assess their credibility in the eyes of the public, thus catalyzing trust and accountability in philanthropy.

In summary, the convergence of AI and philanthropy presents diverse benefits that surpass traditional practices within charitable sectors. AI can raise the capability to personalize donor engagement and communication strategies by utilizing customized landing pages and advanced data analytics to cater to donors' preferences and simultaneously increase their conversion rates by engaging smaller donors as well. The integration of AI in fraud detection, customer service,



predictive analytics for fundraising initiatives, and impact assessment signifies a strategic path toward improved operational effectiveness, informed decision-making, and donor involvement. Thereby, POs embrace a new era marked by increased efficiency, transparency, and alignment with their stakeholders.

### 2.1 Fundraising, operations, and outreach

AI can be employed to enhance donor- and visitor-related activities, for example, foundation’s automating operations such as Shakespeare Birthplace Trust.<sup>2</sup> *Fundraising efforts* can benefit from AI-driven analytics, which can identify potential donors based on their past giving behavior and preferences, enabling targeted outreach campaigns. *Donor outreach* can be enhanced through AI-powered chatbots (Ayanouz et al., 2020), which can provide instant responses to inquiries and offer personalized recommendations for engagement. Especially for the Shakespeare Trust AI-powered chatbot would be very helpful for answering questions and providing information about the trust’s mission and programs.

In practice, the algorithms can analyze vast datasets of donor behaviors, preferences, and trends, enabling foundations to tailor their strategies for more effective *fundraising campaigns*. One example would be a patent application filed by Stamler and Vanvalkenburgh (2023) bearing application number US20230342853A1 for an automated electronic impact platform (see Figure 5.1).

## Impact Platform 10

### Charitable Donations

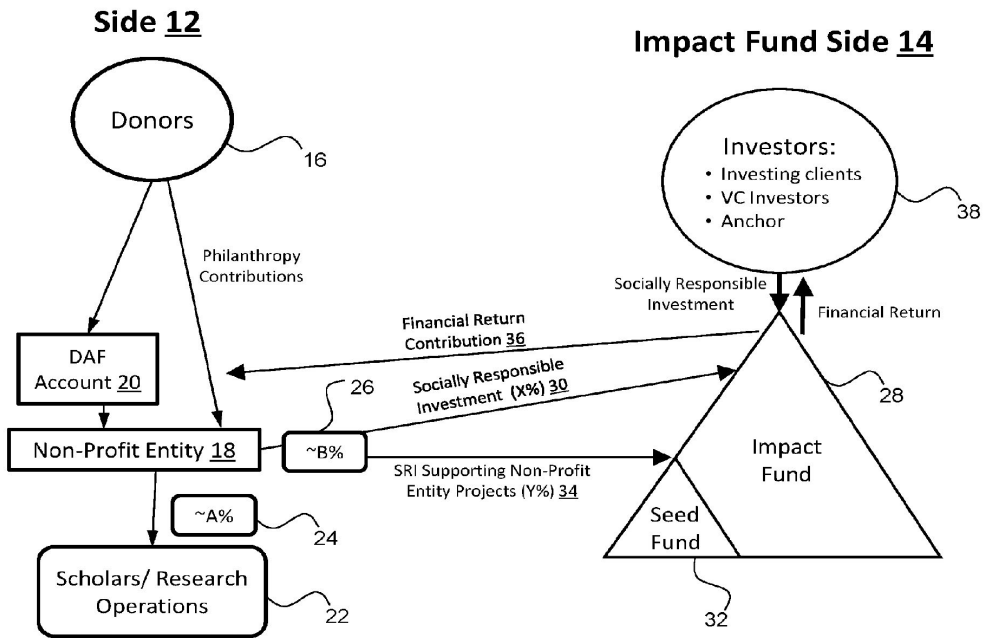


Figure 5.1 Schematic representation of an automated electronic impact platform for sustainable investment in socially responsible endeavors (photograph from the patent application).

The platform system allocates a portion of philanthropic contributions to directly finance charitable activities associated with these initiatives, while another portion is directed toward investing in an impact fund focused on commercial ventures aligned with the same socially responsible endeavors. As the impact fund generates profits, these financial returns are reinvested into the contribution stage. Further, the financial returns are divided into two portions: funding charitable activities and reinvestment in the impact fund to support further commercial endeavors related to the cause. However, the platform system fails to acknowledge the significance of considering the ethical, societal, and human implications of combining charitable funds with impact funds and presents a rather complex landscape. Firstly, the amalgamation of these funds entails a convoluted system characterized by algorithmic fund allocation, computational simulations, and dynamic portfolio adjustments. Furthermore, this approach may fail to account for the evolving needs of projects and donor preferences, risking the misallocation of resources and undermining the efficacy of the initiative. In practice, the reliance on computational simulations to forecast returns and assess investment risks introduces an element of uncertainty and bias, highlighting the need for robust risk management strategies. Secondly, the rigidity inherent in the fixed allocation of funds may impede adaptability to changing contexts, compromising the system's agility in responding to dynamic circumstances.

### *2.1.1 VR-enhanced fundraising and outreach*

Virtual Reality (VR) modules immerse participants in lifelike scenarios, enhancing an experience's realism. VR can immerse donors in the physical environments where charitable operations are conducted, providing them with a firsthand experience of the realities associated with the cause. Such immersive experiences foster a heightened sense of presence and empathy among donors, thereby potentially increasing their inclination to contribute generously to the cause (Kandaurova & Lee, 2019; Sooter & Ugazio, 2023). Compared to a static image, VR – due to its immersive nature – increases realism by allowing the user to become a part of the environment, thereby delivering multiple communication cues and experiential information. These simulations are invaluable tools to nurture essential soft skills such as empathy, communication, and problem-solving, enabling donors, volunteers, or staff to interact more effectively with beneficiaries and stakeholders of any philanthropic cause.

Furthermore, VR platforms offer a dynamic medium to showcase the impact of charitable donations and initiatives in a captivating manner. By providing a visual representation of the outcomes of philanthropic efforts, these platforms inspire increased support and engagement from donors and the wider community. Situations where physical access is limited due to limiting or dangerous factors such as war or geographical distance is another essential use case for VR tools in the philanthropic sector. For example, *Charity: Water*, through their initiative “*The Source*” (Green, 2018), transported 400 attendees on a journey to East Africa during an annual black-tie fundraising event at the Metropolitan Museum of Art. The VR tools enabled participants to immerse themselves in a virtual reality film depicting the transformative impact of clean water access for the first time on the life of a 13-year-old girl and her family. The donors that day alone committed \$2.4 million (Green, 2018) to the event.

Another similar concept is that of a *digital twin* – a virtual model that near-precisely replicates real-world systems or environments (Ukko et al., 2022). By integrating data from various sources, including on-ground sensors, social media, and transaction records, POs can build digital twins of the on-the-ground reality of the causes they support, gaining real-time insights into their

initiatives' progress, challenges, and outcomes. This serves to enhance monitoring and evaluation, thereby improving impact assessment. They could offer donors access to the digital twin, allowing them to virtually experience the impact of their contributions and witness the progress of projects in real time. This can increase the connection between donors and their impact, improving loyalty and support for the cause.

In essence, this transformative capability enhances the effectiveness of various philanthropic activities, such as conducting *virtual community assessments*, organizing *immersive fundraising events*, or delivering *virtual aid and support services to beneficiaries*. These events can offer donors a more engaging and interactive experience, enabling them to participate in virtual tours of project sites, witness the impact of their contributions firsthand, or engage in virtual reality-based fundraising games and activities. Adopting VR technologies can significantly amplify the efficacy and influence of philanthropic endeavors, catalyzing positive change and advancing social progress on a broader scale. Through these immersive experiences, all stakeholders can develop greater empathy and understanding for the communities they serve.

### 2.1.2 NLP-powered virtual assistants for operations and engagement

Natural Language Processing (NLP)-powered chatbots and virtual assistants contribute to simplifying operational and organizational processes, for example, **VAIOT**,<sup>3</sup> which provides AI assistants for sales, marketing, customer service, and legal services. These assistants are poised to play an essential role in the evolving digital realm, offering personalized support, content generation, and smart companion capabilities. Similarly, **Kore.AI**<sup>4</sup> offers a comprehensive suite of features to enhance customer engagement and streamline operational workflows. This ensures seamless communication across diverse channels, ensuring consistent and effective customer service delivery regardless of the chosen communication medium. The platform's automation features alleviate the burden of repetitive tasks and workflows, such as data entry and appointment scheduling. Moreover, Kore.AI prioritizes security and compliance – from data encryption to access controls and audit trails; the platform upholds stringent security protocols to fortify data integrity and confidentiality, which are paramount in the non-profit sector which is often the target of cyberattacks.

POs can leverage tools like VAIOT and Kore.AI to improve not only their *internal administrative operations* but also their *engagement with external stakeholders*, whether visitors, donors, clients, or beneficiaries, dispensing critical advice and information. Conversational interfaces offer real-time interactions that cater to donors' needs, queries, and donation preferences, fostering immediate engagement, reinforcing donor loyalty, and strengthening the bond between donors and organizations. Such a process can also facilitate grant applications by guiding applicants through forms, verifying information, and providing real-time updates, expediting the process, and enhancing the overall user experience. Such chatbots can be deployed across multiple platforms, such as websites, social media, and messaging apps. Then, albeit not directly a virtual assistant or a chatbot, using NLP to automate large-scale natural language aggregation from across the web and leveraging "collective intelligence" (Lee & Jin, 2019) capitalizes on pooling insights from large cohorts (often termed the "wisdom of crowds"). By capturing information from collective intelligence, POs could identify "under-the-radar" groups or organizations operating within specific causes or localities, proactively reaching out rather than solely awaiting grant applications. This can help engage and connect organizations to their causes and surpass barriers imposed by geographical region and language.

## **2.2 Optimized HR practices**

AI integration into HR has become increasingly prevalent in all sectors. POs face challenges similar to those of their corporate counterparts, including the need to efficiently source, recruit, and assess candidates while mitigating selection bias. AI-driven recruitment solutions offer promising remedies to these challenges by automating and enhancing critical procedures such as candidate sourcing, screening, and interviewing.

Like other sectors, the non-profit sector incurs significant expenses during the recruitment process. Consequently, organizations may prioritize the retention of existing talent as a more economically viable strategy compared to the endeavor of replacing personnel. One approach to achieve this retention objective involves exploring opportunities to redistribute role responsibilities among current staff members. Concurrently, organizations may offer training initiatives to equip existing employees seeking skill expansion and professional growth with the requisite capabilities. Such measures can be augmented by incorporating AI technologies, facilitating streamlined processes and enhancing operational efficiencies within POs. AI can curate tailored learning materials, aligning with individual roles, preferences, and skill gaps to ensure access to pertinent content and further enhance the learning experience by recommending courses, articles, and resources, fostering engagement and customization. Additionally, AI can conduct assessments to evaluate employee skills and knowledge, offering valuable insights into areas necessitating improvement.

In addition, recruitment teams within the POs encounter challenges in handling high volumes of applications, mitigating biases in candidate selection, and coping with a scarcity of qualified applicants. AI-driven recruitment remedies these obstacles by automating and enhancing critical recruitment procedures such as candidate sourcing, screening, and interviewing. For example, platforms such as **LinkedIn Recruiter**<sup>5</sup> have gained significant traction within the philanthropic sector, enabling POs to identify and engage with potential candidates more effectively. Such platforms streamline talent scouting by leveraging advanced algorithms to sift through vast databases of user profiles and analyze their data, including work experience, skills, and endorsements, to identify potential candidates matching specific job requirements. By automating candidate searches and recommendations, AI-powered platforms streamline the talent acquisition process for POs, allowing them to identify and engage with top talent efficiently.

### *2.2.1 Transferable tools – Unilever’s recruitment and training approach*

AI tools extend beyond *recruitment* and into *employee training*, where simulated work scenarios allow employees to refine their skill sets within controlled environments. These simulations, tailored to individual employee needs, provide realistic experiences that enhance learning outcomes. These simulations exhibit adaptive proclivity, tailoring experiences to the requisites of individual employees. Taking Unilever as an example (Hu, 2023), in 2016, they partnered with **HireVue**<sup>6</sup> and **Pymetrics**<sup>7</sup> to create an AI-driven system for recruitment, selection, and onboarding. It utilizes a Natural Language Processing (NLP) bot named **Unabot** to streamline the employee orientation process and gather essential insights, effectively addressing their queries. This orientation approach yields valuable insights into new employees’ primary concerns, enabling recruiters to enhance external job postings based on internal feedback to meet applicant expectations. POs can adopt similar tools not only for internal employee training but also to facilitate temporary volunteers’ training queries.

**Pymetrics** employs a comparative approach, aligning candidate data with job requirements and successful employee profiles within Unilever to enhance recruitment accuracy. Prospective

candidates are tasked with understanding Unilever's criteria through gamified elements, reshaping recruitment into an interactive experience focusing on key competencies rather than conventional qualifications. Additionally, Unilever introduced video analysis software for candidates moving on to the second interview stage. The analysis is based on scrutinizing facial expressions, body language, and linguistic cues. Subsequently, an ML algorithm evaluates these inputs using NLP and body language analysis to assess candidate suitability. This approach transforms the recruitment process into an interactive experience, focusing on key competencies rather than conventional qualifications. By integrating gamification, recruitment becomes more dynamic, providing immediate feedback and incentives to sustain applicant engagement. The example of Unilever can be utilized by the POs that can deploy video analysis software, such as **HireVue's** platform, to streamline the hiring process.

Nevertheless, adequate due diligence must be undertaken to ensure such tools have adequately addressed issues surrounding algorithmic bias. While Unilever has implemented video analysis software for candidate assessments, this approach has inherent difficulties. Challenges of using algorithmic body language analysis to assess candidate suitability involve complexities in accurately interpreting non-verbal cues and potential biases in the evaluation process. While algorithms are not inherently biased, the complexity of algorithmic recruitment often results in unintended biases. Mitigating bias in AI-driven recruitment ML algorithms may be influenced by biases in the training data, leading to unjust assessments of candidates based on gender, ethnicity, or age (Chen, 2023). To address this, it is crucial for POs to develop comprehensive strategies to understand that dataset construction, target formulation, and feature selection play pivotal roles in shaping algorithmic bias and to mitigate bias at each stage of the process. Integrating AI into decision-making raises concerns about biases and privacy implications that necessitate thorough evaluation and human intervention.

The integration of AI into recruitment processes within POs presents both opportunities and challenges. By leveraging AI-driven solutions, organizations can optimize talent acquisition processes, aligning candidate data with job requirements and successful employee profiles. However, using algorithmic body language analysis and machine learning algorithms necessitates thorough evaluation and strategies to mitigate bias and ensure privacy. By addressing these challenges, POs can harness the power of AI to enhance recruitment accuracy and drive positive social impact.

### *2.2.2 Empowering skill development with AI-enhanced VR simulations*

Numerous companies utilize AI-driven training simulations using VR modules to immerse employees in lifelike scenarios to enhance practical skills in a risk-free environment. Employees can practice responding to practical challenges in real time, such as active listening, conflict resolution, and understanding the diverse perspectives of fellow employees. The AI simulations can incorporate various perspectives, characters, and crises that require empathy and effective communication under pressure.

VR coupled with AI promotes a personalized learning approach by crafting an immersive learning environment for the users to upskill themselves. Take as an example **Zenarate**,<sup>8</sup> a platform that prioritizes soft skill enhancement through a diverse range of techniques encompassing visual, auditory, and kinesthetic learning modalities. The platform furnishes instantaneous feedback to recipients, catering to auditory learners by accommodating their preferred learning mode while emphasizing the significance of tone. Zenarate harnesses AI capabilities for kinesthetic learners to formulate practical, scenario-driven role-playing exercises for interactive and experiential learning. However, these customized training modules of the platform may not completely adapt the content to align with individual employees' precise needs and learning preferences. Another useful

platform designed to provide hands-on training for technical skills is **Transfr**'s<sup>9</sup> VR training platform. Transfr's is designed to provide hands-on training for technical skills, focusing on vocational career exploration and pre-apprentice training. The platform offers a range of VR simulations that allow employees to practice and master technical skills in a safe and controlled environment. Thereby, using VR in training in any organization can improve learning outcomes, increase efficiency, and reduce risk.

In the philanthropic landscape, integrating such tools presents a major shift in how philanthropic organization employees engage in their work. Using platforms such as Zenarte, staff can practice crisis response communication and conflict resolution skills, empathy, and engagement with multicultural and diverse scenarios. In contrast, platforms like Transfr's can enable practical manual or technical skills (for example, dexterity needed in de-mining activities or other manual labor charity work). Combining the two, POs' staff can be better trained to respond to challenging situations they might not have encountered before, enhance the effectiveness of philanthropic goals and, above all else, reduce risks due to better preparedness.

### *2.2.3 Improving diversity and inclusion*

The application of AI in performance management systems aligns with the philanthropic goal of promoting fairness, inclusivity, and continuous employee development. Funding or endorsing projects that integrate AI-powered technologies into organizational processes will contribute to establishing more equitable and inclusive work environments. Specifically, they can provide financial assistance to research institutions or startups developing AI algorithms that mitigate bias in performance evaluations. Furthermore, POs can collaborate with businesses to test and implement AI-driven performance management systems (Jha, 2023), fostering a workplace culture of fairness and inclusivity. AI can identify and rectify unconscious biases, resulting in fairer and more impartial evaluations and enhancing overall decision-making processes. It thus ensures that all qualified candidates have equal access to job opportunities.

An example tool addressing Diversity and Inclusion initiatives is **Textio**,<sup>10</sup> which harnesses NLP to ensure that job descriptions resonate with diverse and qualified candidates. Textio analyzes language utilized in job postings, emails, and employer branding content to identify patterns that may deter women and minorities from engaging or applying to those specific jobs. Textio scrutinizes language patterns and historical data, equipping users with foresight into how effectively their text will resonate with targeted audiences. Users can make informed adjustments to ensure their messaging aligns with diversity and inclusion objectives by identifying and flagging areas prone to bias or exclusionary language. This language barrier removal, which may otherwise hinder engagement from diverse candidates, increased the applicant pool to a more diverse group. By proactively removing bias from communication systems, Textio is pivotal in fostering more inclusive and diverse hiring practices. Its capabilities extend to predicting the performance of job descriptions and offering recommendations to enhance their effectiveness.

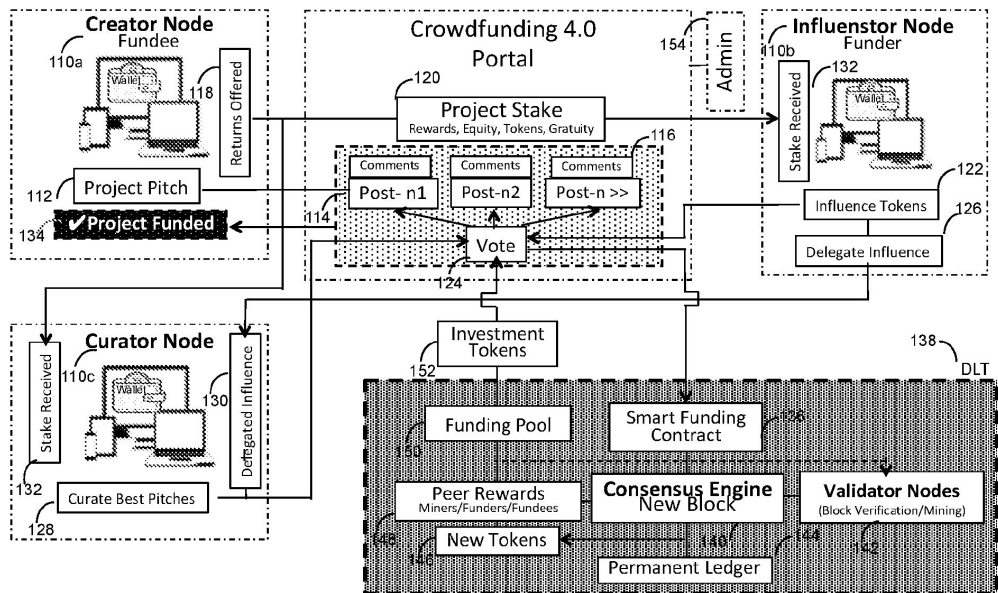
However, organizations employing AI-powered tools in their hiring processes must prioritize strategies to prevent the perpetuation of existing biases. It is crucial to rigorously assess the quality and representativeness of the data used to train AI algorithms, given that incomplete or biased datasets can inadvertently reinforce existing prejudices. Equally, **transparency and accountability** are paramount in developing and deploying AI algorithms. By understanding how these algorithms make decisions, organizations can effectively identify and rectify biases and establish mechanisms for algorithmic accountability, ensuring that AI systems uphold fairness and equity in hiring processes. Regular audits and monitoring are essential to continuously evaluate AI tool performance,

allowing organizations to address biases as they arise swiftly. Next, **diversity** within development teams creating and implementing AI tools is another critical factor that has been shown to reduce the risk of bias (Pantelakis, 2023) in algorithm design and implementation. Assembling teams with diverse perspectives helps reduce the risk of bias in algorithm design. Implementing bias detection and mitigation techniques, such as debiasing algorithms and adversarial testing (Alabdulmohsin & Lucic, 2022), helps organizations combat biases in real time, fostering fair and equitable hiring outcomes. Lastly, involving HR professionals, recruiters, and hiring managers in the development and deployment of AI tools promotes more awareness of potential biases while equally providing ongoing training on bias recognition and mitigation to equip stakeholders with the necessary skills.

PO’s role in this space is multi-faceted. By funding due diligence, regular audits by human intervention of the different algorithms and processes, and bias mitigation efforts, POs will be able to identify and rectify biases or inconsistencies in AI systems. Moreover, they can contribute to establishing diverse development teams with a broad range of expertise and perspectives. Stakeholder engagement and training programs can be funded to educate hiring managers, recruiters, and other personnel about the importance of diversity and equity, equipping them with the necessary tools and knowledge to identify and address biases effectively.

### 3 Blockchain and DAOs for improved governance

International patent application WO2020141360 (Raheman, 2020) titled “Crowdfunding 4.0: a novel influence-based global fundraising platform and system” (see Figure 5.2) discusses how a decentralized crowdfunding platform would operate between fundee, project creators, funders, and influencers. This unique approach facilitates the sharing and monetization of influence among



**Influence-Based Crowdfunding Network Architecture**

Figure 5.2 Block diagram illustrating the network architecture of the next-generation influence-based crowdfunding infrastructure (photograph from the patent application)

peers, thereby raising funds and kick-starting projects without the need for traditional monetary contributions. This platform allows participants to leverage their influence to generate or mine funds in various tokens or cryptocurrency denominations, eliminating the requirement for cash donations from funders to beneficiaries, through self-executing smart contracts.

Participants on the platform, whether project creators, funders, or influencers, engage in activities such as pitching projects, voting on funding pitches, or delegating their influence on one or more project curators. It provides a space for individuals to pitch their projects and share ideas while democratizing fundraising and reducing barriers to entry for marginalized communities. Project creators are empowered to access funding opportunities by leveraging their influence and engaging with the platform’s community. These interactions are tokenized based on the participants’ stake in the platform’s token economy, measured in cryptocurrency tokens, hashing power, reputation scores, intellectual property ownership, or platform activities. The platform utilizes distributed ledger technology (DLT) to ensure security, privacy, and anonymity, with consensus protocols such as proof-of-work or proof-of-stake validating transactions and smart contracts. A self-executing smart contract defined within the platform facilitates funding agreements between fundees, funders, and curators. This contract outlines the terms for funding projects through votes and defines the delegation of influence between funders and curators. Additionally, the platform’s consensus engine is crucial in verifying and validating smart contracts, transactions, and events through peer nodes, adding them to the DLT or blockchain’s permanent ledger (Figure 5.3).

Incorporating the principles of decentralized finance (DeFi) and blockchain, this crowdfunding platform revolutionizes traditional fundraising methods by empowering individuals to contribute to projects through their influence rather than monetary donations. By tokenizing influence and

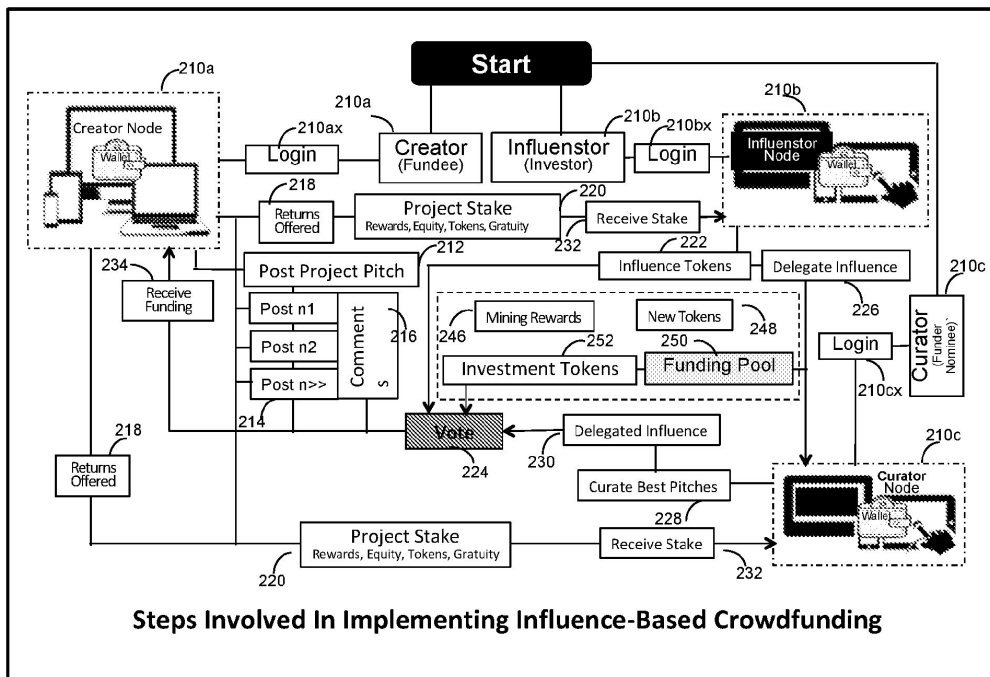


Figure 5.3 Block diagram illustrating the steps involved in implementing the influence-based crowdfunding method (photograph from the patent application).



leveraging smart contracts, the platform promotes transparency, efficiency, and inclusivity in fundraising. Further, this crowdfunding platform acts as a catalyst by providing a level playing field for individuals from all walks of life to access funding and support for their projects. Additionally, in regions characterized by inadequate banking infrastructure or prevalent corruption, the adoption of blockchain-driven fundraising mechanisms, as exemplified in the international patent application WO2020141360, presents substantial benefits for POs. The decentralized architecture of blockchain technology furnishes a transparent and secure fundraising environment, diminishing corruption risks (e.g., see Sarker et al., 2021) and ensuring the direct flow of donations to their intended recipients. Through the integration of blockchain solutions, POs operating in such environments can surmount the constraints of conventional banking systems and mitigate the perils associated with corrupt practices.

A pivotal advantage of blockchain-based fundraising platforms lies in the transparency and accountability of transactions (Almaghrabi & Alhogail, 2022). With all financial activities recorded on a public ledger, donors gain assurance that their contributions are utilized as intended. Blockchain technology functions as a safeguard against fraud and corruption by ensuring that records of transactions are encrypted and stored across a network of computers. This decentralized approach prevents unauthorized alterations or theft of sensitive information, thereby fostering greater transparency and accountability in managing charitable funds. Furthermore, blockchain technology's utilization can enhance fundraising accessibility for marginalized populations. Blockchain-powered platforms empower individuals from diverse backgrounds to secure funding for their initiatives by democratizing the fundraising landscape and lowering entry barriers. This inclusive approach not only addresses disparities but also fosters social and economic progress in underserved communities.

A notable example of blockchain's impact on philanthropy is demonstrated by Dublin-based startup **AID:Tech**.<sup>11</sup> Recognizing the need for enhanced integrity in charitable contributions and social welfare payments, AID:Tech developed a groundbreaking platform built on blockchain technology. By leveraging blockchain's inherent transparency and security features, AID:Tech's solution empowers charitable organizations and governments to effectively manage and track the flow of funds, thereby minimizing the risk of fraud and mismanagement.

In essence, integrating blockchain-driven fundraising tools can transform philanthropic endeavors in regions grappling with deficient banking infrastructures or pervasive corruption. By furnishing a transparent, secure, and equitable fundraising ecosystem, these tools empower POs to amplify their impact and drive positive societal transformations within their communities.

Moving on, **Decentralized Autonomous Organizations (DAOs)** are poised to revolutionize traditional community and business operations globally and offer a decentralized platform for individuals and stakeholders to collaborate based on shared rules encoded on the blockchain. In the context of philanthropy, DAOs present an innovative model for pooling and distributing cryptocurrency using blockchain, functioning as a transparent and community-run financial infrastructure and governance instrument. *Philanthropic DAOs*<sup>12</sup> leverage decentralized decision-making power to enable the transparent and efficient distribution of funds to support social causes and impact-driven projects. The deployment of DAOs to existing blockchain has several advantages, and DAOs can reduce the functional expense and delays associated with traditional charities; automate and streamline multiple steps like fundraising, distribution, and reporting using smart contracts; and provide a democratic, transparent voting mechanism to the peers in the node.

United States patent application 20220391797 titled "*Distributed platform for the development of attracting and scaling innovation*" (Stein et al., 2022) describes a distributed platform that can be utilized to foster innovation at scale, attracting various stakeholders, including DAOs, to

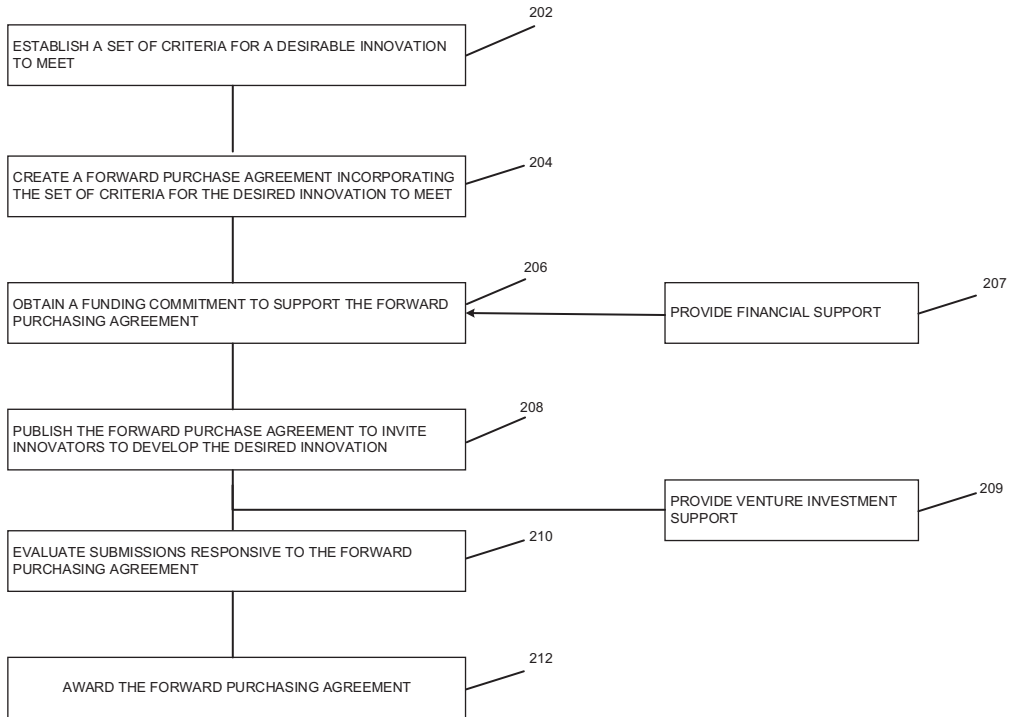


Figure 5.4 Flow diagram illustrating the method of developing, attracting, and scaling an innovation (photograph from the patent application).

contribute to the development and scaling of innovative solutions. Mainly, the platform serves as a mechanism for promoting partnerships among diverse entities (Figure 5.4).

The secure distributed system is designed to facilitate the division of collaborative efforts among multiple entities into future obligations and potential rewards. The system can process these obligations and rewards, consolidating them into a legally binding forward purchase agreement. The obligations and rewards outlined within the agreement can be segmented, encoded, and executed as a smart contract within a distributed ledger system. The system’s capabilities encompass a series of interconnected processes, beginning with the receipt of calls for innovation from commercialization entities. Innovators then submit proposals aligned with these calls, which are evaluated against predefined goals, such as SDGs. However, incorporating the functionality of the forward purchase agreement into a DAO amplifies the system’s reach and impact. Notably, this DAO integration empowers individuals and investors to engage with the agreement by acquiring tokens, thereby securing voting rights within the organization. These voting rights allow participants to influence the selection process for innovation submissions linked to the original forward purchase agreement and potentially subsequent agreements tied to it. Through DAO, selected innovations can garner optional financial support from participants and eventual commercialization. This collaborative approach fosters a dynamic ecosystem where stakeholders actively contribute to the advancement of promising innovations.

DAOs present a novel avenue for aggregating and distributing funds to foster accountability in the philanthropic sector. Philanthropic DAOs apply the benefit of decentralized decision-making

to facilitate transparent and efficient allocation of resources. By integrating DAOs within established blockchain frameworks, philanthropic entities stand to gain from reduced operational costs, streamlined procedures, and heightened transparency to build trust. Through using smart contracts, DAOs automate and optimize various stages of the philanthropic cycle, encompassing fundraising, distribution, and reporting, while offering a democratic and transparent voting system for stakeholders. Through their innovative governance models and collaborative ethos, DAOs can redefine the operational landscape of POs and amplify their societal impact.

#### 4 AI and philanthropy for the SDGs

Based on POs' improved internal operations' efficiency and governance, using tools and methods such as those presented in Sections 2 and 3, POs stand in a unique position to enhance their social impact. Philanthropy and the Sustainable Development Goals (SDGs) are intricately interconnected, with philanthropic efforts playing a crucial role in advancing progress toward these goals. AI can play a significant role in contributing to most SDGs (Nahar, 2024), from *SDG1 – No Poverty*: using AI-enhanced satellite imagery analysis to predict poverty risk (Hall et al., 2023), to *SDG17 – Partnership for the Goals*: automatically deriving partnership suggestions between organizations with similar missions (Tudor et al., 2024). Although of equal importance, it can also hinder progress on some SDGs where data is not yet adequate to create unbiased and reliable models (Vinuesa et al., 2020).

Various initiatives and organizations are leveraging AI to accelerate progress toward the SDGs. For example, Google has launched a \$25 million<sup>13</sup> open call for organizations using AI to advance the SDGs, and the AI for Good Foundation is bringing together the best minds and technologies to solve urgent global challenges. Interestingly, the Sustainable Development Goals Philanthropy Platform (SDGPP),<sup>14</sup> led by the United Nations Development Programme (UNDP)<sup>15</sup> and various POs, provides real-time data and information on the initiatives and solutions that funders support for each SDG. These efforts demonstrate the potential of AI and philanthropy in driving positive social impact through the SDGs.

A leading philanthropic actor in this space is the AI for Good Foundation. The foundation is determined to help facilitate the achievement of the goals through various projects such as the Climate Trend Scanner<sup>16</sup> and the SDG Data Catalog.<sup>17</sup> The Climate Trend Scanner is an initiative commissioned by the United Nations Development Programme (UNDP) to monitor and identify cutting-edge climate solutions worldwide in real time. This endeavor is part of the broader Climate + SDG Scanners initiative, aimed at tracking global progress toward sustainability. Developed by the AI for Good Foundation in collaboration with the Research Institute of Sweden (RISE) and BWA,<sup>18</sup> the Climate Trend Scanner employs advanced AI algorithms to sift through vast datasets on climate change, identifying emerging trends and innovations instantaneously. The project's primary objective is to support researchers and policymakers by providing deeper insights into the complex landscape of climate change through innovative research methods.

Another prominent example is The AI Forward Alliance (TAIFA)<sup>19</sup> that is dedicated to empowering 25 million girls and young women by equipping them with the necessary skills, knowledge, tools, and mentorship to comprehend, create, and implement AI models and transformative technologies (thereby tackling *SDG5 – Gender Equality*). This initiative aims not only to educate participants on the workings of AI models but also to enable them to develop and deploy their own machine learning models to address real-world challenges within their communities. Exponentially increasing the representation of girls and women in science and technology will meet the demand for future skilled jobs while impacting social well-being. Educating and empowering girls

and women yield significant positive impacts on entire communities and economies. Poverty can be mitigated by bolstering the earning potential of women while simultaneously fostering lifelong learning opportunities, stimulating innovation in every field, mitigating overall inequalities, and contributing to the realization of SDGs 1, 4, 5, 8, 9, and 10.

Overall, POs can invest in AI innovation hubs, research centers, and technology incubators to drive technological innovation, infrastructure development, and economic diversification in developing countries, enhancing, among others, *SDG9 – Industry, Innovation, and Infrastructure*. One such application is the international patent WO2021215906, entitled “*Artificial Intelligence-Based Method for Analyzing Raw Data*” (Samantaray, 2021), which presents a novel approach harnessing AI to facilitate computer software development (see Figure 5.5). The methodology involves a series of steps, starting with connecting multiple users through communication devices. These users are then assigned specific tasks based on provided instructions, and their activities and behaviors are systematically captured and stored as a comprehensive training dataset, forming the basis for subsequent stages. The next critical phase is to deploy the training of a deep reinforcement learning (DRL) ensemble neural network (ENN) using the amassed training data. Once trained, this DRL-ENN system becomes adept at executing automated tasks, including generating computer software. This AI-driven methodology presents a promising asset for POs. Using AI in software development, POs can optimize their processes, boost productivity, and amplify their impact on societal and environmental initiatives. In line with the same SDG, the Best Available Charitable Option (BACO) model, pioneered and developed by the Acumen Fund (2007), offers a framework for assessing the societal impact of philanthropic investments. The evaluation process quantifies the social benefits generated by each investment and compares them against various charitable options. BACO ensures the efficient allocation of philanthropic resources toward initiatives that drive innovation, spur industrial growth, and enhance infrastructure development. This strategic approach mirrors the overarching objectives of SDG9, which aims to foster inclusive and sustainable industrialization, stimulate innovative practices, and fortify infrastructure resilience.

## **5 Conclusion – the road ahead**

The intersection of AI and philanthropy presents many advantages for POs seeking to enhance their operational efficiency and social impact. The utilization of AI in customer service, predictive analytics for fundraising campaigns, and impact tracking mechanisms provide POs with the capability to make informed decisions, improve transparency, and foster trust among donors. Our work exemplifies that AI can play a crucial role in fraud detection, personalized messaging, and impact assessment, enhancing transparency and trust, marking a shift toward informed decision-making and amplifying donor involvement, thereby illustrating its transformative impact on philanthropic endeavors. In parallel, advancements in NLP and LLMs offer promising opportunities for POs to enhance their engagement, operations, and donor engagement. While prioritizing security measures to ensure data integrity and regulatory compliance, POs can deliver personalized experiences to all stakeholders.

With context to HR, AI deployment offers significant benefits for POs, including improved decision-making, predominantly through data collection and analysis for recruitment. This analysis serves to identify issues of bias, ensure judicious selection of candidates, and subsequently assist in the formulation of benefited structures. The strategic deployment of AI provides cost-effectiveness, expeditiousness, and quality augmentation. By prioritizing data quality assessment, transparency, and accountability in AI algorithms, POs can promote diversity within their teams, ultimately driving organizational growth and innovation. However, organizations need to remain

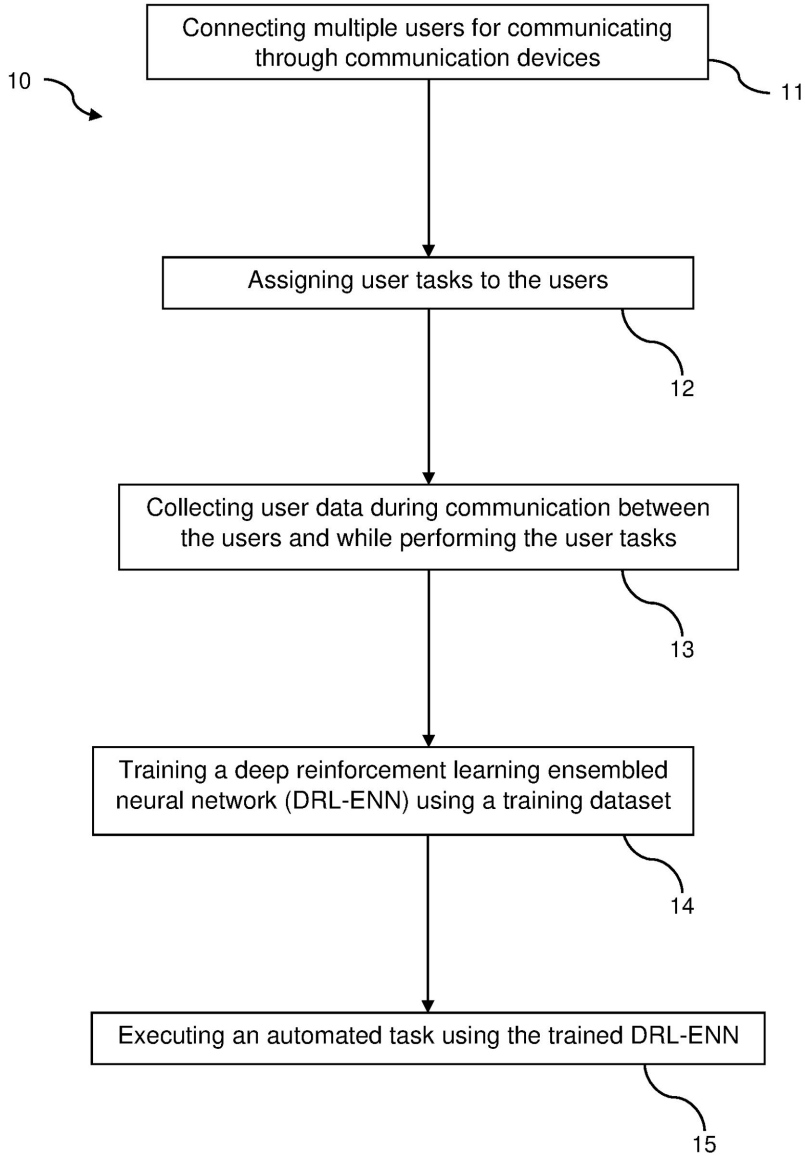


Figure 5.5 Flow diagram of the artificial intelligence-based method for processing raw data (photograph from the patent application).

vigilant about potential biases and ethical considerations associated with AI deployment, ensuring that their practices align with their philanthropic values and goals. Nevertheless, it is imperative to acknowledge that in philanthropic endeavors, the ethical considerations surrounding data privacy and cybersecurity are a cause of concern, especially as AI becomes increasingly relied upon for data management. Therefore, POs must prioritize building trust among stakeholders and safeguarding confidential data.

Our work has also highlighted that the utilization of DAOs significantly impacts the performance of crowdfunding platforms by enabling the exchange and monetization of influence among peers, facilitating fund generation and project initiation without conventional monetary inputs. Through the tokenization of influence and the deployment of self-executing smart contracts, the platform enhances transparency, efficiency, and inclusivity within philanthropic initiatives. Users can actively partake in project pitching, fund pitch voting, and influence delegation. It has been observed that users' engagements, which are tokenized according to their involvement in the platform's token ecosystem and supported by distributed ledger technology (DLT), are able to uphold their security, privacy, and anonymity. Additionally, it has been observed that this platform demonstrates the positive impact of deploying blockchain and leveraging DeFi principles. This will drive progress across sectors, while aligning with SDG10's goal of reducing inequalities and democratizing access to funding opportunities, thereby fulfilling philanthropic goals. Our chapter delved into the effects of the platform's consensus engine and its role in validating smart contracts, transactions, and events, enriching the fundraising landscape, and breaking down barriers for underrepresented communities to engage and thrive within the ecosystem. Further, we highlight the synergistic relationships between AI and blockchain's inherent immutability, which significantly enhances transparency and accountability within POs. This enables the creation of a traceable ledger of donations and transactions that adheres to their authenticity, thereby reducing the risk of funds being misappropriated, fostering trust and accountability among stakeholders. However, the strategic arrangement of AI's capabilities requires careful planning to fully unleash its future potential.

In essence, our findings underscore the importance for stakeholders and institutional investors to leverage the transformative capabilities of AI, DAOs, and blockchain to drive sustainable development and organizational advancement. Through strategic utilization of these technologies, POs can progress toward achieving the SDGs by 2030, paving the way for a future characterized by enhanced transparency, inclusivity, and impact in the philanthropic landscape. Moving forward, it is imperative for POs to embrace these technologies and explore their full potential in driving positive impact. Additionally, it is essential to continue investing in research and development to advance these technologies further and unlock new opportunities for growth and development.

## **Notes**

- 1 <https://dataprot.net/statistics/charity-scam-statistics/>
- 2 <https://www.shakespeare.org.uk/support-us/>
- 3 <https://vaiot.ai/en/vaiot-solutions>
- 4 <https://kore.ai/>
- 5 <https://business.linkedin.com/>
- 6 <https://www.hirevue.com/>
- 7 <https://www.pymetrics.ai/>
- 8 <https://www.zenarate.com/>
- 9 <https://transfrinc.com/>
- 10 <https://textio.com/>
- 11 <https://www.aid.technology/>
- 12 E.g., see World Economic Forum's White Paper on DAOs for Impact: [https://www3.weforum.org/docs/WEF\\_DAOs\\_for\\_Impact\\_2023.pdf](https://www3.weforum.org/docs/WEF_DAOs_for_Impact_2023.pdf)
- 13 <https://globalgoals.withgoogle.com/globalgoals/>
- 14 <https://www.undp.org/policy-centre/istanbul/sustainable-development-goals-philanthropy-platform-sdgp#:~:text=A%20global%20and%20national%20facilitator,supported%20by%20the%20Conrad%20N>

- 15 <https://www.undp.org/policy-centre/istanbul/sustainable-development-goals-philanthropy-platform-sdgp-0>
- 16 <https://ai4good.org/blog/climate-trendscanner-blog-post/>
- 17 <https://ai4good.org/what-we-do/sdg-data-catalog/>
- 18 <https://www.bwa.design/our-work/trend-scanner>
- 19 <https://www.technovation.org/taifa/>

## References

- Acumen Fund (2007). *The best available charitable option*. <https://apsocialfinance.wordpress.com/wp-content/uploads/2013/01/2007-baco-concept-paper.pdf>
- Ahmed, I., Fumimoto, K., Nakano, T., & Tran, T. H. (2024). Blockchain-empowered decentralized philanthropic charity for social good. *Sustainability*, 16(1), 210. <https://doi.org/10.3390/su16010210>
- Alabdulmohsin, I., & Lucic, M. (2022). *A near-optimal algorithm for debiasing trained machine learning models* (arXiv: 2106.12887). arXiv. <https://doi.org/10.48550/arXiv.2106.12887>
- Alkamoua, Z. (2023). *The impact of artificial intelligence on donor engagement for nonprofit organizations*. <https://doi.org/10.13140/RG.2.2.27126.78409>
- Almaghrabi, A., & Alhogail, A. (2022). Blockchain-based donations traceability framework. *Journal of King Saud University – Computer and Information Sciences*, 34(10, Part B), 9442–9454. <https://doi.org/10.1016/j.jksuci.2022.09.021>
- Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020). A smart chatbot architecture based NLP and machine learning for health care assistance. *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 1–6. <https://doi.org/10.1145/3386723.3387897>
- Chalopathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey* (arXiv: 1901.03407). arXiv. <https://doi.org/10.48550/arXiv.1901.03407>
- Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 10(1), 1–12. <https://doi.org/10.1057/s41599-023-02079-x>
- Ferris, J. (2021). *A generation of impact: The evolution of philanthropy over the past 25 years*. <https://search.issuelab.org/resource/a-generation-of-impact-the-evolution-of-philanthropy-over-the-past-25-years.html>
- Green, C. (2018). *Five ways charities are using virtual reality*. Charity Digital. <https://charitydigital.org.uk/topics/five-ways-charities-are-using-virtual-reality-4911>
- Hall, O., Dompae, F., Wahab, I., & Dzanku, F. M. (2023). A review of machine learning and satellite imagery for poverty prediction: Implications for development research and applications. *Journal of International Development*, 35(7), 1753–1768. <https://doi.org/10.1002/jid.3751>
- Hsu, C.-W., Chang, Y.-L., Chen, T.-S., Chang, T.-Y., & Lin, Y.-D. (2021). Who donates on line? Segmentation analysis and marketing strategies based on machine learning for online charitable donations in Taiwan. *IEEE Access*, 9, 52728–52740. <https://doi.org/10.1109/ACCESS.2021.3066713>
- Hu, Q. (2023). Unilever’s practice on AI-based recruitment. *Highlights in Business, Economics and Management*, 16, 256–263. <https://doi.org/10.54097/hbem.v16i.10565>
- Jha, N. (2023, December 3). *How to use ai for employee performance management system?* Avado. <https://www.avadolearning.com/blog/ai-for-performance-management-system/>
- Kandaurova, M., & Lee, S. H. (Mark). (2019). The effects of Virtual Reality (VR) on charitable giving: The role of empathy, guilt, responsibility, and social exclusion. *Journal of Business Research*, 100, 571–580. <https://doi.org/10.1016/j.jbusres.2018.10.027>
- Lee, J.-Y., & Jin, C.-H. (2019). How collective intelligence fosters incremental innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(3), 53. <https://doi.org/10.3390/joitmc5030053>
- Nahar, S. (2024). Modeling the effects of artificial intelligence (AI)-based innovation on sustainable development goals (SDGs): Applying a system dynamics perspective in a cross-country setting. *Technological Forecasting and Social Change*, 201, 123203. <https://doi.org/10.1016/j.techfore.2023.123203>
- Odeyemi, O., Mhlongo, N. Z., Nwankwo, E. E., & Soyombo, O. T. (2024). Reviewing the role of AI in fraud detection and prevention in financial services. *International Journal of Science and Research Archive*, 11(1), 2101–2110. <https://doi.org/10.30574/ijrsra.2024.11.1.0279>
- Pantelakis, A. (2023, May 30). *Can AI help beat unconscious bias in hiring? Yes, it can*. Recruiting Resources: How to Recruit and Hire Better. <https://resources.workable.com/stories-and-insights/overcome-unconscious-bias-hiring-ai>

- Philanthropy, I. U. L. F. S. of. (2022). *The giving environment: Understanding how donors make giving decisions*. <https://hdl.handle.net/1805/27562>
- Raheman, F. (2020). *Crowdfunding 4.0: A novel influence-based global fundraising platform and system* (World Intellectual Property Organization Patent WO2020141360A1). <https://patents.google.com/patent/WO2020141360A1/en?q=WO%2f2020%2f141360>
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>
- Samantaray, S. (2021). *Artificial intelligence-based method for analysing raw data* (World Intellectual Property Organization Patent WO2021215906A1). <https://patents.google.com/patent/WO2021215906A1/en?q=WO2021215906>
- Santana, C., & Albareda, L. (2022). Blockchain and the emergence of Decentralized Autonomous Organizations (DAOs): An integrative model and research agenda. *Technological Forecasting and Social Change*, *182*, 121806. <https://doi.org/10.1016/j.techfore.2022.121806>
- Sarker, S., Henningson, S., Jensen, T., & Hedman, J. (2021). The use of blockchain as a resource for combating corruption in global shipping: An interpretive case study. *Journal of Management Information Systems*, *38*(2), 338–373. <https://doi.org/10.1080/07421222.2021.1912919>
- Sooter, N. M., & Ugazio, G. (2023). Virtual reality for Philanthropy: A promising tool to innovate fundraising. *Judgment and Decision Making*, *18*, e16. <https://doi.org/10.1017/jdm.2023.15>
- Stamler, J., & Vanvalkenburgh, P. (2023). *Electronic impact platform for sustainable investment in socially responsible endeavors* (United States Patent US20230342853A1). <https://patents.google.com/patent/US20230342853A1/en?q=US2023342853>
- Stein, L., Beynon, E., & Borenstein, N. (2022). *Distributed platform for the development of attracting and scaling innovation* (United States Patent US20220391797A1). <https://patents.google.com/patent/US20220391797A1/en?q=20220391797>
- Tudor, M. C., Gomez, L., Giovampaola, C. D., Halopé, H., & Ugazio, G. (2024). Leveraging AI to map SDG coverage and uncover partnerships in Swiss philanthropy. In T. Walker, S. Wendt, S. Goubran, & T. Schwartz (Eds.), *Artificial intelligence for sustainability: Innovations in business and financial services* (pp. 175–206). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-49979-1\\_9](https://doi.org/10.1007/978-3-031-49979-1_9)
- Ukko, J., Saunila, M., Nasiri, M., Rantala, T., & Holopainen, M. (2022). Digital twins' impact on organizational control: Perspectives on formal vs social control. *Information Technology & People*, *35*(8), 253–272. <https://doi.org/10.1108/ITP-09-2020-0608>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, *11*(1), 233. <https://doi.org/10.1038/s41467-019-14108-y>



# 6

## THE USE OF AI AND TECHNOLOGY BY CIVIL SOCIETY ORGANIZATIONS AND ITS INTERNATIONAL IMPLICATIONS

*Anita Budziszewska and Oto Potluka*

### 1 Introduction

Recent technological developments (up to and including artificial intelligence) are proving ever more popular with and necessary for organizations operating in civil society (CSOs), including those of a philanthropic nature.

Recent research shows that the use of AI by business organizations increases efficiency (Nielsen, J., 2023; Noy & Zhang, 2023) and reduces the time needed to complete defined tasks (Brynjolfsson et al., 2023; Noy & Zhang, 2023). In particular, it concerns effects in terms of larger and easier flows of money, greater intensity of interaction, dependence on digital exchanges, and, ultimately, the overcoming of barriers in general that impose limits, including geographical borders (Dutta & Lanvin, 2023).

The contemporary “Technological Revolution” has penetrated all areas of political, economic, and public life (Miguélez et al., 2021; Mustard, 1997), including the systems through which work is organized. The question is whether CSOs are ready to reach out for technological solutions, even though this may seem an unavoidable necessity when new policies and strategies are being developed (not least in the “to be or not to be” circumstances of the COVID-19 pandemic). A new challenge has, thus, arisen when it comes to the digital and virtual operations of the non-governmental and civil society sectors. Charitable organizations in the UK may serve as a positive example here, in line with UK leadership in developing and using AI. The most recent *Charity Digital Skills Report* (Amar & Ramsay, 2023) clarifies that the charities operating there tend to use digital and technological tools more and more in their work. In particular, the 2023 Report notes that, out of 100 charitable organizations, 27 stated that AI was now in everyday use, with another nearly 30% to do so in the case in the near future. A year earlier, as many as 56% of organizations surveyed confirmed that digital tools were being used in their work (Amar & Ramsay, 2022). Nevertheless, most charitable organizations still lag far behind when it comes to the use of digital tools (Latorero, 2018). Access and the cost of access seem to be the biggest challenges of AI, together with the lack of equality and transparency (Global Symposium: Artificial Intelligence and Inequality, 2017).

Beyond that, Plunkett (Legraine, 2023) urges today’s charitable organizations to be much faster at adopting AI, given its potential to shape the entire philanthropic sector. In Plunkett’s view, organizations with a charitable profile have so far been too slow to embrace the first wave of Internet

and innovation – to the extent that one can already speak of missed opportunities (i.e., many potential benefits – see more in Legraine, 2023).

The status of AI as an inseparable part of civil society must be recognized, irrespective of ethical doubts that accompany the development of and are also present in public opinion (Fine & Kanter, 2020).

In our chapter, we investigate the success and failure of the digital civil society in Switzerland – as an example of an economically stable country – and the international and global implications of the (non-)use of AI, through the prism of CSOs dealing with digitalization and AI tools. Based on the past 20 years, Switzerland has one of the most developed philanthropic sectors in the world, which has made an unprecedented leap forward in society as a whole – very much in line with the key role technology has been able to play. Moreover, civil society in Switzerland is very sizable and engaged (Helmig et al., 2017, p. 131).

Following the discussion of Swiss digital civil society, we focus on the international and global implications of AI being (not)used. It is clear that access in the international dimension is far from equal and that the use of AI is often even impossible due to a major global imbalance in AI creation and restricted accessibility in digital technologies (Kowalski, 2021). The disparities may even deepen steadily as AI continues to evolve (Korinek & Stiglitz, 2021).

This may also create a broader impact linked to challenges of developing an effective international civil society. The part in question will thus signal the different challenges for developing this version of civil society and the (im)possibilities of AI being used to restore balance.

On the one hand, we observe that AI helps increase efficiency and effectiveness and that CSOs can raise their performance using technological tools. On the other hand, we observe that even well-developed CSOs like those in Switzerland, with their access to resources and capacity to use AI, still struggle to some extent with the shaping of digital policies. This may signal that if well-developed CSOs can face certain difficulties, the CSOs from the Global South are likely to be in a much worse situation, which has further global implications at different levels.

In terms of terminology, the use of the terms “Global South” and “Global North” is based here on the definition used in the Encyclopedia of the UN Sustainable Development Goals, which clarifies as follows:

The Global South – a term that refers to developing countries located mostly in the southern hemisphere, with generally low-income levels and facing different structural problems. The Global North – a term that refers to developed countries concentrated in the northern hemisphere, characterized by high levels of income, technological advancement, well-developed infrastructure, and macroeconomic and political stability.

(Kowalski, 2021)

The terms Global South and Global North are used interchangeably in the text with developing and developed countries. In terms of CSOs, we consider digital civil society CSOs to be those that meet the requirements of the classic definition of CSOs by Salamon and Anheier (1998), expanded by the fact that these CSOs deal primarily with digitalization. Thus, these organizations are (i) voluntary-based, (ii) pursue the common good, (iii) do not perform governmental tasks, (iv) are nonprofit, and (v) fulfill the non-distribution constraint (by not distributing financial surpluses to owners or managers). Moreover, (vi) they deal primarily with digitalization. We have relaxed the condition on formal structures because informal groups are also very active in digitalization (like hackers, although the common good condition limits participation to “good” hackers).

## 2 Digital technologies and Swiss civil society – an overview

Nowadays, almost all CSOs use digital technology (see, for example, the increasing use of digital technologies and social media as described by Özüpek, 2017, p. 103). It is also the case in Switzerland, the country we chose to study the role of CSOs in shaping digital policy. This country has a large number and a high level of volunteer engagement (Helmig et al., 2017, p. 131). This sizable structure of civil society in Switzerland provides an excellent opportunity to study CSOs oriented toward digital technologies. Furthermore, Switzerland ranks among the most digitally advanced countries (Chakravorti et al., 2020, p. 35; Hantrais & Lenihan, 2021), while its hub status when it comes to international CSOs and the proximity of UN organizations seated in Switzerland contributes to the global importance of the Swiss civil-society landscape.

In Switzerland, 126 organizations and groups in civil society have been identified as primarily dedicated to digitalization (Potluka et al., 2022). These organizations are either foundations that fund digital projects or nonprofits that promote the use of digital technologies, the safety of digitalization, training in digital skills, etc. Most Swiss digital civil society organizations were founded in the last 20 years. Compared to other organizations, those active in the field of digital civil society prove to be significantly younger. The average digital civil society organization was founded in 2010. The other organizations are significantly older, with an average founding year of 1997 (Potluka et al., 2022). This gives hope that they will be more dynamic in using digital technologies in their activities to increase efficiency.

When we refer to data and data analysis in this chapter, we are referring to an online survey we organized among the above-mentioned 126 digital civil society organizations in Switzerland. The survey was conducted on the Qualtrics platform from July 19 to November 17, 2021.

## 3 Efficiency of Swiss civil society through digital technologies

### 3.1 Financial capacities and efficiency in digital civil society

Efficiency, financial capacity, and volunteerism are among the most crucial aspects of civil society capacity. Moreover, resources are perceived as the main obstacle to access various technologies when comparing the Global South and North (Aly, 2022; Arun, 2019; Mannuru et al., 2023). When speaking about financial capacities, long-term growth of assets is perceived as the essential attribute of financial health within CSOs (Bowman, 2011; Fagan, 2006; Potluka et al., 2017) – something that can be seen as enabling efficient work. It provides resources for other activities in civil society, such as operational knowledge, expert knowledge, political networks, and staff recruitment (Carmin, 2010). The latter resource is a particular challenge in civil society because it concerns not only paid staff but also volunteers (Studer & von Schnurbein, 2013; von Schnurbein et al., 2022). In addition to financial and human resources, there are also assets related to networks and network governance. For a network and co-creation to be successful, trust, legitimacy, learning, power, and fairness should be present in networked governance (Larsson, 2019; Wegner & Verschoore, 2022). Civil society is regarded as a crucial stakeholder that adds to the legitimacy of the whole process and counterbalances the interests of the business and public sectors (Estermann et al., 2018; Larsson, 2019). These resources can also be increased where digital technologies and AI are deployed.

CSOs are often financed from sources other than public funding, including donations or grants. This approach fosters civil society's independence while ensuring that its interests are represented and that the absence of bureaucracy enhances the quality of activities. In addition, a further benefit of such an approach for civil society is that it is easier to attract donors and grants because the

collaboration or activity is validated by the public sector, which raises confidence in project outcomes. Thus, the use of AI has the potential to increase the efficiency and effectiveness of CSOs (Fine & Kanter, 2020), just as the use of AI has increased efficiency in companies by applying it to repetitive tasks and automated activities (Nielsen, S.B., 2023; Raftree, 2023).

Data from Swiss digital civil society shows us that the average annual budget of the organizations in our survey is between CHF 100,000 and CHF 500,000. The main funding source is secured through its own resources, closely followed by public and private donors. Less than half of the organizations surveyed complained about their financial situation (see Table 6.1). These are mainly organizations where either self-funding or funding from small donors plays a primary role. This shows the importance of combining different sources of funding.

We see positive aspects in the data, primarily in regard to the distribution of the resource portfolio for all digital civil society respondents. We identified five primary funding sources, with almost a third of the organizations naming their funds as the most important source. Overall, the fact that Swiss digital civil society is not dependent on just one or two funding sources reduces the risk of financial problems.

In terms of organizations' budgets and their ability to achieve financial sustainability, the data confirms that larger organizations, with annual budgets of over half a million Swiss francs, generally have fewer problems than very small organizations (with a yearly budget of up to CHF 10,000). From this perspective, digital civil society is in a similar position to ordinary civil society (Potluka et al., 2022). This finding confirms how larger organizations have an advantage in terms of efficiency, professionalization, and economies of scale, with similar findings from another financial research on CSOs. See, for example, the uneven distribution of EU funding to the most skilled CSOs, as detailed in Potluka et al. (2017), or the existence of some 100 professionalized bodies throughout the whole civil-society sector in the Czech Republic (Kundrata, 2007) (Table 6.1).

### **3.2 Fragmentation of the networks as a resource within digital civil society**

Human resource is another important resource in organizations. The practice of digital civil society shows many digital specialists in the sector are concentrated in hubs (see, for example, Lausanne, Zurich, Bern, Basel, and Geneva, in Switzerland, as noted in Potluka et al., 2022). These individuals are often employed by a company but are also engaged with a CSO. Our data shows

*Table 6.1* Financial management in digital civil society in Switzerland

		<i>How do you manage to come through financially?</i>				
		<i>With no problems (%)</i>	<i>We manage (%)</i>	<i>With difficulty (%)</i>	<i>With severe difficulty (%)</i>	<i>Total (%)</i>
Resource of funds	Public sector	2.9	22.9	8.6	0.0	34.3
	Small benefactors	0.0	14.3	8.6	2.9	25.7
	Foundations	0.0	17.1	17.1	0.0	34.3
	Business sector	0.0	14.3	2.9	0.0	17.1
	Own revenues	0.0	20.0	20.0	5.7	45.7
Total		2.9	54.3	37.1	5.7	100.0

*Source:* Authors' calculation based on Potluka et al. (2022), Chi-Square p-value = 0.000, *N* = 35.

that the vast majority of these people are volunteers, not employees of CSOs (Potluka et al., 2022). However, this knowledge and these skills are fragmented across multiple organizations that are aware of each other, even if their collaboration tends to be bipolar. Collaboration in several groups of a few organizations does not allow for more pronounced synergies.

Swiss digital civil society organizations are relatively well networked with each other. About nine out of ten are in regular contact with at least three other organizations (while 51.6% of respondents have more than five contacts). Nevertheless, the structure of the networks is relatively flat. While several names appeared to be partners of others, these names were rarely repeated. The most common name, “Digitale Gesellschaft,” appeared in only five responses, followed by a second organization with three responses. Other names appeared only twice or once. This highlights the predominantly informal structures present within digital civil society in Switzerland.

Two factors may explain the flat structure referred to. The first relates to the topic of digital civil society. Informal groups (e.g., hackers) can hardly be expected to formally organize themselves or unite under a formal umbrella organization. The second factor is the length of time that digital civil society organizations existed. The organizations are young. In our sample, 56.7% of respondents’ organizations and platforms were founded after 2012. Thus, the governance of co-creation processes in digitalization is dominated by the other two sectors outside civil society (Estermann et al., 2018; Wang & Ran, 2021).

Such a flat structure limits civil society’s strength in relation to other sectors. Regarding the public sector, it is possible to invite stakeholders with knowledge of a specific field. However, it proves challenging to find a representative who can engage in policy dialogue and, at the same time, represent digital civil society as a whole.

### ***3.3 The question of the use of AI to achieve increased efficiency***

Even with limited resources, using them efficiently can help all organizations achieve better results and outcomes. Civil society is oriented toward pursuing and ensuring the common good. At the same time, evaluation raises the question of whether CSOs can improve their activities. How can AI help civil society improve its performance and impact on target groups?

Evaluations use a variety of criteria, depending on the type and purpose of the evaluation. The standard set of evaluation criteria concerns relevance, coherence, efficiency, effectiveness, impact, and sustainability (OECD, 2021). AI is a tool that can help to achieve efficiency and effectiveness by assisting with routine and monotonous activities (Raftree, 2023) and data analysis. In the case of large data sets, the costs of data collection, storage, and processing can be efficiently saved (Nielsen, S.B., 2023). This is especially true for tasks such as reviewing documents for specific information and writing minutes or summaries of meetings.

On the other hand, bottlenecks and problems remain due to the use of AI, which in its current version is unreliable and still requires human supervision (Nielsen, S.B., 2023). Moreover, AI does not address the issue of ethics. Ethical rules governing the use of AI have yet to be developed (Head et al., 2023). However, this is a crucial issue in the case of CSOs, especially when working with vulnerable target groups (Reid, 2023), as stereotypes from the Internet do much to bias the results provided by AI (Head et al., 2023). The data gap is still a challenge when using AI, although the data accessibility has improved due to various data scraping techniques and digital data collection methods (Global Symposium Artificial Intelligence and Inclusion, 2017).

It emerges that great expectations are sustained regarding AI, even when people fail to understand either the possibilities of AI or its limitations. For example, Natural Language Processing is a tool that works in a certain way based on the strings it combines. It is then clear that, for AI

to deliver reliable results, there must be knowledge behind the strings. AI does its job, but it lacks deeper knowledge. If the task is implemented in a different context, then the AI will not understand the situation and will provide standardized results that do not accentuate the specifics of the situation or environment. AI does not yet offer such knowledge.

While AI will supplement technical aspects (especially those involving routine and monotonous activities), interpersonal and especially contextual responsive aspects of work are least likely to be taken over by AI (Azzam, 2023; Mason, 2023). Especially in civil society, the context of the work in question plays an important role, as the needs of target groups vary. Moreover, ethical considerations related to safety, transparency, accountability, inclusivity, and equity remain a weak point of AI and its use (Raftree, 2023), including by civil society organizations.

On the one hand, it is essential to note that CSOs offer their services as a value-added contribution to society without anticipating reciprocal benefits. From this perspective, CSOs invariably contribute positively. On the other hand, the question arises: is doing good enough? Should CSOs strive to be more efficient? The dilemma of utilizing AI for efficiency while ensuring ethical and inclusive development closely mirrors the debate on the business-like approach of CSOs (Hersberger-Langloh, 2020). It is the task of CSO leaders to respond to this dilemma.

Looking at the problems faced by Swiss civil society, even though it has resources that must be considered sufficient when set against those of CSOs operating in the Global South, questions must be asked about the level of success achieved by CSOs in that Global South – when it comes to their use of digitalization, and AI in particular.

## **4 AI, digital technologies, and international civil society – Global South versus Global North**

### ***4.1 A global perspective on the use of digital technologies***

The 2023 Network Readiness Index report seeks to assess the impact of ICTs on society and social development. It confirms how well-developed economies show solid network readiness across all dimensions. The research was based on a survey of 134 economies around the world, taking into account 58 indicators related to the four dimensions of digital readiness termed “technology, people, management and impact” (Dutta & Lanvin, 2023). Research, among other things, revealed the dominance of specific countries and economies from a global and regional perspective. In particular, it is noteworthy that, on the list of top 10 countries, the highest position in the index was achieved by the United States. Singapore and South Korea remain the only top 10 countries in the Asia-Pacific region, with the rest situated in Europe (i.e., Finland, the Netherlands, Sweden, Switzerland, Denmark, Germany, and the UK). Switzerland ranked 6th (Dutta & Lanvin, 2023).

The first aspect made clear by the survey was that the trends in the development of digital technology and network infrastructure differ significantly from one region to another, while a second aspect was that, from a global perspective, it is Europe and the European (Western) economies that dominate the world in terms of broad digital readiness (Dutta & Lanvin, 2023). This tendency is also confirmed by Moreno (2023), who, in addition, stresses that the main challenge facing the global community when it comes to digital technology and AI is access to AI tools. Such access is almost entirely in the hands of a few highly developed countries – namely the United States, China, and the UK, which control nearly half of all AI-related patents (Crawford, 2021; Moreno, 2023).

This inequality negatively influences the formation and cooperation characterizing civil society internationally and globally (Aly, 2022; Arun, 2019; Mannuru et al., 2023). The consequences of this disparity may include poor development, a weak presence, limited participation,

and opportunity to exert influence – on the part of civil organizations operating in the societies in the Global South that do not have adequate resources. In the long term, this ensures a lack of a presence of CSOs from the Global South in global negotiations to set international legal and ethical standards and limited participation in global flows and governance, including regarding AI (Chinen, 2023). Standards and norms might again favor – and reflect what is in place in – the Global North.

This result has far-reaching implications for the global effectiveness of international civil society organizations in addressing hybrid and global challenges that require joint action and solidarity with societies in all states and regions.

A report from the Brookings Institution and the Centre for European Policy Studies makes it clear that the priority role in addressing or resolving disasters and challenges of a global nature today is precisely that of international cooperation, which also extends to the use of AI (Kerry et al., 2021). The report notes that, although the standards of cooperation regarding AI as such are mainly developed by states and governments, the architecture and governance need to be open to active and real joint action with CSOs (Baldoni et al., 2020; Kerry et al., 2021).

As the case of Switzerland shows, digitalization is business-driven, while the public sector shapes the policies (Estermann et al., 2018). While civil society can be a moderator between the two, it has so far had little to say, as it has not yet achieved the status of an influential partner.

#### ***4.2 Universality, equality, and CSOs from the Global South in global fora***

Being aware of the existing challenges, the international community strives to reduce disparities in the level of use of and access to AI and new technologies for developing countries, to ensure greater universality. One example is the standards provided by the Global Partnership on Artificial Intelligence (GPAI), as a 2020 outcome of the work of the G7 member states. One of the main tasks to be pursued (as set out in these standards) involves founding international projects cooperating, among other things, with civil society, with particular attention to the interests and involvement of countries from the Global South. The GPAI thus goes some way toward implementing the principles that lie at the heart of global governance, where both countries of the Global South and representatives of civil society are offered a seat at the table (Goralski & Tan, 2020; Tallberg & Uhlin, 2011). This would seem to be a non-standard solution (Chinen, 2023).

On the other hand, irrespective of whether organizations in society ever have fundamental and far-reaching influence in international standard-setting, practice does not seem to give full effect to this idea. Organizations from civil societies in the Global South are not always assured of full participation in the shaping of international standards, and there is no equality in the representation of their CSOs or lobby groups in international organizations. CSOs from these regions are typically smaller (and have more limited resources) than their counterparts from the developed regions, and they also lack expertise in AI tools. Under these circumstances, it is impossible for their actual influence to be as significant (Tallberg & Uhlin, 2011), including the further development of the technologies and the use of AI. Latonero (2018) puts it straight: “It can be difficult for civil society organizations, especially smaller ones in the Global South, to find ways to engage with AI. Therefore, organizations in developing countries may see the AI field dominated by powerful countries.”

From another point of view, this state of affairs seems to be influenced by the very accreditation policies and mechanisms that the UN and others apply and run. This is made clear by the structuring to be noted among the CSOs that participate in UN international conferences discussing universal norms – showing the representation of CSOs limited to particular states (Chinen, 2023). For example, at one of the most important series of international meetings on digitalization and

digital transformation in the broadest sense (i.e., the UN-mandated *Internet Governance Forum*), in the period 2006–2019, among the 2,830 accredited NGOs and CSOs from as many as 155 states, 1,113 came from only six states: the United States, Brazil, Germany, the UK, India, and France (Chinen, 2023).

### **4.3 AI's cultural dimension**

A further issue and challenge regarding the universalized application of new technologies and AI relates to matters of cultural and axiological relativism, including the use of AI and its ethical assessment. In other words, even if international efforts ensure that the civil societies from countries in the Global South have greater access to AI tools, relevant training resources, and broader participation in international social movements generally, the question would still remain as to whether these organizations would be able to make full use of all this – and whether they would even want to do so – to the extent that can be seen among charitable organizations in developed countries. There is a risk that the tools in question – and indeed all that AI offers – will not have the potential impact or efficacy in line with the cultural and axiological context that can be observed for them in the West.

This fact has been pointed out, among others, by Ravit Dotan (2023), who underlines that the whole AI system is founded upon and constructed within the cultural and organizational context of the West. She noted: “AI systems are designed for Western contexts, and AI systems are trained on Western data” (Dotan, 2023). Furthermore, even the principles underpinning the application and use are grounded in Western values and the axiological system of the West. As Dotan further stresses: “AI Ethics Principles May Reflect Western Values” (Arun, 2019; Dotan, 2023; Kumar et al., 2021).

About the activities of CSOs and the use of AI, one example is the different accounts that governments take of organizations in society when it comes to the process by which state policy is shaped and created, including in terms of support for the digitalization of civil society *per se*. There are differences in the cultural appreciation of what is ethical and what is not, as well as in the understanding of social order, including the so-called individualism-collectivism dichotomy. These differences can be seen very clearly when comparing the culture of giving and the culture of a society in general between Western and African cultures (Carman & Rosman, 2021; Dotan, 2023; Floridi et al., 2018). Thus, cultural values influence how people help organize their societies, become socially active, and – ultimately – use (or effectively use) AI tools (Arun, 2019; Davies, 2018).

Additionally, in African countries, for example, an ethical dimension may also be at play, as well as a general lack of goodwill toward AI, given the ongoing exploitation of the continent by companies from the Global North, including many that experiment with AI right there in Africa itself (Dotan, 2023; Gestoso, 2022). In other words, the ethical dimension may play a specific role in the context of North-South relations, leading to a lack of trust in these tools – and thus their limited use, as a morally inspired opposition to it (on the ethical aspects concerning CSO reports see Schiff et al., 2021).

In a broader sense, it also seems that CSOs – particularly those from developing countries – may face a dilemma between the intense pressure to constantly increase the efficiency of their activities, including through the use of modern technologies, and their social role and responsibility to look after ethical dimensions – including in the field of human rights (Davies, 2018). At some point, this may lead to an impasse in the shaping of appropriate policies regarding activities due to the need to meet society's expectations when it comes to taking care of the moral aspects of AI.



Such considerations may especially affect smaller and less developed CSOs, whose abandonment of the widespread use of AI tools must inevitably denote a gradual slowing down in their activities and a falling behind if the decision is made to prioritize the ethical dimension (as in the case of CSOs in countries of the Global South, for whom the ethical dimension of AI and trust in its tools may appear particularly important).

Thus, the attempt to universalize and unify the use of AI tools by CSOs worldwide may not be an optimal solution due to the growing lack of trust in AI tools in general, as well as the cultural, axiological, and technological differences that exist in different regions of the world, and thus the risk of incomplete use of AI's full potential.

Instead, the solution could be the steady transfer of know-how while keeping in mind regional specificities and a constant effort to maintain the permanent change in the logic of global governance (toward a more inclusive international civil society). In the long term, this would mean the adaptation of AI tools to the requirements and cultural conditions of a particular region or state. Such a more individualized approach would better address the needs and unleash the full potential of a given country and region.

As the example of Swiss CSOs has shown, CSOs from the Global North countries may face different challenges and, therefore, need a different approach than those from the Global South.

Despite the international efforts being made, the idea of unifying access to and global use of AI tools by CSOs worldwide seems rather difficult. At the same time, the constant evolution of AI and the global technological shift cannot be avoided in any region, so a total resignation from using AI in the long term may also be impossible.

When it comes to operational policies of CSOs from the Global South, certain challenges might also be related to the limited ability to implement AI principles, tools, and regulations at the political, social, and state levels, due to different priorities and urgent needs to solve problems of a different nature to that of transferring the latest technology.

## **5 Conclusion**

The era of digitalization and the new technologies often used by charitable organizations ushered in a change in how they reshape civil society. At the national level (as the Swiss case clearly shows), digitalization and new technologies are becoming an inseparable part of the innovative civil society. However, at the global level, the differences in digitalization and the introduction of technology in general and AI in particular, as used or not used by organizations in society, are such as to ensure an uneven (unequal) dynamic to the development of the social and charitable sector around the world.

At the most general level, the differences and disparities between civil societies and their organizations are between those of the countries of the Global South and those of the Global North – and an obvious reason for this lies in the more limited possibilities of developing countries, as well as the smaller financial outlays supplied in support of their charitable and societal sectors.

From a global perspective, this means that organizations from the Global South are more limited in their participation in shaping international civil society and global aid initiatives.

Thus, the lack of equal access to and efficient use of AI, and hence the absence of equal use, has at least several global implications. First, the progress of civil society dynamics in developing (as opposed to developed) countries is more limited, given that the technological tools up to and including AI are factors that support this development, and improved efficiency in general. Second, and as a side-effect of the first, there is the matter of the potential slow but steady marginalization of institutions from societies in the Global South when it comes to the global agenda.

As mentioned previously, even the idea of universalizing access to AI tools may not solve the problem since the civil society sector is also seen as revolving around cultural and axiological dimensions. We can observe a tendency toward regionalization rather than the universalization of the use of AI tools in the civil society sector. Such a recommendation can be found, among other things, in the NRI report (Dutta & Lanvin, 2023), which reads that “the relative standings of individual counties in their respective regional rankings reflect the importance of creating and pursuing tailored strategies and policies to address the specific and unique digital needs and challenges faced by each region.”

At the same time, international efforts should focus on the disparities mentioned above, considering not only economic aspects but also other non-material differences. The aim is to better implement and use AI’s potential in shaping the dynamics of international civil society.

As the Swiss example shows, the economic aspect of a country cannot be the only one considered when talking about the use of digitalization, as the use of AI and the development of civil society may still face struggles. Our research shows that civil societies in countries that have a strong base of CSOs and are also technologically advanced face challenges. These lie primarily in the fact that while a substantial proportion of the population has access to AI, they have yet to learn how to use it and how to use it effectively. Paradoxically, the result is similar to that of nonprofits in the Global South – untapped AI potential in the civil society sector – though for different reasons.

Although the highly developed countries with a high level of technological development have certain areas that could be improved, they still have a better chance of filling these gaps than CSOs from the Global South. The international community will have to take into account the growing global doubts about the challenges of using AI and digital technologies, as mistrust seems to be more visible. At the same time, there is an issue of human rights in the new digital era, as well as an urgent need to understand the link between trust and digital inclusion (Dutta & Lanvin, 2023). Nevertheless, the dedicated solutions should be based on a voluntary basis, as there is no global authority that can impose any regulation on the accessibility and use of AI.

## References

- Aly, H. (2022). Digital transformation, development and productivity in developing countries: Is artificial intelligence a curse or a blessing? *Review of Economics and Political Science*, 7(4), 238–256, available at: <https://doi-org.libproxy.library.unt.edu/10.1108/REPS-11-2019-0145>
- Amar, Z., & Ramsay, N. (2022). *The Charity Digital Skills Report*, available at: <https://charitydigitalskills.co.uk/wp-content/uploads/2022/07/Charity-Digital-Skills-Report-2022.pdf>.
- Amar, Z., & Ramsay, N. (2023). *The Charity Digital Skills Report*, available at: <https://charitydigitalskills.co.uk/the-charity-digital-skills-report-introduction>
- Arun, C. (2019). AI and the Global South: Designing for other worlds. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 588–606). New York: Oxford University Press.
- Azzam, T. (2023). Artificial intelligence and validity. *New Directions for Evaluation* (178–179), 85–95. <https://doi.org/10.1002/ev.20565>
- Baldoni, J., Begoli, E., Kusnezov, D., & MacWilliams, J. (2020). Solving hard problems with AI: Dramatically accelerating drug discovery through a unique public-private partnership. *Journal of Commercial Biotechnology*, 25(4), 42–48, available at: <https://www.osti.gov/biblio/1765477>
- Bowman, W. (2011). Financial capacity and sustainability of ordinary nonprofits. *Nonprofit Management and Leadership*, 22(1), 37–51. <https://doi.org/10.1002/nml.20039>
- Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at Work*. National Bureau of Economic [Research Working Paper 31161], available at: <https://www.nber.org/papers/w31161>
- Carman, M., & Rosman, B. (2021). Applying a principle of explicability to AI research in Africa: Should we do it? *Ethics and Information Technology*, 23(2), 107–117. <https://doi.org/10.1007/s10676-020-09534-2>
- Carmin, J. (2010). NGO capacity and environmental governance in Central and Eastern Europe. *Acta Politica*, 45(1–2), 183–202. <https://doi.org/10.1057/ap.2009.21>

- Chakravorti, B., Chaturvedi, R. S., Filipovic, C. & Brewer, G. (2020). *Digital in the Time of COVID: Trust in the Digital Economy and Its Evolution Across 90 Economies as the Planet Paused for a Pandemic*. The Fletcher School at Tufts University.
- Chinen, M. (2023). *The International Governance of Artificial Intelligence*. Northampton, MA: Edward Elgar Publishing.
- Crawford, K. (2021). *The Atlas of AI*. New Haven, CT: Yale University Press.
- Davies, R. (2018, September 17). *Sorting Algorithms: The Role of Civil Society in Ensuring AI Is Fair, Accountable and Transparent*. Charity Aid Foundation, available at: <https://www.cafonline.org/about-us/blog-home/giving-thought/the-future-of-doing-good/sorting-algorithms-the-role-of-civil-society-in-ensuring-ai-is-fair-accountable-and-transparent>
- Dotan, R. (2023, March). *The Impact of AI on Developing Countries. Why AI May Impact Developing Countries More Negatively & What They Can Do about It* [Paper Presentation]. 26th CSTD side event UNCTAD, Geneva, available at: [https://unctad.org/system/files/information-document/ppt\\_ai26cstd\\_Dotan\\_en.pdf/](https://unctad.org/system/files/information-document/ppt_ai26cstd_Dotan_en.pdf/)
- Dutta, S. & Lanvin, B. (2023). *Network Readiness Index. Trust in a Network Society: A Crisis of the Digital Age?*, available at: [https://download.networkreadinessindex.org/reports/nri\\_2023.pdf](https://download.networkreadinessindex.org/reports/nri_2023.pdf)
- Estermann, B., Fraefel, M., Neuron, A. C., & Vogel, J. (2018). Conceptualizing a national data infrastructure for Switzerland. *Information Polity*, 23(1), 43–65. <https://doi.org/10.3233/ip-170033>
- Fagan, A. (2006). Transnational aid for civil society development in post-socialist Europe: Democratic consolidation or a new imperialism? *Journal of Communist Studies and Transition Politics*, 22(1), 115–134. <https://doi.org/10.1080/13523270500508437>
- Fine, A., & Kanter, B. (2020). *AI4Giving – Unlocking Generosity with Artificial Intelligence: The Future of Giving Report*, available at: <https://bethkanter.org/ai4giving/>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C. Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Gestoso, P. (2022, September 24). How artificial intelligence is recolonising the Global South. *The Mint Magazine*. <https://www.themintmagazine.com/how-artificial-intelligence-is-recolonising-the-global-south/>.
- Global Symposium: *Artificial Intelligence and Inclusion (2017, November 8–10), Rio de Janeiro*, available at: <https://dotplot.berkmancenter.org/projects/ai>
- Global Symposium: Artificial Intelligence and Inequality. (2017). AI and Inclusion: Global Symposium, Pre-Event Survey Responses, available at: <https://drive.google.com/file/d/1xXynk73DPxlcw7iD5f2ZsuNH2I1izRfe/view>
- Goralski, M. A., & Tan, T. K. (2020). Artificial intelligence and sustainable development. *The International Journal of Management Education*, 18(1), 1–9. <https://doi.org/10.1016/j.ijme.2019.100330>
- Hantrais, L., & Lenihan, A. T. (2021). Social dimensions of evidence-based policy in a digital society. *Contemporary Social Science*, 16(2), 141–155. <https://doi.org/10.1080/21582041.2021.1887508>
- Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdson, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, 2023(178–179), 33–46. <https://doi.org/10.1002/ev.20556>
- Helmig, B., Gmür, M., Bärlocher, C., von Schnurbein, G., Degen, B., Nollert, M., Sokolowski, S. W. & Salamon M. L. (2017). Switzerland: A liberal outlier for Europe. In L. M. Salamon, S. W. Sokolowski, & M. A. Haddock (Eds.), *Explaining Civil Society Development* (pp. 131–142). Baltimore, MD: Johns Hopkins University Press.
- Hersberger-Langloh, S. (2020). The marketization of nonprofits: Four essays on stakeholder management and market orientation in nonprofit organizations. *CEPS PhD Series*, N. 2.
- Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A. C., & Fanni, S. (2021, October 25). *Strengthening International Cooperation on AI, Progress Reports*. Brookings, available at: <https://www.brookings.edu/articles/strengthening-international-cooperation-on-ai/>.
- Korinek, A., & Stiglitz, J. E. (2021, February). Artificial intelligence, globalization and strategies for economic development [Working Paper 28453]. *National Bureau of Economic Research*. Cambridge, available at: <https://www.nber.org/papers/w28453>
- Kowalski, A. M. (2021). Global South-Global North differences. In W. Leal Filho, A. M. Azul, L. Brandli, A. Lange Salvia, P. G. Özyay, & T. Wall (Eds.), *No Poverty. Encyclopedia of the UN Sustainable Development Goals*. Cham: Springer. [https://doi.org/10.1007/978-3-319-95714-2\\_68](https://doi.org/10.1007/978-3-319-95714-2_68)

- Kumar, S., Raut, R. D., Queiroz, M. M., & Narkhede, B. E. (2021). Mapping the barriers of AI implementations in the public discourse system: The Indian experience. *Technology in Society*, 67(C), 1–9. <https://doi.org/10.1016/j.techsoc.2021.101737>
- Kundrata, M. (Ed.) (2007). *Dopady členství ČR v EU na NNO v programovacím období 2004–2006*, Prague: The Government Council for Non-Governmental Non-Profit Organisations.
- Larsson, O. (2019). A theoretical framework for analyzing institutionalized domination in network governance arrangements. *Critical Policy Studies*, 13(1), 81–100. <https://doi.org/10.1080/19460171.2017.1393440>
- Latonero, M. (2018). Governing artificial intelligence: Upholding human rights and dignity. *Data and Society*, 20, 1–37.
- Legraine, L. (2023, October 6). *Charities urged to adopt AI faster after 'very slow' uptake of 'first wave of internet'*. Civil Society, available at: <https://www.civilsociety.co.uk/news/charities-urged-to-adopt-ai-faster-after-very-slow-uptake-of-first-wave-of-internet.html#sthash.1cPww8o8.qY1VyX7a.dpuf>
- Mannuru, N. R., Shahriar, S., Teel, Z. A., Wang, T., Lund, B. D., Tijani, S., Pohboon, C. O., Agbaji, D., Alhassan, J., Galley, J., Kousari, R., Ogbadu-Oladapo, L., Saurav, S. K., Srivastava, A., Tummuru, S. P., Uppala, S., & Vaidya, P. (2023). Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. *Information Development*, 1–19. <https://doi.org/10.1177/02666669231200628>
- Mason, S. (2023). Finding a safe zone in the highlands: Exploring evaluator competencies in the world of AI. *New Directions for Evaluation*, 2023(178–179), 11–22. <https://doi.org/10.1002/ev.20561>
- Miguélez, F., Planas, J., & Benítez, P. (2021). Digital revolution and sociocultural change. In P. López-Roldán, & S. Fachelli (Eds.), *Towards a Comparative Analysis of Social Inequalities between Europe and Latin America* (pp. 141–167). Cham: Springer.
- Moreno, P. M. (2023, December 6). The opening speech – Deputy Director of UNCTAD, UNCTAD *eWeek* 2023, Geneva, High-level panel discussion on the digital economy in the age of AI: Implications for developing countries, available at: <https://unctad.org/osgstatement/unctad-eweek-high-level-panel-discussion-digital-economy-age-ai-implications>
- Mustard, F. J. (1997). The economy and social equity in a period of major technoeconomic change. *Scandinavian Journal of Work, Environment & Health*, 23, 10–15.
- Nielsen, J. (2023, July 16). *AI Improves Employee Productivity by 66%*. Nielsen Norman Group, available at: <https://www.nngroup.com/articles/ai-tools-productivity-gains/>
- Nielsen, S. B. (2023). Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. *New Directions for Evaluation*, 2023(178–179), 47–57. <https://doi.org/10.1002/ev.20558>
- Noy, S., & Zhang, W. (2023). *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, available at SSRN: <https://ssrn.com/abstract=4375283>
- OECD (2021). *Applying Evaluation Criteria Thoughtfully*. OECD Publishing. <https://doi.org/10.1787/543e84ed-en>
- Özüpek, N. (2017). Digitalization and civil society. In Ayhan, B. (Ed.), *Digitalization and Society* (pp. 91–112). Frankfurt am Main: Peter Lang GmbH.
- Potluka, O., Meier, D., Wolf, R., Giardina, F., & Ramacci, R. (2022). Mapping Digitale Zivilgesellschaft in der Schweiz. *CEPS Forschung & Praxis, Band 29*, available at: [https://ceps.unibas.ch/fileadmin/user\\_upload/ceps/2\\_Forschung/Publicationen/CEPS\\_Forschung\\_und\\_Praxis/Forschung\\_Praxis\\_29\\_Digitale\\_Zivilgesellschaft.pdf](https://ceps.unibas.ch/fileadmin/user_upload/ceps/2_Forschung/Publicationen/CEPS_Forschung_und_Praxis/Forschung_Praxis_29_Digitale_Zivilgesellschaft.pdf)
- Potluka, O., Spacek, M., & von Schnurbein, G. (2017). Impact of the EU Structural Funds on financial capacities of non-profit organizations. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 28(5), 2200–2223. <https://doi.org/10.1007/s11266-017-9845-1>
- Raftree, L. (2023, November 16). A Just Transition: What does it mean for AI and Evaluation? *European Evaluation Society*, available at: <https://merltech.org/a-just-transition-what-does-it-mean-for-ai-and-evaluation/>
- Reid, A. M. (2023). Vision for an equitable AI world: The role of evaluation and evaluators to incite change. *New Directions for Evaluation*, 2023(178–179), 111–121. <https://doi.org/10.1002/ev.20559>
- Salamon, L. M., & Anheier, H. K. (1998). Social origins of civil society: Explaining the nonprofit sector cross-nationally. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 9(3), 213–248. <https://doi.org/10.1023/A:1022058200985>
- Schiff, D., Borenstein J., Biddle, J., & Laas, K. (2021). AI Ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1), 31–42. doi: 10.1109/TTS.2021.3052127

- Studer, S., & von Schnurbein, G. (2013). Organizational factors affecting volunteers: A literature review on volunteer coordination. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 24(2), 403–440. <https://doi.org/10.1007/s11266-012-9268-y>
- Tallberg, J., & Uhlin, A. (2011). Civil society and global democracy: An assessment. In D. Archibugi, M. Koenig-Archibugi, & R. Marchetti (Eds.), *Global Democracy: Normative and Empirical Perspectives* (pp. 210–232). Cambridge: Cambridge University Press.
- von Schnurbein, G., Hollenstein, E., Arnold, N., & Liberatore, F. (2022). Together yet apart: Remedies for tensions between volunteers and health care professionals in inter-professional collaboration. *Voluntar*, 1–13. <https://doi.org/10.1007/s11266-022-00492-5>.
- Wang, H., & Ran, B. (2021). Network governance and collaborative governance: A thematic analysis on their similarities, differences, and entanglements. *Public Management Review*, 25(6), 1187–1211. <https://doi.org/10.1080/14719037.2021.2011389>.
- Wegner, D., & Verschoore, J. (2022). Network governance in action: Functions and practices to foster collaborative environments. *Administration & Society*, 54(3), 479–499. <https://doi.org/10.1177/00953997211024580>

# THE IMPACT OF ARTIFICIAL INTELLIGENCE AND INFORMATION TECHNOLOGY ON PHILANTHROPIC ORGANIZATIONS

Case studies of non-profit and cultural associations

*Luca Barzanti, Lia Benvenuti and Enrico Gaudenzi*

## 1 Introduction

Connections between Artificial Intelligence (AI) and philanthropy are gaining increasing interest since the rigorous quantitative approach to the problems and the challenges in this field have been developed following the success gained in the profit sector. This chapter examines the impact of AI and Information Technology (IT) in the philanthropic context by illustrating two distinct applications in very different sub-domains. Case studies use a variety of AI and mathematical methods that precisely fit the considered problems and demonstrate either the flexibility of the tools and their adaptability or the great utility of AI and IT in the philanthropic field.

The first considered application context regards the Decision Support Systems (DSSs) for fundraising management. In the context of philanthropy, non-profit organizations (NPOs), whose main activity is fundraising (FR), have a considerable role to play (Rosso et al., 2004). FR strategies are crucial for achieving the mission and reaching the goal of the ongoing campaign (Sargeant, 2001). The Donor's role and efficient management are of great importance (Duncan, 1999). For this reason, both econometric and operational literature dealt with (potential) Donors' profiles that match some specific gift inclination (Duffy et al., 2007) to support the effectiveness of the process. Economists agree that information on potential Donors plays a crucial role in achieving the improvement of fundraising strategies (Melandri, 2004b; Nudd, 2003).

Quantitative studies in econometric and economic modeling have shown the main factors influencing individuals in their choice of giving. For example, Andreoni (2006) characterizes altruism's economic and social foundations as individuating factors such as the own community or the social network and the so-called "enlightened self-interest." These variables are also modeled by Smith and Chang (2002). Lee et al. (1999) argue that an individual tends to assume a role identity as Donor that depends on their network of social relationships. They identify several variables that can

impact role identity, influence individual preferences and attitudes, and affect the utility people get from their decision about how and to what extent they donate (Cappellari et al., 2011).

In general, several factors should be considered to individuate an optimal fundraising strategy: the interests of (potential) Donors, their social network and personal profile, the operational literature and rules of thumb of the experts in the field, and the information on past campaigns. Practitioners claim that the 70%–80% success rate of a fundraising campaign is determined by choosing the appropriate target (the set of Donors to whom the strategy is addressed) and only 20%–30% from motivations and creativity (Melandri, 2004a). These factors strongly influence the gift probability, affected by individual aptitudes and economic constraints (Cappellari et al., 2011): age, instruction level, place of origin, financial situation, number of children, social network, and religious involvement. Therefore, integrating all of this information to find an optimal fundraising strategy is very complex. In this context, the need for advanced DSSs, targeted for different aims and association types, is highly felt.

The second considered domain is Financial Literacy, which is related to philanthropy through the activity of the Cultural Philanthropic Organizations. The subject of financial education is broad and varied and includes very different approaches. In the international institutional sphere (Organisation for Economic Co-operation and Development – OECD, 2011), there is a need to monitor the level of financial education of a specific population. This involves statistical aspects of sampling, the design of questionnaires, and the analysis of their results. Behavioral economics has also studied the links between personal finance choices and notions of financial education. Economic psychology also influences monetary choices and can be useful in developing financial education pathways.

On a more practical level, financial education plays a role in savers' choices of pension funds. The financialization of welfare is also analyzed in Caselli and Ruocco (2018), where the intersections between economic actors and philanthropic and financial instruments are explored in an international context, particularly in the Italian scenario, with specific reference to Social Impact Investing.

Financial literacy in the adult population is also extensively analyzed in OECD (2020). A more didactic approach is developed in Houghton Budd (2016), where specific strategies for teaching financial education are proposed, examining a concrete case study. However, financial education for children has received less attention over time. Anyway, interesting tools for a qualitative approach to financial literacy have recently been developed, such as the Europoli app, developed by the EduFin Committee, with the contribution of the European Commission.

The approach proposed and developed in this study fits into the framework of tools aimed primarily at young persons. Its originality is in analyzing contexts frequently observed in personal finance choices and visualizing them in a guided and interactive computer lab, respecting the directions of research in the didactics of financial mathematics. Each model is implemented by constructing a visualization diagram that captures the financial situation and its evolution, which is a strength of this approach. The use of computer tools, even the most advanced ones, makes it possible to graphically model quite complex situations and to make effective financial choices. This technique is particularly useful for those people who don't have specific notions of financial mathematics (Barzanti & Pezzi, 2019a, 2019b). In this context, collaboration with a philanthropic cultural organization is in development, in order to make available a high-end IT solution with the construction of a web application and increase the usability of the proposed implementation.

The chapter is organized as follows: Section 2 considers the field of fundraising management and the employment of IT. In particular, Section 2.1 overviews the most advanced DSSs, which make use of AI, soft computing techniques, and advanced mathematical methods. In contrast,

Section 2.2 describes in detail one of these systems, with a rigorous approach completed with the help of diagrams and explanatory figures. Some very recent approaches, challenges, and ongoing developments are illustrated as well. Section 3 considers the context of Financial Literacy and the need for an advanced IT tool for the financial education of children. Section 3.1 examines two recent applications in Excel. Section 3.2 shows a very recent demo of an ongoing high-end IT solution with a web application developed in collaboration with a philanthropic cultural organization. Section 4 concludes the chapter.

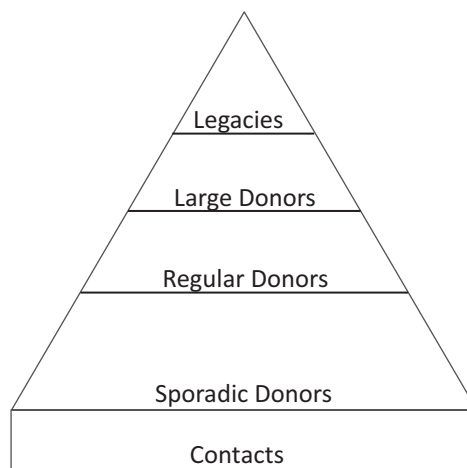
## **2 Decision support systems for fundraising management**

The strategies commonly employed use the so-called giving pyramid, an operative structure that advises the fundraisers in their macro-decisions. The giving pyramid is well studied by the operational literature (Melandri, 2017). Economic literature also refers to this concept (Lange et al., 2007). Figure 7.1 shows the Donors' segmentation determined by the pyramid (a further subdivision is possible).

The Contacts constitute the ground of the pyramid, that is, the potential Donors known by the association. One goal of a loyalty campaign is to involve new people in the mission of the association by their first donation. It is essential that a Contact becomes a Donor at the first gift request (therefore, it is significant to solicit the gift in a suitable campaign for that Contact). The effectiveness of achieving the step at the first gift request makes the new Donor more prone to become a regular Donor, establishing in this way a positive drift (the Donor becomes "hot") toward a subsequent step to the second level of the pyramid.

Quantitative methods employing database (DB) technologies have been studied and developed in the pertaining literature to make these strategies more effective (Barzanti & Pieressa, 2006; Kercheville & Kercheville, 2003). The effective use of information on Donors and Contacts (i.e., potential Donors) is crucial for optimizing the resources for the campaign by selecting the most promising Donors/Contacts for the considered context.

However, tools using a classical database approach (Flory, 2001a, 2001b) are not able to elaborate the knowledge available in the econometric and operational literature as an expert in the field



*Figure 7.1* The giving pyramid. The pyramid explains the Donors' segmentation.



does (Duffy et al., 2007) and are not able to suggest suitable fundraising strategies. The problems that these systems can solve are limited by the potential of such a technology. The support to the fundraiser is limited to giving general indications about specific claims without adequately managing all data about people, integrating qualitative information with quantitative data.

For these reasons, an innovative approach was performed in this field by Barzanti et al. (2007) who introduced the use of mathematical modeling and Decision Support System (DSS) techniques to help associations to decide the kind of campaign they have to organize, which features to implement, and which Donors of the DB list must be contacted, to maximize the expected return of the campaign, satisfying time and cost constraints.

### **2.1 Artificial intelligence and soft computing for fundraising DSSs**

The need for an advanced approach has led to comprehensive development in this field. In particular, quantitative methods have been specialized for different types of organizations. On the one hand, Barzanti et al. (2009) and Barzanti and Mastroleo (2013) (their corresponding systems are hereafter called Knowledge Fuzzy Mathematical - KFM by Barzanti et al. (2009) and Advanced Mathematical - AM by Barzanti and Mastroleo (2013)), have dealt with large-sized associations, including international ones, owing lists of millions of contributors and a powerful organizational system that requires a very sophisticated DSS.

On the other hand, Barzanti and Giove (2012) introduced Knowledge Choquet – KC, which also considers small-sized organizations and developed a DSS based only on essential information without an organized DB. This approach has been validated in the operational world by associations that test it (as documented in Barzanti & Giove, 2012; Barzanti & Mastroleo, 2013; Barzanti et al., 2009) and in the pertaining literature (Melandri, 2017; Verhaert & Van den Poel, 2012).

Medium-sized organizations are considered in Barzanti et al. (2017), Fuzzy System – FS, and Barzanti and Giove (2018), where a DSS based on a specific mathematical model and targeted for this kind of association has been created and enhanced.

More generally, a process of evolution, strengthening, and specialization of the proposed methods and algorithm has been developed. In this context, Barzanti et al. (2020) introduced the Recommended System Fuzzy – RSF, and in Barzanti et al. (2021), it was considered that one of the goals of a loyalty campaign is involving new people in the mission of the association by their first donation. In operational language, the goal is to make some Contacts going up from the ground of the giving pyramid to the first level.

The innovative approaches to fundraising are characterized by significant use of mathematical modeling, suitably implemented according to the considered purpose or the particular focus of the process. As regards the available information, associations are classified according to the existence of a structured DB and the presence in the DB of specific qualitative information of Donors' profiles (like personal interest and relationship network), in addition to the usual information on the gifts and the typical personal profile. Usually, this classification strictly depends on the organization's size. Along with the modeling approach, the methods differ in the use of advanced mathematical and statistical techniques (probability, linear algebra, utility functions, similarity measures, Choquet integral, nonparametric estimation) or soft computing and artificial intelligence (AI) techniques (fuzzy logic, knowledge-based approach). Table 7.1 summarizes the classification of the DSSs by association size and quantitative techniques employed.

KFM method is based on a knowledge approach and extensively uses Fuzzy techniques<sup>1</sup>; it is targeted at large-sized associations with a well-structured complete DB. AM method is an improvement of KFM, which uses Advanced Mathematical techniques instead of soft computing.

Table 7.1 Classification of the DSSs by association’s size and quantitative techniques employed

	<i>Advanced mathematical techniques</i>	<i>Soft computing and AI techniques</i>
Large-sized associations	AM, RSF	KFM
Medium-sized associations		FS
Small-sized associations	KC	

In particular, the fuzzy evaluation of the gift probability is replaced by an entirely mathematical estimation based on the Choquet integral with rigorous upper and lower bounds. Moreover, the utility function approach allows to contemplate not only the immediate maximum expected return objective but also a long-term perspective considering the evolution of the Donor lifetime as well.

Medium-sized organizations are considered in FS. FS only uses the quantitative information of the DB to rank and select the most promising Donors for a specified campaign by a Fuzzy technique. RSF is focused on the strategic goal of involving new Donors in the association with their first gift. The Recommender system is based on a Fuzzy approach and uses a specific Similarity measure developed for this aim, starting from a general similarity formula. It is targeted at medium-sized organizations and uses quantitative and qualitative information from a medium-structured DB.

KC uses the features of the organization profile to select the best fundraising strategy. This approach is necessary when structured DB information is unavailable, and it is therefore targeted at small associations. The knowledge of one or more experts is extracted by hierarchically organizing the organization’s domain and identifying a non-additive measure coupled with the Choquet integral. DSSs have been variously validated, tested, cited, and referred. In addition to the remarks above, see Moro et al. (2018) and Ruixia et al. (2010).

In general, the advantage of the DSSs targeted for large-sized organizations is the completeness of the considered process, obtained by managing both Donors and Contacts, by considering the total expected return of the campaign and by allocating an eventual budget saving. The disadvantages concern the necessity of frequent maintenance of DB and the non-immediate interaction with the system by the fundraisers, who are required to be trained on quantitative notions, particularly for AM.

Concerning medium-sized associations, FS is relatively easy to use and considers both the Donors’ ranking and the total expected return, with simple management of the budget saving. Contacts are not considered, and consequently, no long-term strategy is examined. On the other hand, RSF considers the Contacts, analyzing the best method to acquire Donors from Contacts. Although the mathematics is nontrivial, the DSS use is quite simple, and the results are specific and effective.

Finally, KC is easy to use and needs no DB (and consequently no information maintenance), but the process of identification of the measure is elaborate, with the fulfillment of an articulate questionnaire by the association fundraisers. For a wide review, see Barzanti (2021).

An idea of articulating such systems is shown in Figure 7.2, where the architecture of KFM is displayed. In the first part, the feasible strategies with their Donors’ targets are selected by an expected gift model and fuzzy rules in compliance with the marginal gain condition. Then, the possible Contacts selection based on fuzzy rules is performed if a budget saving occurs. Then, an articulated strategies evaluation is completed, with a dynamic estimation based on a mathematical

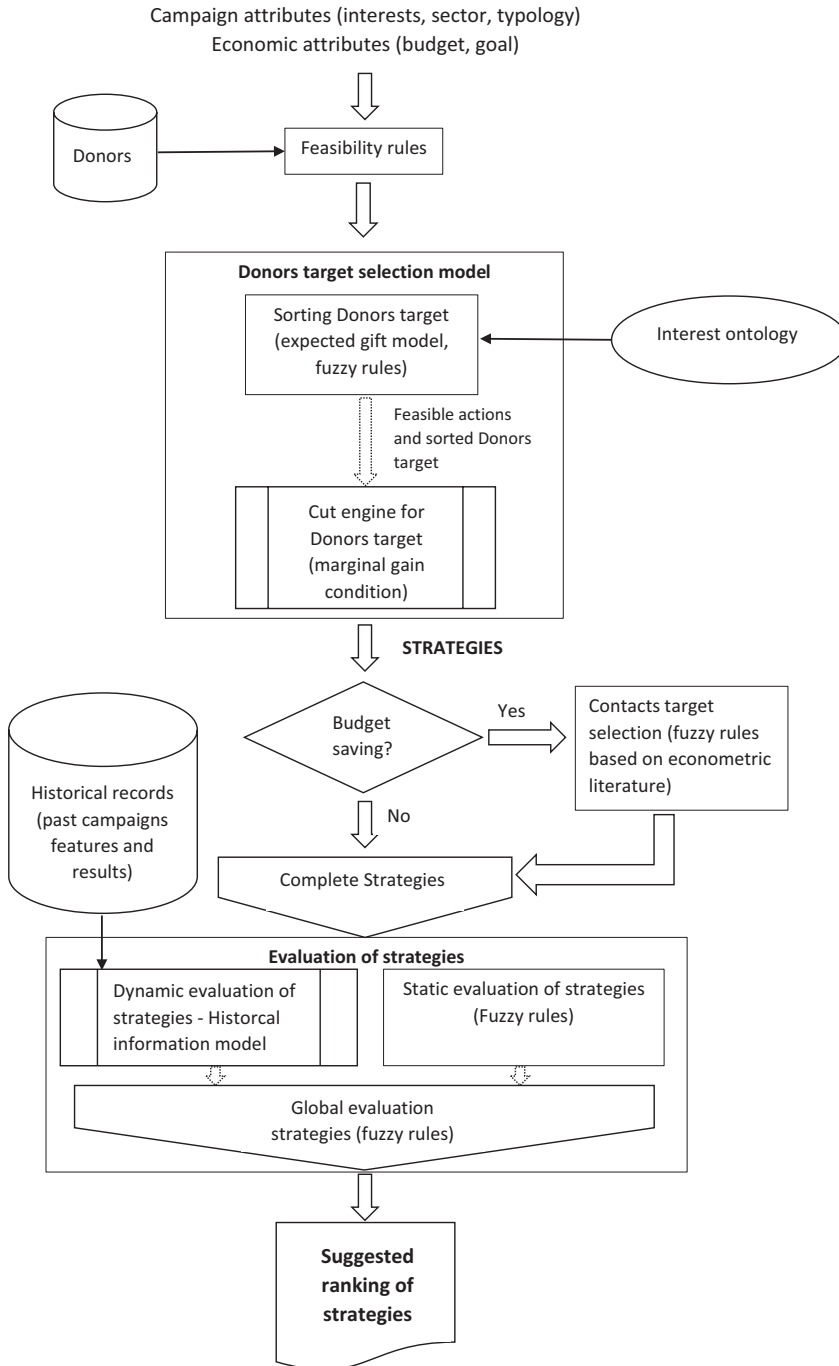


Figure 7.2 The architecture of KFM. Each part of the system is highlighted, with the explication of all the interactions.

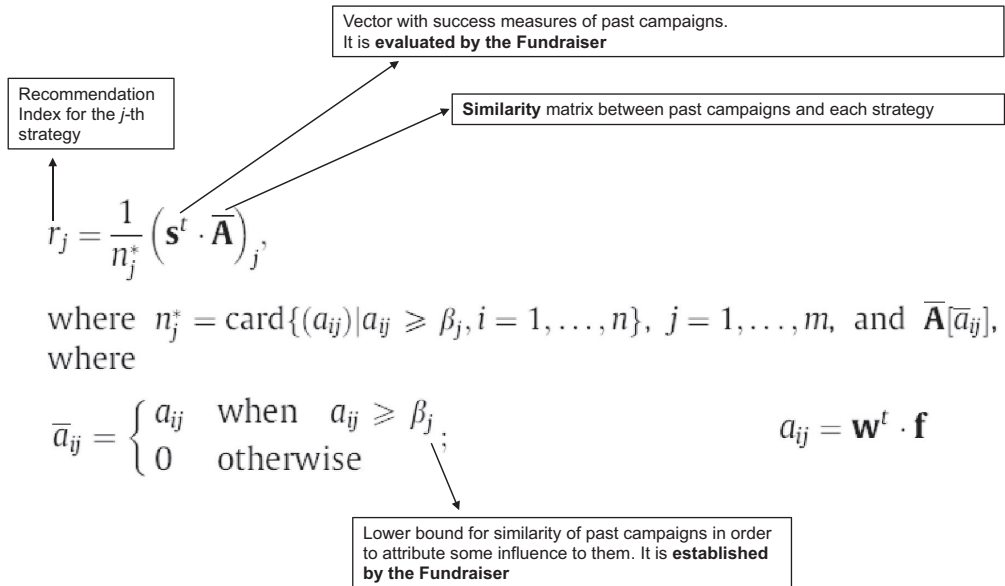


Figure 7.3 The mathematical model of dynamic evaluation of strategies in KFM. In this linear algebra model, “card” denotes the cardinality (i.e., the number of elements) of the considered set.

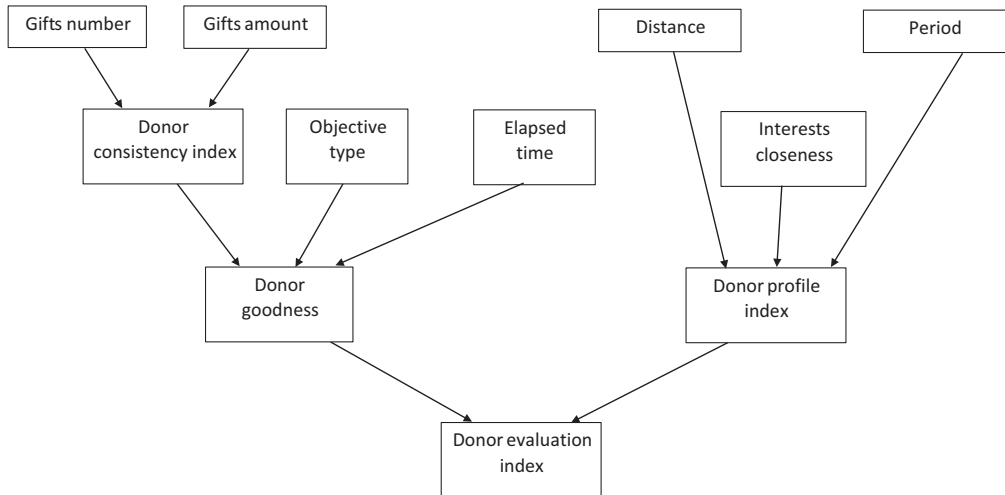


Figure 7.4 Fuzzy evaluation of the gift probability in KFM. The fuzzy structure of the variables is shown.

model (Figure 7.3) and a static fuzzy model. The global evaluation of the strategies produces their ranking. Figure 7.4 shows the details of the fuzzy evaluation of the gift probability.

## 2.2 Case study: the system FS

FS is developed for medium-sized associations and uses fuzzy techniques to maximize the global return of a campaign. As for the data, it uses the classical quantitative information of Donors (i.e.,

gift amount, recency, and frequency). Data is organized in a simulated DB constructed from experts' knowledge based on a realistic composition of a set of Donors.<sup>2</sup> The Donors' segmentation was determined following the giving pyramid. A set of 30,000 Donors is considered, and about 400,000 gift requests are collected. In the set, about 6% are Large Donors, 19% are Regular, and are further subdivided into "stable" (SRD) and "dynamic" (DRD). The remaining 75% are Sporadic Donors; about 25% made only one donation (OSD), and the rest made more than one donation (MSD). Legacies are not present in the considered sample.

Concerning the gift history, the DB includes the number of donations for each Donor; the gift amount for each donation (measured in thousands of euros); the elapsed time or recency, that is, the number of units of time since the last donation (measured in semesters); the number of past gift requests, including those in which the donation was not subsequently made. Other personal profile variables collected are risk aversion, measured as numbers of insurance policies signed by the Donor; education, subdivided into four categories: Master and Ph.D., Bachelor, High School, other/lower school level; age and number of children; wealth, measured in thousands of euros.

Tables 7.2 and 7.3 report a synthesis of the data collected in the DB. Table 7.2 illustrates the segmentation of the Donors population in the Giving Pyramid related to some characteristics: the minimum and maximum Donation amounts are shown. Large Donors have high gift amounts, Regular Donors from low to medium gift amounts and from medium to high frequency, whereas the Sporadic ones are characterized by low amount and frequency.

Only 10% of the Large Donors have "low wealth," and the percentage increases respectively to about 40% for Regular and 70% to Sporadic Donors. In the last column, the percentage of Donors who subscribed to at least one insurance contract is reported; this is a commonly used measure of the risk aversion of a person.

In addition, Table 7.3 shows some statistics of Donors' profile characteristics and gift history.

Table 7.2 Some Donors' characteristics with respect to their segmentation

<i>Donors</i>	<i>Min D. amount</i>	<i>Max D. amount</i>	<i>Low wealth (%)</i>	<i>Ins. policies <math>\geq 1</math> (%)</i>
Large	300	1,000	10	65
Regular (DRD)	100	500	40	65
Regular (SRD)	50	400	40	65
Sporadic (MSD)	30	100	70	35
Sporadic (OSD)	10	50	70	35

Table 7.3 Statistics regarding Donors' profiles

	<i>Min</i>	<i>Max</i>	<i>Mean</i>	<i>Std. Dev.</i>
Gifts' number	1	28	6.40	5.20
Elapsed time	0	119	94.72	25.76
Amount	20	1,000	133.65	158.20
Wealth	10	1,000	398.47	310.17
Risk aversion	0	5	1.07	1.67
Age	18	89	53.43	20.85
Childrens' number	0	3	1.50	1.12

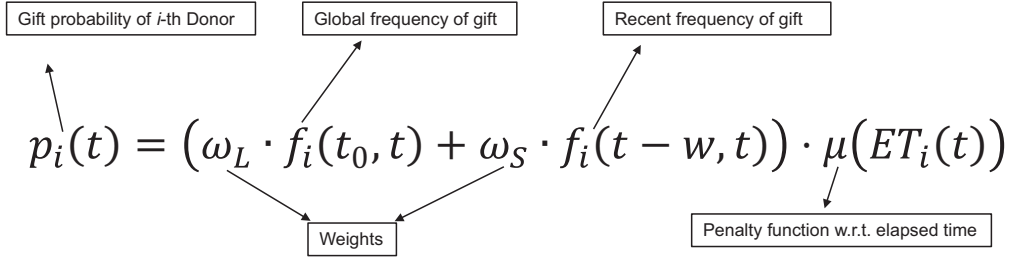


Figure 7.5 Evaluation of the gift probability in FS. The role of each mathematical object of the evaluation is explicated.

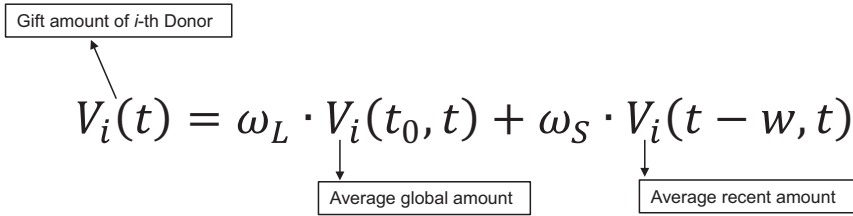


Figure 7.6 Evaluation of the amount in FS. The role of each mathematical object of the evaluation is explicated.

The system implements a model, whose key points are the particularly in-depth estimation of the giving probability (Figures 7.5 and 7.6) and the amount (Figure 7.6) expected from each Donor and the fuzzification process (Figure 7.7).

The two values  $f(p)$  and  $g(V)$  are subsequently aggregated using a suitable aggregation operator. In this case, the choice of a conservative strategy suggests the adoption of the MIN operator.

Summarizing, for each Donor in the DB, represented by the ordered couple  $(p_i(t), V_i(t))$ , a score is computed, as in (7.1):

$$Score_i = MIN\{f(p_i(t)), g(V_i(t))\} \tag{7.1}$$

The Donors are ordered using the values  $Score_i$ , and  $D = \{d_{(1)}, d_{(2)}, \dots, d_{(n)}\}$  is the ordered list of all the Donors with  $Score_i > 0$  ( $N \leq M$ ), i.e.,  $\{d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n)}\}$ .  $U_{cost}$  is the cost for a request and  $K_{MAX} = INT(Budget/U_{cost})$  is the maximum request number given the available *Budget*. The system proposes to the decision-maker (DM) the first  $K^*$  Donors in the ordered list, with  $K^* = MIN\{K_{MAX}, N\}$ . An estimation of the total gain  $TG$  is computed through the expected values of the first  $K^*$  expected gifts, see (7.2):

$$TG(K^*) = \sum_{i=1}^{K^*} p_i(t) \cdot V_i(t) \tag{7.2}$$

The System computes the total average score

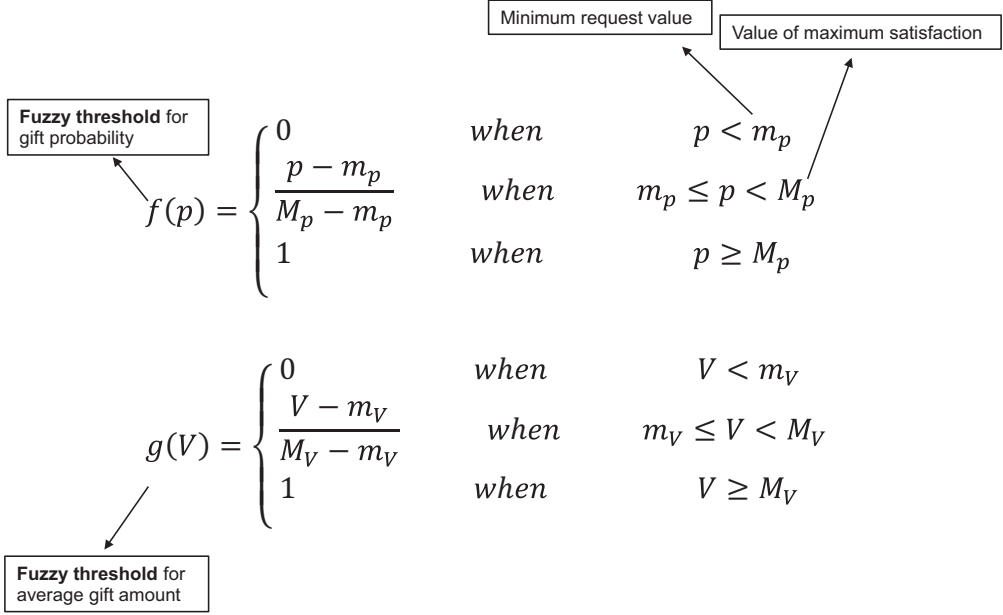


Figure 7.7 The fuzzification process in FS. The mathematical process implemented by the fuzzy system is specified.

$$TScore(K^*) = \frac{1}{K^*} \sum_{i=1}^{K^*} Score_i$$

as well.

If  $TG - Budget \geq G$  the target is reached. Anywise, if the average score is low and the expected total gain is satisfactory, the DM can decide to decrease the value of  $K^*$  and the system recomputes  $TG(K^*)$  and  $TScore(K^*)$  until a good compromise is reached.

An ongoing development in this framework regards the parametric statistical approach in the evaluation of  $p_i(t)$  and  $V_i(t)$  on the basis of Barzanti and Nardon (2022), with the aim to individuate a complete distribution estimation of these quantities and therefore a lower and an upper bound of the total return of the campaign. Another challenge can be the employment of a nonlinear fuzzification process, instead of the formulas applied in Figure 7.7, with the aim to capture pessimistic or optimistic evaluations of the Fundraiser. With respect to the gift probability, the fuzzification formula becomes like in (7.3), with  $k > 0$ , where  $0 < k < 1$  denotes optimism and  $k > 1$  pessimism.

$$f(p) = \begin{cases} 0 & \text{when } p < m_p \\ \left( \frac{p - m_p}{M_p - m_p} \right)^k & \text{when } m_p \leq p < M_p \\ 1 & \text{when } p \geq M_p \end{cases} \quad (7.3)$$

From an operational point of view, some challenges regard the interaction between Fundraiser and DSS, in terms either of the use of the software and interpretation of its results or the appropriate setting of the parameters to incorporate his/her own valuations and fine-tune the system. The key point is to make the Fundraiser confident with the software in order to use it appropriately as an instrument that can be personalized on the basis of the objectives and the professional evaluations.

As for the implementation of the system, the tool is designed in Visual Basic, using SQL language on a MS Access DB. The graphical interface allows the DM to set up the Campaign Parameters (*Target*, *Budget*, and *Unit Cost*), the Algorithm Parameters (*omega L*, *omega S*, *Time Horizon*, and *Time Window*), the Preliminary Selection Parameters (*Robustness*, *Frequency*, and *Average Past Gift*), as also the Elapsed Time Thresholds (*Min Elapsed Time*, *Max Elapsed Time*). Fuzzy aggregation parameters are also set by the indication of the bound of both thresholds (*Min Probability*, *Max Probability*, *Min Expected Value*, *Max Expected Value*). Figure 7.8 shows the whole interface.

Figure 7.8 Graphical user interface. The interaction mask between the system and the user is shown. In particular, the input parameters are classified by their meaning.



Notice that the results are widely described. First, a synthesis of the ranking by probability and expected value is presented (the Probability and Expected Value Ranking Tab). Then, the ranking obtained by the fuzzy algorithm is shown (the Fuzzy Ranking Tab), as well as the synthetic indicators for the campaign results (Campaign Results).

The Campaign Parameters are typical of a medium-sized organization; the Technical Ones are determined based on the function of the characteristics of the DB and the preliminary choices of the DM. The results of the first ranking phase (gift value and probability estimation) are presented in Figure 7.9.

In this case, after the preliminary selection, the considered Donors decreased from 30,000 to about 6,900. Consider that the DM can choose how deep to explore each single ranking. This gives a general idea of the features of data with respect to the ranking criteria. In this case, the selected step is 300 for the Probability and 600 for the Expected Value. The fuzzy aggregation implementing the algorithm (Donors' selection) is shown in Figure 7.10.

The global results of the Campaign (Expected Campaign Value), with the Total Average Score, are presented in Figure 7.11.

The target is well achieved, with a satisfactory average score. The whole budget is used as specified by the algorithm. However, suppose the DM wants to decrease the considered Donor number to slightly increase the Average Score, thus maintaining the goal of reaching the Target; in that case, there is the possibility to recompute the Campaign Results with the new Donor number

Probability and Expected Value Ranking			Fuzzy Ranking	Campaign Results		
Position	Donor Name	Probability		Position	Donor Name	Expected Value
1	D 19897	95,00%	▲	1	D 2073	121
301	D 8631	72,50%		601	D 11744	96
601	D 9469	63,33%		1201	D 21409	89
901	D 28938	55,00%		1801	D 27044	86
1201	D 15463	44,32%		2401	D 22234	82
1501	D 26357	35,00%		3001	D 24209	79
1801	D 5628	26,79%		3601	D 19807	77
2101	D 27163	17,50%		4201	D 16882	74
2401	D 23715	13,94%	▼	4801	D 26472	72

Figure 7.9 First ranking phase. The expected gift probability and amount are computed for each Donor.

Probability and Expected Value Ranking			Fuzzy Ranking	Campaign Results	
Position	Donor Name	$f(p)$	$g(V)$	<b>Score</b>	
1	D 2834	1,00	1,00	1,00	
51	D 9605	1,00	0,95	0,95	
101	D 18362	0,87	0,96	0,87	
151	D 5678	1,00	0,82	0,82	
201	D 26650	1,00	0,79	0,79	
251	D 6007	1,00	0,76	0,76	
301	D 10401	1,00	0,73	0,73	
351	D 20076	0,70	0,89	0,70	
401	D 3802	1,00	0,68	0,68	

Figure 7.10 Fuzzy aggregation and whole ranking. The score of each Donor is computed.

Probability and Expected Value Ranking	Fuzzy Ranking	Campaign Results
<b>Expected Campaign Value</b>	<input type="text" value="€ 51.068,545"/>	
<b>Budget saving</b>	<input type="text" value="0"/>	
<b>Total Average Score</b>	<input type="text" value="0,64"/>	
DM Donors Number	<input type="text" value="1000"/>	<input type="button" value="Recompute"/>

Figure 7.11 Campaign results. The expected total return of the campaign and the robustness of the result are calculated.

Probability and Expected Value Ranking	Fuzzy Ranking	Campaign Results
<b>Expected Campaign Value</b>	<input type="text" value="€ 41.366,63"/>	
<b>Budget saving</b>	<input type="text" value="€ 1.250,00"/>	
<b>Total Average Score</b>	<input type="text" value="0,71"/>	
DM Donors Number	<input type="text" value="750"/>	<input type="button" value="Recompute"/>

Figure 7.12 Campaign results with budget saving. The robustness of the result is enhanced.

through the parameter DM Num Donors and the Recompute function. In this way a budget saving is also obtained, which will be added to the Expected Campaign Value. Figure 7.12 shows the results with 750 Donors.

### 3 The use of information technology in financial literacy

The subject of financial education is broad and varied and includes very different approaches. In the international institutional sphere (OECD, 2011), the level of financial education of a specific population needs to be monitored. It involves statistical aspects of sampling, design of questionnaires, and results analysis. Financial education programs are then classified in a five-tier framework concerning needs, accountability, fine-tuning, micro impacts, and macro impacts, to overcome difficulties in their evaluation (O’Connel, 2009). Behavioral economics has also examined the links between personal finance choices and notions of financial education. Economic psychology also influences financial choices and can, therefore, be helpful in developing financial education pathways. More practically, financial education may influence savers’ behavior when choosing pension funds.

The phenomenon of the financialization of Welfare is also analyzed in Caselli and Ruocco (2018), where the intersections between economic actors and philanthropic and financial tools

are analyzed in an international context, mainly in the Italian scenario, with particular reference to Social Impact Investing. Opening new spaces for investment in social policies presents challenges in governance, responsibility, public finance, and incentives for the non-profit sector. Philanthropic associations, mainly banking foundations, support a wide range of social policy actions carried out by third-sector organizations through non-refundable grants. These third-sector organizations often partner with each other or local administrations. In this context, the phenomenon of impact finance is essential between philanthropic action and the creation and management of financial assets based on goods and/or services dedicated to fulfilling fundamental social rights.

Financial literacy in the adult population is also extensively analyzed by OECD (2020); starting from the assumption endorsed by G20 leaders, financial literacy is one of the essential ingredients for the financial empowerment of individuals and the overall stability of the financial system; a comprehensive study with descriptive statistical tools is developed in 26 countries. The study provides financial literacy information that go beyond knowledge, covering aspects of financial behavior and attitudes. In the Italian context, in Cucinelli et al. (2019), the three indicators defining the financial literacy index are taken into account and analyzed using multilevel regression statistical methods of OCSE – Financial Attitude Index (FAI), Financial Knowledge Index (FKI), and Financial Behavior Index (FBI) whose dynamics depend on the region. A more didactic approach is developed in Houghton Budd (2016), where specific strategies for teaching financial education are proposed through the analysis of real case studies.

Financial education for children and young persons has received less attention over time. In Rinaldi and Todesco (2012), research is developed concerning possible gender differences in terms of economic socialization patterns and, consequently, of financial literacy and attitude to money. The study – focused on the adolescent population of 1,635 students in Northern Italy – was carried out using statistical methods through ad hoc questionnaires. It emerged that although there is a gender difference regarding attitude to money, there is no difference compared to financial literacy. Interesting tools for a qualitative approach to financial literacy have recently been developed, such as the Europoli app, developed by the EduFin Committee, with the contribution of the European Commission.

The purpose is to educate young people about financial education through play, coming into contact with the world of finance and financial instruments. The app leads young people into a virtual world where they are invited to realize how to earn and manage money. They interact with parents, bank counselors, and teachers who show them how to manage money-related activities through games and challenges such as first savings, house purchase, child support, and retirement. Day-to-day account management, investments, financial management of contingencies, and procurement planning are some topics examined. There is also a first approach to insurance, banking, and social security instruments. A synthesis of the state of the art can be found in Lusardi (2019).

The approach proposed and developed in this study fits into the framework of tools aimed primarily at young people. However, adult users can usefully adopt it. Its originality consists in reproducing in a guided and interactive digital lab real contests frequently observed when making one's financial choices, thus respecting the guides of research in the didactics of financial mathematics (Barzanti & Pezzi, 2019a, 2019b).

The use of digital tools, even the most advanced ones, allows not only the graphic modeling of quite complex situations but also the achievement of an effective financial choice even in the absence of specific notions of financial mathematics. The interactive environment *unloads* all calculations to the Excel worksheet, bypassing all technical issues (Barzanti & Benvenuti, 2023) and

following the development of a quantitative approach aimed at financial choices in a concrete operational stage: a guided lab is implemented to solve personal finance problems and to suggest the most appropriate decisions in various contexts, depending on the financial variables involved. This ambitious goal may be reached by using digital and automatic calculation tools, which are also used to make access to the lab easier.

### 3.1 An interactive Excel financial lab with the use of graphic modeling

The following examples concretely show how the use of Excel provides solutions to complex problems in financial economics, even where specific notions of financial mathematics are not present, as demonstrated by the *Goal Seek* function used in the first example and the use of well-selected financial functions within the broader context of the second example, where a quite complex personal finance problem is solved in an evolving scenario. Both examples are focused on methodological and financial literacy aspects, while the technical approach is ensured by using Excel. In this way, students, appropriately guided in the computer lab, can understand step by step the different phases of problem-solving. The examples were proposed to students and validated in specific teaching lessons cooperating with high schools.

#### 3.1.1 Problem 1

It is a typical economic aspect example listed among financial literacy problems: The price of a good results from the balance between supply and demand. In other words, price determination relates to its supply and demand, as it follows specific market rules. In particular, the demand curve, a decreasing function, and the supply curve, an increasing function, describe the quantities demanded and offered of a good according to its price. The economic theory ensures that the balance price is where supply and demand coincide, that is why it is necessary to find the matching point of the two functions (i.e., the zero of their difference). We consider the following problem, where the demand and supply functions of a particular good (expressed in millions of units), as a function of its price, are respectively:

$$d(p) = 15e^{(-p-2)}, \quad s(p) = 6\ln(p+1)$$

We determine the break-even price (in hundreds of euros) of the good and its quantity traded on the market.

The functions considered (exponential, in this case negative, and logarithmic), typical in the representation of economic processes, show nonlinear or transcendent trends and the resulting equation can be efficiently solved using Newton's method (Barzanti & Benvenuti, 2023).

In this case, however, we proceeded using Excel (*Goal Seek* tool), which interactively implements Newton's method as mentioned above.

We must solve the following equation:

$$15e^{(-p-2)} - 6\ln(p+1) = 0$$

to find the value of the variable  $p$  at the intersection of the two curves, as shown in Figure 7.13.

We set the problem in an Excel worksheet and use the *Goal Seek* simulation tool, which can be found in the Data menu, *What if Analysis > Goal Seek* icon. We must now set up the tool: the first cell must contain the references of the target function (\$C\$5), 0 must be entered in the second

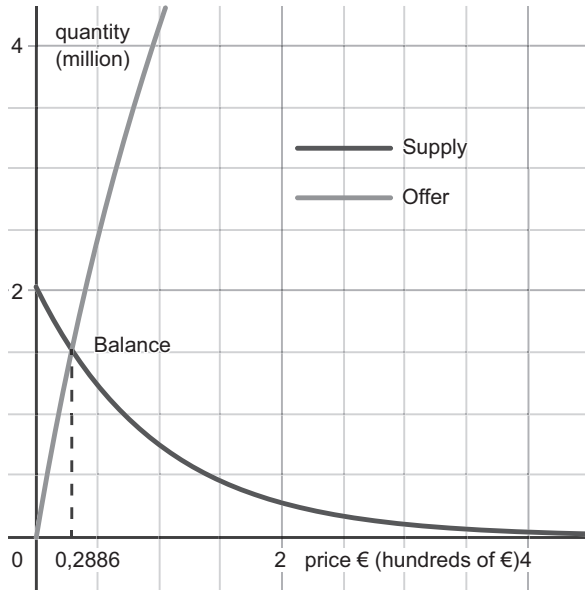


Figure 7.13 Supply and demand quantities as a function of price. The equilibrium point is shown.

cell as we want the target function to be equal to 0, and in the third cell its reference (\$C\$6) must be entered, in order to obtain the price value that will lead the supply and demand functions into balance, as shown in Figure 7.14.

Once OK is clicked, overwriting in cells C5 and C6 the results obtained from the processing, *Goal Seek* will return a balance price value (€ 28.86), corresponding to a quantity of traded goods of 1.5212 million units as shown in Figure 7.15. We can observe that in cell C5, just as in the Current Value of *Goal Seek*, an approximate result of zero is obtained due to the numerical nature of the method.

In this way, Excel can bypass the direct, specific, and analytical knowledge of Newton’s method and allows for solving a macroeconomic problem related to Financial Literacy using software that requires this numerical method. Therefore, a young student without specific quantitative knowledge of this subject can easily understand the problem and its solving strategy.

### 3.1.2 Problem 2

It offers the possibility of dealing with a personal finance problem. It is quite a complex problem evolving, and once again, it can be successfully solved using Excel without an in-depth knowledge of specific financial mathematics notions. Consider the following problem:

As rent for her garage, Samantha will receive regular advance annual installments of €2,000 each for five years. These sums will be invested in a deposit account that initially yields a yearly compound interest rate of  $i = 3.70\%$ .

- a If after two years, when the third installment is paid, the compound annual interest rate rises to 4%, what will be the amount available at the end of the fifth year?

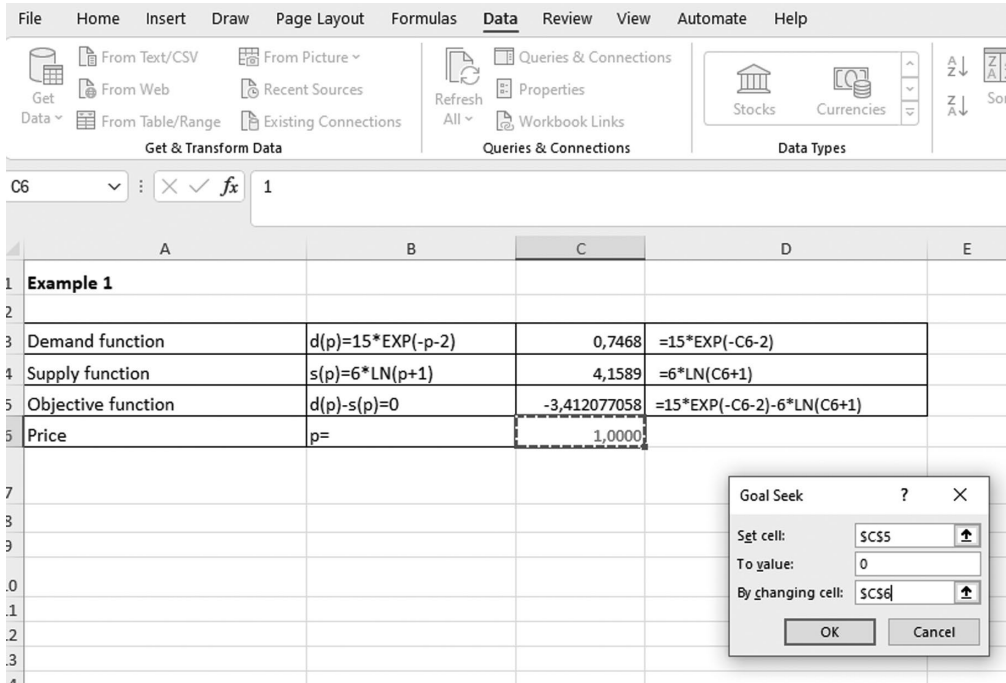


Figure 7.14 Implementation of Problem 1 in Excel with the Goal Seek tool.

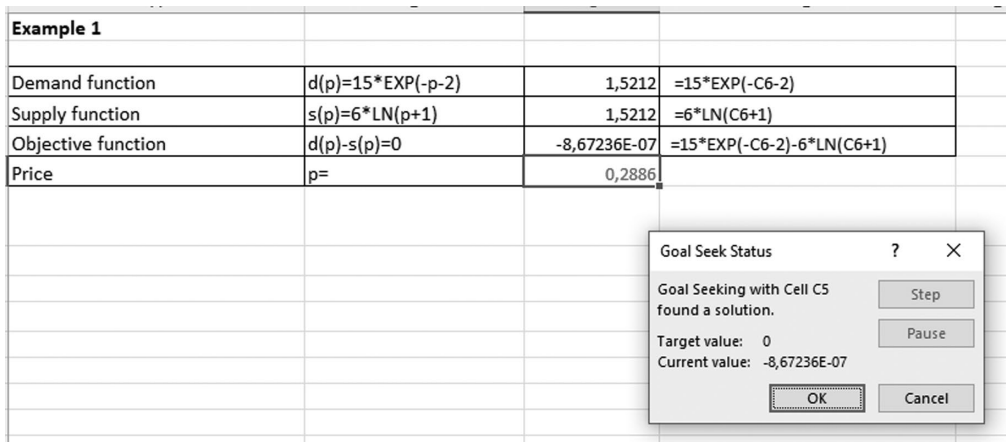


Figure 7.15 Solving Problem 1 using the Goal Seek tool.

- b) After 2.5 years, the bank proposes Samantha to disinvest up to €5,000 from her deposit account and invest it in a new product providing a higher yield (4.50%). Since Samantha decides to accept the transaction for the maximum amount allowed (€5,000), what overall amount will she get at the end of the five years, considering that the remaining sum is left in her deposit account?

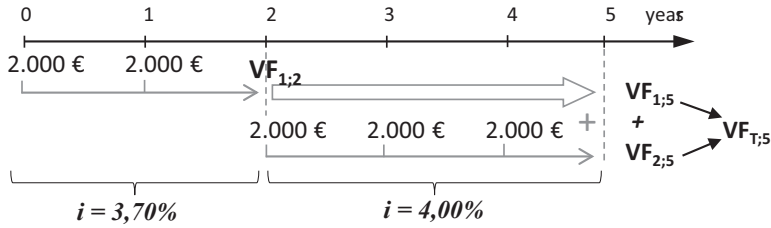


Figure 7.16 Model of Problem 2, point (a). The graphical model facilitates the resolution.

- c Calculate the global investment rate, considering both the rate change after two years and the disinvestment and further reinvestment of €5,000 after 2.5 years.
- d If the disinvestment and reinvestment of €5,000 generated trading commissions, what would be the highest amount Samantha could afford to consider the investment operations still profitable compared to her deposit account yield?

### 3.1.2 Solution

- a The problem model is described in Figure 7.16, where the first two installments at the rate  $i = 3.70\%$  and the subsequent ones at the rate  $i = 4.00\%$  are highlighted.

With the proper use of the Excel worksheet, the student can try to solve the financial problem by setting up calculations using the financial formulas (*FV* and *SUM*) available in the software, as shown in Figure 7.17.

Note that  $FV_{i;k}$  refers to the future value of the  $i$ th annuity at time  $k$ ; e.g.,  $FV_{1;2}$  represents the future value of annuity 1 at time two years. A proper use of Excel allows calculating the future value  $FV_{T;5}$  after five years (€ 11.254,19), as shown in cell B10.

- b At this point, a new investment proposal may occur. The model of the problem, which now considers the new scenario arising from the bank's proposal, is described in Figure 7.18.

Although the new model is more complex compared to (a), it may be easily managed by methodically following the graphical scheme step by step until the final solution is determined through the calculations of the intermediate future values and then of the final future value, by properly applying the yield rates for each financial operation obtained from the disaggregation of the whole problem (Figure 7.19).

The nontrivial articulation of the problem is because the new investment opportunity is proposed at a time outside the original schedule, generating on the computational side the resolution of three sub-problems (solved in cells B23, B24, and B25, respectively) and leading to the final result of  $FV_{T;5} = €11.311,71$  (cell B26), also using *ad hoc* methods, due to the specificity of the problem and the Excel functions (*FV* and *SUM*) that are correctly activated.

- c The model for calculating the overall investment rate is shown in Figure 7.20.

Note how neither the rate change at the two-year period nor the disinvestment and reinvestment of €5,000 at the 2.5-year period are shown in the model since these events do not directly affect the calculation of the overall rate (in particular, with the disinvestment and reinvestment on €5,000 calculation the entire amount available at that time is still kept invested): the only indirect influence visible in the model is the final future value, which includes the effects of both operations.

1	<b>REPLY a)</b>		
2	Installment	€ 2.000	
3	Initial compound annual rate	3,70%	
4	New annual compound rate	4,00%	
5	Annuity duration 3,7% rate (years)	2	
6	Annuity duration 4% rate (years)	3	
7	$FV_{1,2}$ = Annuity future value 1 (to 2 years)	€ 4.224,74	=FV(B3;B5;-B2;;1)
8	$FV_{1,5}$ = Annuity future value 1 (to 5 years)	€ 4.752,26	=B7*(1+B4)^B6
9	$FV_{2,5}$ = Annuity future value 2 (to 5 years)	€ 6.492,93	=FV(B4;B6;-B2;;1)
10	<b><math>FV_{T,5}</math> = Total future valute (to 5 anni)</b>	<b>€ 11.245,18</b>	=SUM(B8:B9)
11			

Figure 7.17 Solving Problem 2, point (a) using Excel. Use of Excel functions *FV* and *SUM* is highlighted.

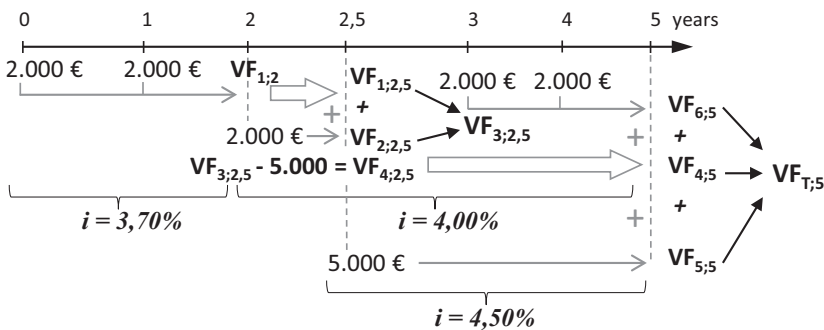


Figure 7.18 Evolution of the model in relation to a new investment opportunity. The graphical model is very effective to capture financial changes.

Figure 7.21 shows the problem setup and the Excel function *RATE* needed to determine the solution, exhibited in cell B32.

The problem resolution proposed in Figure 7.21 is only one of the hypotheses that the Excel sheet offers; the same result could be achieved using the *Goal Seek* tool, as in *Problem 1*, by constructing the Discounted Cash Flow (DCF) function and seeking the value that resets the function to zero. The math solution, obtained through Newton's method, is implemented in this case by the interactive *Goal Seek* tool, which again allows seeking the solution in a rapid way.

- d At this point, the economic operator has to decide because the new investment opportunity (which is supposed to be more profitable) has generated additional costs.

To calculate the maximum value of the expenses that Samantha could accept to decide to disinvest and reinvest €5,000, it is necessary to compute the threshold value that makes the two investment options equivalent. In other words, we need to calculate the amount of the expenses that, once subtracted from the €5,000 investment, still generate the same future value as if €5,000 were left in Samantha's deposit account, as shown in Figure 7.22 The variable  $E_{max}$  refers to the maximum expense Samantha can afford.

Figure 7.23 shows the Excel solution, where  $E_{max} = 59.59$  €.





Impact of AI and IT on philanthropic organizations

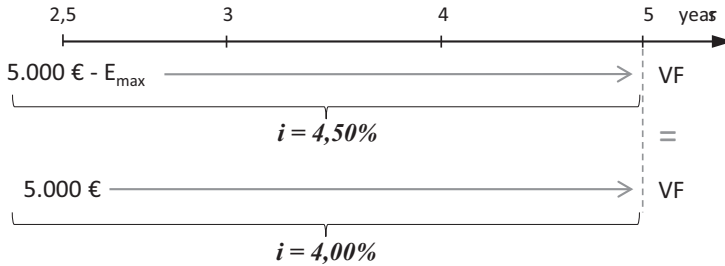


Figure 7.22 Modeling of the decision problem. The graphical model is very effective in showing the choice variable.

34	REPLY d)		
35	Annual compound rate deposit securities	4,00%	
36	New investment compound annual rate	4,50%	
37	Invested capital	€ 5.000	
38	Investment duration (years)	2,5	
39	$E_{max}$ = Maximum expense	€ 59,59	=B37*(1-((1+B35)/(1+B36)))^B38)

Figure 7.23 Solving Problem 2, point (d) using Excel.

innovative, and high-level easy-to-use solution was developed in collaboration with *Nuova Civiltà delle Macchine* (NCdM), a philanthropic association with cultural and social promotion aims.

The association aims to promote, enhance, and disseminate scientific and humanistic culture, spread awareness, and promote the development of relations between the experts in these studies, facilitating their cooperation at local, regional, national, and international levels. Its priorities are training teachers and students and ongoing training of adults. To this end, the association carries out educational research and experimentation projects with the cooperation of the education system to achieve a coherent integration with training, upskilling, and cultural promotion activities. Therefore, the successful collaboration and efforts to spread the culture of financial literacy, primarily among young students, allowed matching each other's knowledge to reach a wider audience of young people.

The proposed solution allows the user to easily learn through direct training on the job and through effective self-learning. Questions propose closed answers; in case of a wrong selection, a comment will be shown to help the user understand why and orient him to the correct answer.

Starting from the solution given by Excel in *Problem 2* (Figure 7.24), we used the XML meta-language, a marker or tag language able to create real instructions for the representation of data by a digital support, in order to edit a shared syntax readable by both man and machine. The MFCompile.exe program (created by Claudio Ricci, NCdM) generates, after inputs (Excel and XML), an HTML file executable from the Web, which is, in fact, a high-level problem-solving solution allowing a wide range of students to enter the world of Financial Literacy easily.

This executable program also solves any kind of problem by easily building the corresponding web/IT solution. When requested by a PC's browser (Chrome, Edge, Firefox, Safari), the problem page is displayed in about three seconds. In addition to the text, it contains the input elements for interacting and the control logic associated with the elements, so it can check the exact answer to a question and/or make suggestions. Input elements can be taken from the associated Excel sheet.

	A	B	C
1	<b>REPLY a)</b>		
2	Installment	€ 2.000	
3	Initial compound annual rate	3,70%	
4	New annual compound rate	4,00%	
5	Annuity duration 3,7% rate (years)	2	
6	Annuity duration 4% rate (years)	3	
7	FV <sub>1,2</sub> = Annuity future value 1 (to 2 years)	€ 4.224,74	=FV(B3;B5;-B2;;1)
8	FV <sub>1,5</sub> = Annuity future value 1 (to 5 years)	€ 4.752,26	=B7*(1+B4)^B6
9	FV <sub>2,5</sub> = Annuity future value 2 (to 5 years)	€ 6.492,93	=FV(B4;B6;-B2;;1)
10	<b>FV<sub>7,5</sub> = Total future valute (to 5 anni)</b>	<b>€ 11.245,18</b>	=SUM(B8:B9)
11			
12	<b>REPLY b)</b>		
13	Installment	€ 2.000	
14	New investment amount	€ 5.000	
15	Start annual compound rate	3,70%	
16	New annual compound rate	4,00%	
17	New investment c.annual rate	4,50%	
18	FV <sub>1,2</sub> = Annuity future value (first two installments)	€ 4.224,74	=FV(B15;2;-B13;;1)
19	FV <sub>1,2,5</sub> = Annuity future value 1 (to 2,5 years)	€ 4.308,40	=B18*(1+B16)^0,5
20	FV <sub>2,2,5</sub> = Future value 3 <sup>a</sup> installment (to 2,5 years)	€ 2.039,61	=B13*(1+B16)^0,5
21	FV <sub>3,2,5</sub> = Total future value (to 2,5 years)	€ 6.348,01	=SUM(B19:B20)
22	FV <sub>4,2,5</sub> = Residual (to 2,5 years)	€ 1.348,01	=B21-B14
23	FV <sub>4,5</sub> = Residual future value (to 5 years)	€ 1.486,88	=B22*(1+B16)^2,5
24	FV <sub>5,5</sub> = New invest. Future value (to 5 years)	€ 5.581,63	=B14*(1+B15)^2,5
25	FV <sub>6,5</sub> = Annuity future value 2 (to 5 years)	€ 4.243,20	=FV(B16;2;-B13;;1)
26	<b>FV<sub>7,5</sub> = Total future value (to 5 years)</b>	<b>€ 11.311,71</b>	=SUM(B23:B25)
27			
28	<b>REPLY c)</b>		
29	FV <sub>7,5</sub> = Total future value	€ 11.311,71	=B26
30	Installment	€ 2.000	
31	Installment number	5	
32	<b>i = Global annual compound rate</b>	<b>4,137%</b>	=RATE(B31;-B30;;B29;1)
33			
34	<b>REPLY d)</b>		
35	Annual compound rate deposit securities	4,00%	
36	New investment compound annual rate	4,50%	
37	Invested capital	€ 5.000	
38	Investment duration (years)	2,5	
39	<b>E<sub>max</sub> = Maximum expense</b>	<b>€ 59,59</b>	=B37*(1-((1+B35)/(1+B36))^B38)
40			

Figure 7.24 The overall solution of Problem 2. An overview of the whole Problem's resolution is given.

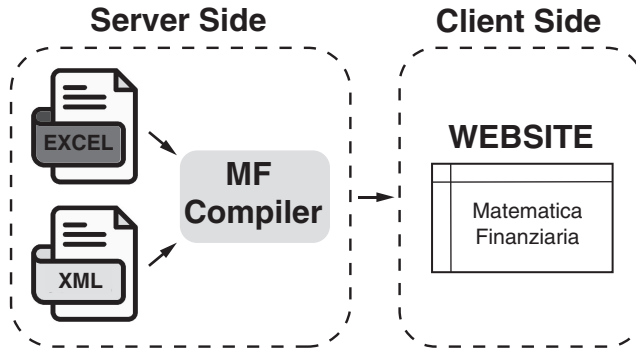


Figure 7.25 Interaction model between the involved software programs. Structure of the Demo Web Interactive Financial Lab.

Each problem, always enclosed in a single web page, is subdivided into sections containing one or more rows with text, images, buttons, or interaction objects such as a text box, radio button, checkbox, and combo-box (drop-down list).

The proposed version is a software demo developed in Italian (see Figure 7.25) and shows the interaction model between the inputs and the compilation program to generate web dialogue pages that the final user can easily operate.

Implementing a first pilot in an Italian high school made it possible to validate the proposed approach and the trial software as an effective tool to facilitate students' learning of Financial Literacy skills and knowledge.

The short-term goal is to develop a fully operative version of the software enriched with several applications offering a reasonably comprehensive overview of Financial Literacy. The next step will see the internationalization of the application and the possibility of disseminating its use to a broader audience of students.

In this way, youngsters can find answers to the most common Financial Literacy problems simply and effectively, either through traditional school learning or by using the most modern self-learning tools.

#### 4 Conclusion

The impact of Artificial Intelligence and Information Technology in Philanthropic Organizations has been investigated, with particular reference to two specific contexts, with the illustration of case studies.

The most advanced DSSs for fundraising management for non-profit organizations have been overviewed, with some mention of ongoing developments. Furthermore, a specific DSS for medium-sized organizations has been illustrated, by the rigorous mathematical modeling approach and the explanation for non-technical readers through visualization diagrams and numerical results.

Philanthropic Cultural Organizations have been considered through their activity of Financial Literacy, in particular to youngsters. A very recent Information Technology approach has been investigated, with the presentation of a visualization diagram modeling method and the development of an interactive Excel Lab. In this context, an ongoing collaboration with the organization *Nuova Civiltà delle Macchine* (NCdM) produced a high-end web application for the Excel Lab, which has been analyzed.

These applications showed either the flexibility of the tools and their adaptability or the great utility of Artificial Intelligence and Information Technology in the philanthropic field.

## Acknowledgments

We would like to thank Dr. Claudio Ricci for computational assistance and Eng. Roberto Campo-resi, President of the Philanthropic Cultural Organization NCdM, for his collaboration.

## Notes

- 1 Fuzzy Logic Systems produce acceptable but definite output in response to incomplete, ambiguous, distorted, or inaccurate (fuzzy) input. See, e.g., Czabanski et al. (2017) for further reading.
- 2 The DB has been fine-tuned in collaboration with ASSIF (the Italian fundraiser association) and Philanthropy Centro Studi, a research center of non-profit, fundraising, and social responsibility operative in the University of Bologna.

## References

- Andreoni, J. (2006). Philanthropy. In S. C. Kolm, & J. Ythier (Eds.), *Handbook the Economics of Giving, Altruism and Reciprocity* (pp. 1201–1269). Elsevier, Amsterdam.
- Barzanti, L. (2021). Decision support systems for the management of the fund raising process: A review. In F. J. Roberts (Ed.), *Decision Support Systems: Types, Advantages and Disadvantages* (pp. 79–132). Nova Science Publishers, New York.
- Barzanti, L., & Benvenuti, L. (2023). Computational mathematics at the service of problem solving: Newton's method. *Nuova Secondaria*, 7, 51–58 (in Italian).
- Barzanti, L., Dragoni, N., Degli Esposti, N., & Gaspari, M. (2007). Decision making in fund raising management: A knowledge based approach. In R. Ellis, T. Allen, & M. Petridis (Eds.), *Applications and Innovations in Intelligent Systems XV* (pp. 189–201). Springer, London.
- Barzanti, L., Gaspari, M., & Saletti, D. (2009). Modelling decision making in fund raising management by a fuzzy knowledge system. *Expert Systems with Applications*, 36, 9466–9478.
- Barzanti, L., & Giove, S. (2012). A decision support system for fund raising management based on the Choquet integral methodology. *Expert Systems*, 29(4), 359–373.
- Barzanti, L., & Giove, S. (2018). A decision support system for fund raising management in medium-sized organizations. *Mathematical Methods in Economics and Finance*, 9(10), 3–12.
- Barzanti, L., Giove, S., & Pezzi, A. (2017). A decision support system for non profit organizations. In A. Petrosino, V. Loia, & E. Pasero (Eds.), *Fuzzy Logic and Soft Computing Applications, Lecture Notes in Artificial Intelligence* (pp.270–280). Springer, Cham.
- Barzanti, L., Giove, S., & Pezzi, A. (2020). An effective fuzzy recommender system for fund-raising management. In A. Esposito, M. Faundez-Zanuy, F. C. Morabito, & E. Pasero (Eds.), *Neural Approaches to Dynamics of Signal Exchanges. Smart Innovation, Systems and Technologies 151* (pp.73–82). Springer, Singapore.
- Barzanti, L., Giove, S., & Pezzi, A. (2021). A recommender system for fund raising management. *Mathematical Methods in Economics and Finance*, 15/16(1), 1–13.
- Barzanti, L., & Mastroleo, M. (2013). An enhanced approach for developing an expert system for fund raising management. In J. M. Segura, & A. C. Reiter (Eds.), *Expert System Software: Engineering, Advantages and Applications* (pp. 131–156). Nova Science Publishers, New York.
- Barzanti, L., & Nardon, M. (2022). Estimation of the gift probability in fund raising management. In M. Corazza, C. Perna, C. Pizzi, & M. Sibillo (Eds.), *Mathematical and Statistical Methods for Actuarial Sciences and Finance, MAF 2022* (pp. 70–75). Springer, Cham.
- Barzanti, L., & Pezzi, A. (2019a). Financial modeling: Method before technique (1). *Nuova Secondaria*, 9, 72–76 (in Italian).
- Barzanti, L., & Pezzi, A. (2019b). Financial modeling: Method before technique (2). *Nuova Secondaria*, 10, 70–73 (in Italian).
- Barzanti, L., & Pieressa, L. (2006). Technological solutions for fund raising management: A comparative analysis. *CLEONP, Facoltà di Economia, University of Bologna, Sede di Forlì, Working Paper*, 30 (in Italian).
- Cappellari, L., Ghinetti, P., & Turati, G. (2011). On time and money donations. *Journal of Socio-Economics*, 40(6), 853–867.
- Caselli, D., & Ruocco, F. (2018). The financialization of welfare. *Quaderni di Sociologia*, 76, 57–80 (in Italian).

- Cucinelli, D., Trivellato, P., & Zenga, M. (2019). Financial literacy: The role of the local context. *Journal of Consumer Affairs*, 53(4), 1874–1919.
- Czabanski, R., Jezewski, M., & Leski, J. (2017). Introduction to fuzzy systems. In P. Prokopowicz, J. Czernia, D. Mikołajewski, L. Apiecionek, & D. Ślęzak (Eds.), *Theory and Applications of Ordered Fuzzy Numbers. Studies in Fuzziness and Soft Computing*, 356 (pp. 23–43). Springer, Cham.
- Duffy, J., Ochs, J., & Vesterlund, L. (2007). Giving little by little: Dynamic voluntary contribution games. *Journal of Public Economics*, 91(9), 1708–1730.
- Duncan, B. (1999). Modeling charitable contributions of time and money. *Journal of Public Economics*, 72, 213–242.
- Flory, P. (2001a). *Building a Fundraising Database Using Your PC*. DSC, London.
- Flory, P. (2001b). *Fundraising Databases*. DSC, London.
- Houghton Budd, C. (2016). In the shoes of Luca Pacioli—Double entry bookkeeping and financial literacy. In C. E. Aprea et al. (Eds.), *International Handbook of Financial Literacy* (pp. 621–637). Springer, Singapore.
- Kercheville, J., & Kercheville, J. (2003). The effective use of technology in nonprofits. In E. Tempel, (Ed.), *Hank Rosso's Achieving Excellence in Fund Raising* (pp. 366–379). John Wiley & Sons.
- Lange, A., List, J. A., & Price, M. K. (2007). A fundraising mechanism inspired by historical tontines: Theory and experimental evidence. *Journal of Public Economics*, 91(9), 1750–1782.
- Lee, L., Piliavin, J. A., & Call, V. R. (1999). Giving time, blood and money: Similarities and differences. *Social Psychological Quarterly*, 62(3), 276–290.
- Lusardi, A. (2019). Financial literacy and the need for financial education: Evidence and implications. *Swiss Journal of Economics and Statistics*, 155(1). <https://doi.org/10.1186/s41937-019-0027-5>.
- Melandri, V. (2004a). *Fundraising Course Materials*. D.U. Press, Bologna (in Italian).
- Melandri, V. (2004b). Intelligent management of fund raising in nonprofit organizations. *Terzo Settore*, 10, 43–49 (in Italian).
- Melandri, V. (2017). *Fundraising*. Civil Sector Press, Toronto.
- Moro, S., Cortez, P., & Rita, P. (2018). A divide-and conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, 35(3). <https://doi.org/10.1111/exsy.12253>.
- Nudd, S. P. (2003). Thinking strategically about information. In E. Tempel (Ed.), *Hank Rosso's Achieving Excellence in Fund Raising* (pp. 349–365). Wiley, New York.
- O'Connell, A. (2009). Evaluating the effectiveness of financial education programmes. *OECD Journal: General Papers*, 2008(3). [https://doi.org/10.1787/gen\\_papers-v2008-art17-en](https://doi.org/10.1787/gen_papers-v2008-art17-en).
- OECD (2011). *Improving Financial Education Efficiency*. OECD Publishing.
- OECD (2020). *OECD/INFE 2020 International Survey of Adult Financial Literacy*. OECD Publishing.
- Rinaldi E., & Todesco L. (2012). Financial literacy and money attitudes: Do boys and girls really differ? A study among Italian preadolescents. *Italian Journal of Sociology of Education*, 11(2), 143–165.
- Rosso, H., Tempel, R., & Melandri, V. (2004). *The Fund Raising Book*. ETAS, Bologna (in Italian).
- Ruixia, Y., Jianguo, Z., & Xiang, W. (2010). Vague set methods of multi-criteria fuzzy decision-making. 2010 Chinese Control and Decision Conference, Xuzhou, 658–661.
- Sargeant, A. (2001). Using donor lifetime value to inform fundraising strategy. *Nonprofit Management and Leadership*, 12(1), 25–38.
- Smith, W., & Chang, C. (2002). Shipping the good apples out: A note on contributions of time and money. *Economic Bulletin*, 10(1), 1–14.
- Verhaert, G. A., & Van den Poel, D. (2012). The role of seed money and threshold size in optimizing fundraising campaigns: Past behavior matters!. *Expert Systems with Applications*, 39, 13075–13084.

# 8

## ON THE IMPACT OF AI-EMPOWERED, GAMING-BASED VIRTUAL WORLDS ON PHILANTHROPY

*Marc Schipper, Manouchehr Shamsrizi and Adalbert Pakura*

### 1 Introduction

In the rapidly evolving landscape of technology and innovation, one of the most striking phenomena of the recent era has been the rise of generative AI like ChatGPT. Developed by OpenAI, this advanced generative AI has transcended beyond mere technological novelty to become a global sensation, transforming industries, education, communication, and daily life. Its ability to (seemingly) “understand,” respond, and interact in human-like manners has not only showcased the potential of generative AI but has also opened new avenues for its application across various sectors (Deng & Lin, 2023). As ChatGPT and similar technologies evolve, they underscore a significant shift toward an increasingly digital and interconnected world, where the boundaries of what is possible are being redefined almost every other month.

Parallel to the technological advancements spearheaded by AI, the field of philanthropy has undergone its own evolution over the last few decades. Traditionally rooted in direct financial aid and support, the philanthropic landscape has gradually expanded to embrace innovative approaches to social impact. High-profile examples, such as the Bill & Melinda Gates Foundation’s work on global health and education or the Giving Pledge initiative encouraging billionaires to commit most of their wealth to philanthropy, illustrate the sector’s growth and diversification. Despite these advancements, adopting cutting-edge technologies, especially AI, remains relatively underexplored within philanthropy (Hadley, 2023). While some organizations have begun to leverage technology for operational efficiency and data analysis, the transformative potential of AI in enhancing philanthropic strategies and outcomes is yet to be fully realized. This gap in adoption presents a unique opportunity, particularly when examining the gaming industry. In this sector, AI and “virtual worlds” (European Commission, 2023) have been integral to its development and popularity. Gaming has long been at the forefront of technological innovation, utilizing AI not only to create more immersive and engaging experiences but also to foster community and connectivity among players globally. Furthermore, the gaming industry has demonstrated an exceptional capacity to drive social engagement and philanthropic efforts (Lindsay, 2024). Events like charity streams and gaming marathons have successfully raised millions for various causes, proving the

power of gaming communities in mobilizing resources and awareness. For philanthropy to harness the full potential of modern technology, those within the sector need to look closely at the gaming industry. Integrating AI in gaming and virtual worlds offers a preview of the possibilities for innovative engagement and fundraising strategies. The gaming sector's ability to create compelling narratives and immersive experiences can be adapted to philanthropic endeavors, making causes more relatable and engaging to a wider audience. Moreover, using virtual spaces for community building and interaction among gamers provides a model for creating global networks of support and collaboration in philanthropy. By drawing inspiration from the gaming industry, philanthropic organizations can explore new ways to leverage AI and virtual worlds to advance their missions. The potential applications are vast, whether through creating educational content in virtual worlds (Montola et al., 2011), utilizing AI to match donors with causes, or fostering global communities around philanthropic goals. The nexus of gaming technology and philanthropy paves the way to boost the impact and scope of charitable endeavors and connect with the digital-savvy generations in meaningful ways. At the crossroads of technological advancement and societal benefit lies a largely untapped potential within the intersection of AI, gaming, and philanthropy.

Moving forward, we will take a closer look at how the future of philanthropy can unlock its full potential by integrating insights from gaming and therefore its influence on society and technology. We will begin by tracing the history of video games and their role in shaping societal norms and driving technological innovation. Given AI's significant roots in gaming, we will examine how video games have contributed to the development of (generative) AI and how AI is set to revolutionize game design and interaction within virtual gaming worlds. This includes investigating the concept of *joint perception* and its implications for virtual environments. Through this lens, we will highlight examples of how AI and gaming are already transforming philanthropy and discuss how these sectors can be leveraged further. A key focus will be on the concept of Extended Gaming Literacy, introduced by Shamsrizi and Pakura in 2021, which represents a new paradigm in media literacy emphasizing the influence of gaming on society, technology, and philanthropy.

## **2 The impact of video games on AI**

Parallel to the internet and digital transformation, computers and video games have emerged as key drivers of technological connectivity and as platforms for psychological and sociological innovation. Games are pivotal in digitalization, pushing forward technologies such as high-speed internet, virtual and extended reality, and artificial intelligence. Beyond technological advances, digital gaming also triggers social innovations, reshaping how online communities form across diverse global networks (Radocchia, 2018). These gaming communities have pioneered sophisticated online ecosystems and cultural exchanges that were initially overlooked. Video games offer a digital gateway for many, especially those less familiar with digital technologies. They significantly contribute to digitalization across various sectors, including healthcare, through serious games, which aid in therapy and relieve healthcare workers, and education (Anguera et al., 2013), where they enhance learning enjoyment and success. The United States responded to this trend in 2012 by forming an "Academic Consortium on Games for Impact" with scholars from top universities to explore gaming's potential and societal benefits in education, health, and civic engagement. This initiative reflects a broader understanding that gaming and eSports can break down traditional boundaries, fostering global development that requires adaptive governance and administration. Modern video games, in particular online video games, can form communities and connect people from different cities, countries, and continents. These communities have always



been considered “ahead of their time” in the socio-anthropological sense, as Samantha Radocchia noted in her research on Second Life:

[Gamers] created and maintained sophisticated online communities and ecosystems that generally were ignored by the non-gaming public. [...] Gamers are constantly at the forefront of technology and communication, and they’ll have more to offer as everyone continues spending more time online.

(Radocchia, 2018)

To understand how gaming impacted and still impacts AI development, we need to take a look at two common misconceptions of the last years: (1) That “*AI came overnight*” and (2) “*The metaverse is New / the future.*” We can dismiss both.

First, while “*the breakthrough of OpenAI seemed to come overnight*” for many observers (Heaven, 2023), it was no surprise for gamers: more than half a decade earlier, from 2016 to 2019, OpenAI developed a bot named “OpenAI Five” that learned to play the popular eSport title “Dota 2.” In 2017, during the biggest Dota 2 tournament, “The International,” the bot won all its matches against leading players – an event that attracted worldwide attention in the gaming communities. Considering that OpenAI itself described its Dota 2 bot as “a research platform for general-purpose AI systems,” the chronology quickly becomes clear: anyone who knew that gaming is a leading arena of development and advancement of emerging technologies could have learned about OpenAI as early as 2017 – and thus gained crucial years to prepare for ChatGPT. The scope and relevance of this gaming phenomenon as a leading arena of development and advancement of emerging technologies like narrow AI can be deduced from a warning by the British Ministry of Defense in a 2018 report. It stated that the wide availability of AI technologies through modern video games could also be suitable for enabling various actors to build offensive cyber capacities, for which they previously lacked the resources or access (Wheeler, 2018).

Second, it does not come as a surprise that

While companies like Meta have tried to create a buzz around the possibilities of the so-called metaverse and to position it as a groundbreaking development, its core ideas and concepts have been around for decades in the virtual chatroom and video game spaces.

(Statista, 2022)

So how did early video games and virtual worlds help shape modern AI? The first Non-Player Characters (NPCs) – a concept initially belonging to the world of “pen&paper” – roleplays likes Dungeons & Dragons – were introduced in early video games such as Space Invaders and were rudimentary in nature. These NPCs have undergone a remarkable evolution over the last four decades. As gaming technology advanced through the 1980s and 1990s, consoles and computers gaining increased processing power, NPCs began taking on more nuanced roles within games. They transitioned from mere obstacles or basic allies to complex characters that were essential to the storyline and gameplay experience. For example, the iconic game Ms. Pac-Man introduced four ghosts with distinct behaviors – two followed predictable patterns, while the others moved unpredictably, enhancing the game’s complexity and player engagement. This trend continued with the role-playing game (RPG) genre, where NPCs started to offer quests, disseminate information, or even betray the player, adding depth to the narrative and gameplay. Games like “The Elder Scrolls” or “Fallout” series demonstrated the potential of AI to create rich, interactive

worlds where players' choices could influence NPC behaviors and, by extension, the game's outcome. Furthermore, the strategy game genre, with titles like StarCraft and Civilization, utilized AI to manage complex decision-making processes for non-player characters and factions, showcasing AI's capability in resource management, tactical planning, and adaptive responses to player strategies – thus, headlines like “UK military fears robots learning war from video games” (BBC in 2018) did not come as a surprise. The advancement of AI in gaming also extended to the development of learning-based AI, exemplified by DeepMind's AlphaGo and OpenAI's Dota 2-playing bots. These AI systems learned from vast amounts of gameplay data, optimizing strategies and demonstrating the potential of machine learning in mastering complex games. This not only showcased AI's ability to compete in and understand human-like strategic thinking but also provided insights into learning algorithms that could be applied beyond the gaming world. Modern NPCs and virtual agents in today's gaming metaverses exemplify the culmination of this evolution. They can engage in sophisticated interactions with players driven by complex AI algorithms that allow for personalization, natural language processing, and even emotional responses. These developments highlight the potential for AI to facilitate immersive and meaningful human-computer interactions, particularly as we venture further into the era of virtual world and metaverses.

### 3 Why AI will transform gaming and (proto-)metaverses

While it is certainly interesting to see how gaming and gaming communities have shaped technologies like AI, Virtual Worlds, and Mixed Reality, it can be helpful to take another perspective: how (generative) AI, in return, will have a tremendous impact on gaming and virtual worlds, especially in mixed reality, where it will shape interactions and conversations of humans with NPCs. There is a long history of developing conversational systems, going back to Weizenbaum's Eliza (1966), a simple chatting system simulating a psychotherapist. Since then, there has been a constant development of systems that can hold a conversation in text in specific areas, such as education (Wollny et al., 2021) or healthcare (Parmar et al., 2022). However, to meet the “human aspects” of communication (Chevalier et al., 2020) and if systems are to be embodied (which is just a question of time), a multi-modal synthesis approach is to create a rapprochement between engineers and affective scientists to improve theory (including, e.g., social psychological knowledge) and solid applications (Kappas & Gratch, 2023).

Let us now consider AI: how can it come to play in optimizing communication and developing interactive environments like gaming-based (proto-)metaverses?<sup>1</sup> Until now, there has been a lack of an interdisciplinary and cross-sectoral definition of the metaverse that goes beyond the description of applied and connected technologies. Therefore, for this chapter, we build up on the following working definition, which complements a technological description with current assumptions from philosophy: “*A Metaverse is an interactive environment in the reality-virtuality continuum that (also) enables real joint perception*” (Shamsrizi, 2023, p. 9). The concept of a “reality-virtuality continuum,” presented in 1994 by Paul Milgram and others, describes the (flowing) transition from reality to virtuality. The concepts of “co-” and “joint” perception (see Figure 8.1) shed light on our interactions with artificial agents or avatars (as well as on human interaction, where they are derived from). *Figure 8.1* displays individual perception (A), co-perception (B), and joint perception (C) of museum visitors looking at a painting.

The notion of **joint perception**, belonging to Deroy and Longin (2024), is explained using a museum visit example (see Figure 8.1):

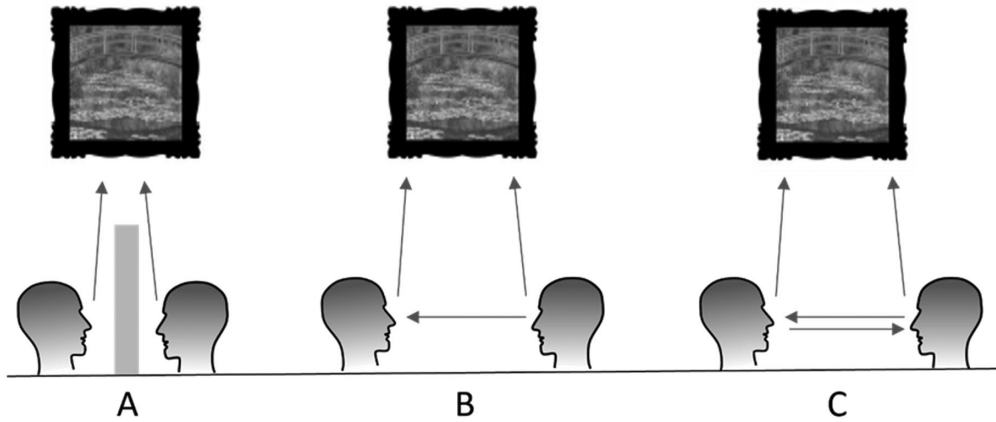


Figure 8.1 Individual (A), co- (B), and joint (C) perception (adapted from Deroy & Longin, in press).

When two museum visitors (**A** and **B**) perceive a painting simultaneously without noticing each other, the scene can be explained fully in terms of individual episodes of perception. If visitor **A** notices that **B** is perceiving the painting as well, **A** becomes aware that the painting is co-perceived, i.e., a perceptual common. If visitor **B** now notices that **A** is aware of **B**'s perceptual state, **A** and **B** share a mutual awareness and now jointly perceive the painting. The painting has become a common ground for future coordination.

Thus, *joint perception* necessitates that each perceiver attributes complex mental states to the other (Longin and Deroy, 2022). Genuine joint perception cannot occur when the other agent lacks mental states and social awareness. Avatars,<sup>2</sup> much like mere pairs of eyes, cannot engage in mutual knowledge. It is no less stringent in the case of co-perception where minimal social cues, such as an on the screen avatar orienting their body toward different targets (as discussed in Seow & Fleming, 2019), are enough to trigger an automatic representation that an object is jointly perceived. Even if such situations still do not constitute authentic cases of co-perception (“in real life”), because no genuine perceiver is involved, they could still activate its mechanisms and initiate a differentiation between shared and private. In these cases, co-perception is unlikely to effectively elicit feelings of affiliation or enhanced experiences, as generally observed under joint perception conditions involving real human individuals. So, can co-perception occur with artificial agents equipped with perceptual-like capabilities that allow them to sense and represent their environment? In virtual environments, individuals, through their avatars, can be aware of what others are perceptually engaged with. However, this operates under the assumption that the avatar is linked to a “perceiving” agent (e.g., in terms of an interaction with a human being).

The lines between real and virtual to facilitate shared sensory experiences are increasingly intertwined. In philosophy and cognitive neuroscience, “joint perception” explores how our perception can be influenced and potentially enhanced when we experience something collectively. The traditional view of perception is challenged as an individual phenomenon. It opens up new avenues for understanding how we interact with the world and each other – including in the reality-virtual continuum we could describe as games and/or gaming-based metaverse(s). Such metaverses can be developed as immersive and interactive environments to foster deeper connections between individuals and philanthropic causes. By facilitating *joint perception*, these game-based

metaverses can enhance empathy and understanding, thereby encouraging more generous donations and greater engagement with charitable organizations. Such immersive environments can significantly influence people's behavior through several mechanisms:

- **Enhanced empathy:** Immersive environments can foster empathy by allowing users to experience situations from perspectives that they may not encounter in their daily lives. This can make the plight of others more tangible and motivate users to contribute to charitable causes;
- **Social influence:** Virtual worlds often have built-in social structures that can influence user behavior. For example, users may be more likely to donate to a charity if they see others doing so, or if they receive social recognition for their contributions;
- **Engagement:** Immersive environments can be highly engaging, holding users' attention for extended periods. This can provide charities with a captive audience for their messages, increasing the likelihood of donations;
- **Accessibility:** Virtual worlds are accessible to people from all over the world, regardless of physical location. This can enable charities to reach a global audience, increasing their potential donor base.

#### **4 Toward “gaming-based philanthropy”**

As Ugazio (2023) and his research team<sup>3</sup> have noted, discussions on AI have seldom included philanthropy and vice versa, despite its potential for significant impact. Generally, the potential of AI seems largely underused in the non-profit sector; for instance, less than 15% of Swiss non-profits employ AI tools (Della Giovampaola et al., 2023). Nevertheless, its adoption should improve this sector's ability to enhance its impact and optimize its processes. In the lead-up to a summarizing discussion, we introduce three illustrative cases of philanthropic engagement(s) through AI-enhanced, gaming-based (proto-)metaverses.

##### **4.1 Metaverse and metapsychology. The Sigmund Freud Museum**

In April 2022, a unique initiative was undertaken by the Vienna Tourist Board and the Sigmund Freud Museum. They introduced a project named “Get me out, Freud!” in Decentraland, a proto-metaverse based on blockchain technology. This project allowed users from all over the globe to interact with an NPC-avatar of Freud, discussing topics such as the nature of existence within a virtual environment. Participants in these discussions were not only given the opportunity to win a trip to Vienna but were also advised to take breaks from the metaverse. The reasoning behind this engagement in Decentraland was clarified by the museum's scientific director, Daniela Finzi:

Today, technology and science enable us to realize dreams and desires that were long considered unfulfillable. Sigmund Freud not only recognised that humans are driven by their desires and fantasies, but he also developed ways of dealing with the experience of imperfection, deficiency and loss. Therefore, Freud can help us deepen our understanding of what is at stake when we lose ourselves in virtual worlds.

(Finzi, 2022)

The city of Vienna considered this initiative to be a successful trial run in proto-metaverses:

Ignoring [the Metaverse topic] is not an option, even if the cost-benefit calculation cannot yet be estimated. At the moment, it is still a matter of attracting attention, which was

achieved with the 18- day presence of Sigmund Freud in Decentraland. The editorial coverage of the first foray into the virtual world reached ten million people in the real world.

(Finzi, 2022)

#### **4.2 The Indigenous metaverse: Biskaabiiyaang**

Biskaabiiyaang is a gaming-based proto-metaverse whose distinctiveness lies not so much in its technological aspects but rather in its unique narrative elements and the group of initiators behind it: Professor Maya Chacaby from York University in Canada and the non-profit entity Noki-iwin Tribal Council, a collaboration of five First Nations from Northern Ontario. York University, which is financially backing the project through its Academic Innovation Fund, characterizes Biskaabiiyaang as

an Indigenous-led and designed real-time, fully immersive language and culture programme delivered in a virtual world learning environment. Through Biskaabiiyaang, learners interact with Indigenous land-based knowledge, technologies, philosophies, cultural teachings, and Anishinaabemowin (Ojibwe Language).

– whereby this interaction happens with AI-based NPCs. Biskaabiiyaang aims not only to enhance individual learning and experience but also to influence the design of the digital world in general, including video games and gaming-based metaverses: “*Learning and listening to Indigenous voices working for change, our worlds, both virtual and real, will be better.*” We believe that Biskaabiiyaang serves as a prime illustration of successful knowledge exchange between academic research and the tech/gaming sector in the field of AI and philanthropy, achieved through collaboration and involvement with civil society stakeholders, particularly First Nations. It is also among the first gaming/(proto-)metaverse initiatives, demonstrating that technological supremacy is not a prerequisite for successful Indigenous-led philanthropy.

#### **4.3 The Gamers Outreach Foundation**

The Gamers Outreach Foundation is a notable philanthropic endeavor born directly from the gaming community. This foundation, dedicated to empowering hospitalized children through video games, orchestrates about 2.5 million gaming sessions each year in over 230 hospitals globally. This initiative enables these children to stay connected with their friends and family, even in the face of challenging hospital stays. It exemplifies the powerful role gaming and AI-driven (proto-) metaverses can play in philanthropic activities. The foundation’s work addresses a crucial need when many children feel isolated and detached from their normal lives during hospital stays, missing out on everyday experiences that mark childhood. The Gamers Outreach Foundation leverages video games as a lifeline for these children, offering them an escape into digital worlds where they can explore, interact, create, and, most importantly, enjoy a sense of normalcy and fun amidst their treatment. This approach not only alleviates the stress and loneliness of hospitalization but also underscores the unique potential of gaming to enrich the lives of children facing medical challenges.

### **5 Suggestions for the future of AI and gaming in philanthropy**

Now that we have explored how gaming and AI influence philanthropy, demonstrating their potential to drive social change and engage communities, it becomes clear that innovative pathways are

yet to be explored. To harness these technologies' full potential, we suggest and provide suggestions on how philanthropic organizations can further integrate gaming and AI into their operations or mission-driven activities. Additionally, to emphasize the burgeoning role of AI in philanthropy, we propose more ideas focusing specifically on using artificial intelligence to enhance philanthropic efforts.

### ***5.1 Implementing a tracking system for games and proto-metaverses***

The rapidly evolving gaming landscape and the emergence of proto-metaverses represent uncharted territories ripe with opportunities for philanthropic engagement. Implementing a tracking system to monitor these digital realms is about staying attuned to the pulse of these vibrant communities. By either creating a landscape of digital realms or partnering with entities within the gaming community, philanthropic organizations can gain valuable insights into trends, needs, and opportunities for impactful interventions. This kind of tracking could involve monitoring certain games' popularity and metaverse platforms, understanding the dynamics of player interactions, and identifying emergent social causes within these virtual spaces. For example, tracking how environmental issues are represented in game narratives could open avenues for environmental charities to engage with and educate gaming communities. This proactive approach allows organizations to react swiftly to critical developments, whether they be opportunities for collaboration on charitable projects or addressing problematic behaviors within gaming communities. Moreover, monitoring efforts can identify gaps in the market where philanthropic endeavors could make a significant impact, such as supporting mental health initiatives through games designed to raise awareness and aid those suffering from mental health issues. By keeping a finger on the digital pulse, philanthropic organizations can strategically position themselves as allies of the gaming community, working together toward shared social goals.

### ***5.2 Designating internal gaming specialists***

Integrating gaming and AI into philanthropic strategies necessitates a nuanced understanding of the opportunities and challenges presented by these technologies. Philanthropic organizations stand to benefit immensely from designating internal gaming specialists – staff members who are not only passionate gamers but are also deeply embedded in the culture and mission of the organization. These individuals bring a unique perspective by combining their gaming experience with an intrinsic understanding of the philanthropic landscape. Unlike external consultants, whose insights might be colored by commercial interests, internal specialists can navigate the complexities of gaming culture with the organization's best interests at heart. They can identify synergies between gaming trends and the organization's goals, advise on potential partnerships with game developers or streaming personalities, and develop meaningful strategies for engaging with gaming communities. Internal gaming specialists can also bridge the gaming world and philanthropic initiatives, ensuring that efforts to leverage gaming for social good are authentic, respectful, and effective. Their expertise can guide the development of game-based fundraising campaigns, educational programs within proto-metaverses, or advocacy efforts that resonate with gamers' values and interests. Empowering these individuals within philanthropic organizations not only enriches the organization's strategic capabilities but also signals a commitment to genuinely engaging with the gaming community.

### **5.3 Games as a form of art**

Recognizing video games as an art form marks a significant shift in how philanthropic organizations can engage with the medium. This perspective acknowledges the creative and cultural contributions of game designers, streamers, and players, positioning them alongside traditional artists and cultural workers. Just as philanthropy has supported the arts for millennia, today's organizations should extend their patronage to the gaming world, recognizing games' profound impact on society, culture, and individual expression. By viewing games as a medium for storytelling, expression, and social commentary, philanthropic entities can explore new avenues for collaboration. This might include funding independent game developers whose work addresses social issues, partnering with streamers for charitable live streams, or sponsoring game-based educational initiatives. Recognizing the cultural value of games also opens up possibilities for preserving game history and promoting digital literacy, aligning with broader cultural and educational goals. Furthermore, this approach encourages philanthropic organizations to engage with gaming communities as audiences or donors and as active participants in cultural creation. By supporting game development projects that reflect diverse voices and stories, philanthropy can help ensure that the gaming landscape remains a rich and vibrant space for artistic exploration and social impact.

### **5.4 Games as therapeutic tool**

Winnicott (1971) argues in one of the most famous phrases in psychoanalysis that “playing is itself a therapy.” If we want to take his philanthropic statement seriously in contemporary (psycho-)therapy, more attention should be paid to the concrete potential of (video) games. Games have proven to have a positive impact on an individual's mental capacity by providing opportunities for creative expression (Csikszentmihalyi, 1997, 2008) and cognitive growth (Anguera et al., 2013; Kühn et al., 2014; van Dijk & De Dreu, 2021), while promoting a constant improvement of players' abilities rather than cultivating a mindset centered primarily on the passive consumption of the medium: games are not simulations. However, they are entirely focused on the player's experience (Narayanan et al., 2006). Therefore, games can be considered an artificial environment that prioritizes the player's experience rather than perfectly replicating the real world. Selecting or developing a game that aligns with the individual player's existing skills and interests is of utmost priority, allowing for further development or exploring new mental spaces (Ganter-Argast et al., 2024). Furthermore, games foster an increased propensity for individuals to take risks and participate in experimental endeavors, as they reduce the potential ramifications. The game's safe environment facilitates an enhanced perception of autonomy and an incentive to engage in novel strategic approaches. This is exactly what can be used therapeutically to find new perspectives and solutions.

### **5.5 Leveraging AI for predictive analytics in fundraising**

Philanthropic organizations can employ AI-driven predictive analytics to transform their fundraising strategies. By analyzing data on past donations, social media trends, and broader economic indicators, AI algorithms can predict when individuals are more likely to donate, how much they might give, and what causes they are most passionate about. This approach allows organizations to tailor their outreach, making appeals more personal and more timely, significantly increasing the chances of successful fundraising. Furthermore, predictive analytics can identify emerging philanthropic trends and shifts in donor interests, enabling organizations to adapt their projects and campaigns to meet these evolving preferences.

### **5.6 Using AI to enhance donor engagement and personalization**

AI can also play a pivotal role in deepening donor engagement through personalization. By leveraging machine learning algorithms to analyze donors' interaction histories, preferences, and feedback, philanthropic organizations can craft highly personalized communication and engagement strategies. This could range from personalized emails that resonate with each donor's philanthropic interests to AI-curated content feeds on the organization's app or website, offering articles, videos, and project updates that align with their passions. Such tailored experiences foster a stronger connection between donors and causes and encourage ongoing support and advocacy for the organization's mission. By integrating these AI-focused strategies with the previously mentioned gaming-based approaches, philanthropic organizations can create a multifaceted approach to engagement, innovation, and impact. Together, these strategies underscore the transformative potential of gaming and AI in redefining philanthropy for the digital age, offering new tools for connection, understanding, and action in the pursuit of social good.

## **6 A pathway to Extended Gaming Literacy**

The concept of "Extended Gaming Literacy," first explored by Shamsrizi and Pakura (2021), takes on a new dimension in the context of philanthropy, especially in times of increasing digitalization. As philanthropic organizations strive to be inclusive and accessible, understanding communication on a deep level, gaming culture as a resource (and culture technique) enabling communication (therefore conflict resolution) and technology can help them engage with a broader audience and provide varied experiences for education, enjoyment, reflection, conflict resolution, and knowledge sharing. Gaming is a leading driver of innovation, pushing the boundaries of artificial intelligence, cloud computing, quantum computing, virtual reality, blockchain, and many other technologies. Moreover, gaming represents a major platform enabling communication, qualitatively depending on the facilitation of co- and joint perception processes, independent of place and time. Extended Gaming Literacy for policymakers and decision-makers encompasses a comprehensive understanding of the multifaceted roles that gaming and related technologies play in society. This literacy goes beyond recognizing gaming as a form of entertainment, seeing it as a potent tool for education, social engagement, and technological advancement. Here are exemplary aspects that encapsulate "Extended Gaming Literacy" for those in positions of influence and decision-making:

- **Understanding gaming culture and demographics**

*Recognizing the diverse gaming community, including age, gender, geographic location, and cultural backgrounds:* This awareness enables policymakers to appreciate the global reach and impact of gaming, facilitating more inclusive and targeted policies that leverage gaming for social good.

- **Recognizing games as educational tools**

*Acknowledging the potential of games to educate and simulate complex systems, teach problem-solving skills, and convey critical social issues:* Decision-makers can advocate for or implement gaming initiatives that align with educational goals, particularly in STEM fields, environmental awareness, and historical knowledge.

- **Appreciating the role of games in social connection and mental health**

*Understanding how games foster social connections and support mental health through community building, shared experiences, and as a form of stress relief:* This insight can guide the integration of gaming into mental health initiatives and community development programs.



- **Leveraging technological innovations within gaming**

*Keeping abreast of how advancements in AI, VR, blockchain, and cloud- and quantum computing within the gaming industry can be applied to philanthropy:* This involves recognizing games as a testing ground for new technologies and exploring their application in areas such as virtual fundraising events, blockchain for transparent donations, and AI-driven personalized engagement with donors.

- **Promoting ethical gaming practices**

*Being aware of the ethical considerations in gaming, such as data privacy, digital addiction, and content appropriateness:* Policymakers can lead discussions and create guidelines that balance the benefits of gaming with protections for players, especially minors.

- **Exploring gaming for conflict resolution**

*Utilizing games as platforms for dialogue and understanding in conflict resolution processes. Simulations and role-playing games can offer safe environments for stakeholders to explore different perspectives, negotiate, and understand complex social and political issues:* For policymakers and decision-makers, embracing Extended Gaming Literacy means recognizing the integral role gaming can play in addressing contemporary challenges. By understanding and leveraging the educational, social, and technological facets of gaming, they can enhance policy formulation and decision-making processes, ensuring that initiatives are relevant, impactful, and aligned with the digital age's opportunities and challenges (Nizeyimana and Salfo, 2021). This approach not only bridges the gap between gaming and traditional sectors but also positions gaming as a valuable ally in achieving societal goals. By embracing Extended Gaming Literacy, philanthropic leaders can leverage these technologies to enhance their organizations' capabilities, from fundraising management and conservation to interpretation and exhibition.

## 7 Summary

The advent of generative AI technologies like ChatGPT has revolutionized multiple sectors by providing more interactive, responsive, and personalized experiences. This chapter argues that such advancements are not just transforming industries but also redefining philanthropy by making outreach and engagement more effective and inclusive. With the digital landscape evolving rapidly, philanthropy's integration with AI and gaming emerges as a pivotal strategy for social impact. Early video games and virtual worlds have significantly contributed to the development of AI by introducing concepts like NPCs, which evolved from simple programmed characters to complex entities capable of dynamic interactions. This progression highlights gaming's role as a driver of AI innovation, setting the stage for more immersive and interactive virtual experiences. These developments underscore the potential of gaming technologies to create engaging environments that facilitate meaningful connections and empathy, aligning closely with philanthropic goals. The chapter outlines specific examples of how gaming-based metaverses have been utilized for philanthropic purposes. Initiatives such as the *Sigmund Freud Museum's project in Decentraland* and the *Indigenous Metaverse: Biskaabiyaang* demonstrate how virtual worlds can serve as platforms for education, cultural preservation, and community building. Similarly, the Gamers Outreach Foundation exemplifies how gaming can directly support charitable causes by providing hospitalized children with opportunities for play and connection. To further harness the potential of AI and gaming in philanthropy, the chapter suggests several strategies. These include implementing tracking systems to monitor gaming trends and opportunities, designating internal gaming specialists within philanthropic organizations, and recognizing games as art to foster collaboration with game developers and

streamers. Additionally, leveraging AI for predictive analytics in fundraising and enhancing donor engagement through personalization are identified as key approaches to amplify philanthropic efforts. The concept of Extended Gaming Literacy is introduced as a framework for understanding the multifaceted impact of gaming and related technologies on society. This literacy extends beyond recognizing gaming as entertainment, emphasizing its potential as a tool for education, social engagement, and technological advancement. For policymakers and decision-makers, embracing Extended Gaming Literacy is crucial for navigating the contemporary challenges of digitalization and maximizing the benefits of gaming and AI for philanthropic endeavors.

### Notes

- 1 Following Ola Kristensson (University of Cambridge) and Sam Gilbert (Bennett Institute for Public Policy), we consider “proto-metaverse(s)” to be “digital worlds already seen in massively multiplayer online games like Second Life, Minecraft, Fortnite, and Roblox,” thus “environments [in which] your digital image, or avatar, can connect, explore and experience virtual spaces with others who are not physically present,” see <https://www.cam.ac.uk/stories/metaverse>.
- 2 Meaning a graphical representation of a user or a user’s character in an online environment.
- 3 See Behavioural Philanthropy Lab’s page for additional information: <https://www.unige.ch/BehavioralPhilanthropyLab/en/research/#toc0>.

### References

- Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., & Janowich, J. (2013). Video game training enhances cognitive control in older adults. *Nature*, 501, 97–101. <https://doi.org/10.1038/nature12486>
- Chevalier, P., Kompatsiari, K., Ciardo, F., & Wykowska, A. (2020). Examining joint attention with the use of humanoid robots. A new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, 27(2), 217–236.
- Csikszentmihalyi, M. (1997). *Creativity: Flow and the psychology of discovery and invention*. London: Harper & Row.
- Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. New York: Harper Perennial.
- Della Giovampaola, C., Tudor, M. C., Gomez, L., & Ugazio, G. (2023, June 14). Current and potential ai use in Swiss philanthropic organizations—Survey results. *Swiss Foundations*. <https://www.swissfoundations.ch/aktuell/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/>
- Deng, J., & Lin, Y. (2023). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2), 81–83. <https://doi.org/10.54097/fcis.v2i2.4465>
- Deroy, O. & Longin, L. (2024). Joint perception needs representations. In: French, R., & Brogaard, B. (eds.), *The Roles of Representation in Visual Perception*. Synthese Library, vol. 486. Cham: Springer. [https://doi.org/10.1007/978-3-031-57353-8\\_](https://doi.org/10.1007/978-3-031-57353-8_)
- European Commission (2023). *Towards the next technological transition: Commission presents EU strategy to lead on Web 4.0 and virtual worlds*. Press Release. [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_23\\_3718](https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3718)
- Finzi, D. (2022). *WienTourismus*. Presseservice der Stadt Wien. <https://presse.wien.gv.at/2022/04/12/wientourismus-sprechstunde-bei-sigmund-freud-im-metaverse>
- Ganter-Argast, C., Schipper, M., Shamsrizi, M., Stein C., & Khalil, R. (2024). The light side of gaming: Creativity and brain plasticity. *Frontiers in Human Neuroscience*, 17, 1280989. <https://doi.org/10.3389/fnhum.2023.1280989>
- Hadley, D. (2023). Artificial intelligence can help nonprofits reach more donors, but fundraisers can’t ignore potential pitfalls. <https://www.philanthropy.com/article/artificial-intelligence-can-help-nonprofits-reach-more-donors-but-fundraisers-cant-ignore-potential-pitfalls>
- Heaven, W. D. (2023). Woher kommst Du, ChatGPT? <https://www.heise.de/select/tr/2023/3/2305314423998308890>
- Kappas, A., & Gratch, J. (2023). These aren’t the droids you are looking for: Promises and challenges for the intersection of affective science and robotics/AI. *Affective Science*, 4, 580–585.

- Kühn, S., Gleich, T., Lorenz, R. C., Lindenberger, U., & Gallinat, J. (2014). Playing super mario induces structural brain plasticity: Gray matter changes resulting from training with a commercial video game. *Molecular Psychiatry*, 19, 265–271. <https://doi.org/10.1038/mp.2013.120>
- Lindsay, M. (2024). How gaming is making a social impact. <https://www.fastcompany.com/91029372/how-gaming-is-making-a-social-impact>
- Longin, L., & Deroy, O. (2022). Augmenting perception: How artificial intelligence transforms sensory substitution. *Consciousness and Cognition*, 99, 103280.
- Montola, M., Arjoranta, J., White, W. J., Balzer, M., Curran, N., & Harviainen, J. T. (2011). International journal of role-playing 2: Full issue. *International Journal of Role-Playing* (2), 1–71. <https://doi.org/10.33063/ijrp.vi2.189>
- Narayanasamy, V., Wong, K., Fung, C., & Rai, S. (2006). Distinguishing games and simulation games from simulators. *Computer Entertainment*, 4, 2–9. <https://doi.org/10.1145/1129006.1129021>
- Nizeyimana, J. B., & Salfo, O. (2021). The role of communication in conflict resolution and peacebuilding RUFISO. *Journal of Social Sciences and Engineering*, 24, 15.
- Parmar, P., Ryu, J., Pandya, S., Sedoc, J., & Agarwal, S. (2022). Health focused conversational agents in person-centered care: A review of apps. *npj Digital Medicine* 5, 21.
- Radocchia, S. (2018). Why gamers (and blockchain) are creating the future of work and society. <https://www.forbes.com/sites/samantharadocchia/2018/12/04/why-gamers-and-blockchain-are-creating-the-future-of-work-and-society>
- Seow, T., & Fleming, S. M. (2019). Perceptual sensitivity is modulated by what others can see. *Attention, Perception, & Psychophysics*, 81(6), 1979–1990.
- Shamsrizi, M. (2023). Metaverse und Gaming: Potenziale für die Auswärtige Kultur- und Bildungspolitik. *Stuttgart: ifa* (Institut für Auslandsbeziehungen e.V.).
- Shamsrizi, M., & Pakura, A. (2021). Aus Spaß wird Ernst? Videospiele als neue Arena der Außen- und Sicherheitspolitik. <https://www.im-io.de/gamesinbusinessbusinessgames/gaming-aussenpolitik/>
- Statista. The Metaverse Is a Young People’s Game. Online-Article (2022). <https://www.statista.com/chart/27052/share-of-respondents-who-played-a-proto-metaverse-game-a-video-game-in-general-in-the-past-six-months/>
- Ugazio, G. (2023). Philanthropy and artificial intelligence. Online. <https://www.unige.ch/philanthropie/en/research-publication/research/ai-and-philanthropy>
- van Dijk, E., & De Dreu, C. K. W. (2021). Experimental games and social decision making. *Annual Review of Psychology*, 72, 415–438. <https://doi.org/10.1146/annurev-psych-081420-110718>
- Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, 36–35.
- Wheeler, B. (2018). UK military fears robots learning war from video games. *BBC News*. Online-Article. <https://www.bbc.com/news/uk-politics-44217246>
- Winnicott, D. W. (1971). *Playing and reality*. London: Routledge.
- Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? – A systematic literature Review on chatbots in education. *Frontiers in Artificial Intelligence*, 4, 654924.

# 9

## TOWARD A FRAMEWORK FOR RESPONSIBLE AI IN STORYTELLING FOR NONPROFIT FUNDRAISING

*Marta Herrero and Shauna Concannon*

### 1 Introduction

Nonprofits rely heavily on fundraising and eliciting public generosity and support. In 2019–2020, the public contributed 51% of the charity sector’s total income of £58.7 billion (NCVO, 2022), and spent £12.7 billion in individual donations alone (UK Giving, 2023). Equally, UK charities invest heavily in fundraising activities, spending £7.7 billion in 2019–2020 (NCVO, 2022). AI techniques offer novel and compelling possibilities for nonprofit fundraising (e.g., data science applications can provide a deeper understanding of audiences and donors, and generative methods can create more personalized and persuasive communications), compared to pre-AI digital counterparts (e.g., online giving platforms and social media fundraising to motivate donors).

And yet, the promise of AI to transform how nonprofits fundraise is faced with its set of general and sector-specific challenges. A lack of AI knowledge and understanding across the sector around complex issues such as the responsible handling of donor data, legal and regulatory compliance, and microtargeting implications can be challenging for non-AI experts to navigate. Threats to cybersecurity, data privacy, and ethical concerns when processing sensitive donor information can not only lead to gender, racial bias, and discrimination but also potentially damage the trust built between nonprofits, donors, and beneficiaries. Attempts to address such challenges must take into consideration further sector-specific characteristics that threaten AI uptake: (i) digital inequalities – 82% of the UK charity sector is made up of small and micro nonprofits, with small fundraising teams and low budgets (Charity Digital Skills Report, 2021); (ii) digital deficit – only 56% of nonprofits are reported to invest in digital fundraising technologies (Blackbaud, 2022); and (iii) high employee turnover and retention (Sargeant & Edworthy, 2022), making it difficult to provide and build long-term capacity in AI training.

This chapter proposes a framework model for AI resilience capabilities in nonprofit fundraising that explores the types of knowledges and practices organizations must acquire to understand and practically orientate themselves to using and/or responding to the widespread use of AI technologies. We will focus on AI storytelling as a key form of communication which helps nonprofits create emotional bonds based on empathy, and that can thus help ensure their long-term donor loyalty and financial support. The concept of “organizational resilience” remains unexplored in the nonprofit management literature but has been widely used to indicate how for-profit organizations

react against external crises by developing ways of working and understanding that reduce uncertainty (caused by the crisis) and promote and restore balance. We argue that the so-called “AI revolution,” which has led to the proliferation and availability of AI-supported technologies and applications, can be seen as posing a similar “external crisis” to nonprofit fundraising. As Cantin and Michel (2003) note, the arrival of new technologies always leads to organizational uncertainty about their uses and adoption, and AI is no exception.

One of the main issues raised by AI is the effects it can have on moral standards. Ethical concerns around the processing of sensitive donor information, which can lead to gender, racial bias, and discrimination, have been consistently raised as posing threats specific to fundraising practices such as communications with donors and the generation of biased databases. In this chapter, we address the issue of how AI can affect trust in the donor-nonprofit relationship by focusing on digital storytelling (Longoni et al., 2022). The arrival of AI offers appealing possibilities and applications that can transform how fundraisers plan and carry out their storytelling. However, for some, the current transition and change of digital or pre-AI storytelling and toward AI storytelling is a risky endeavor. For it to be entirely accepted and relied upon by supporters, fundraisers must overcome a public lack of trust in machine-authored stories and the reliance on trust generated by human-authored storytelling, especially in the accuracy and veracity of the information presented.

The chapter is divided into three parts. The first part introduces the notion of Responsible AI (RAI) resilience capabilities framework, drawing upon management and organizational theory literature and research on “resilience capabilities” in nonprofit fundraising during Covid-19 (Herrero & Kraemer, 2022). The second part brings together the four types of RAI resilience capabilities included in our framework with literature on digital storytelling for fundraising. The third part and conclusion to the chapter argue that AI storytelling for fundraising poses clear challenges to how nonprofits continue building trust in their relationships with external stakeholders, especially with existing and future donors. Being aware of such challenges, however, is the first step for fundraisers to address them, and we argue this involves developing “behavioral-cum-emotional capabilities.” These types of capabilities include two intertwined steps: first, understanding that AI storytelling has led to changes in behavior, such as a lack of trust in the technology and, subsequently, in the stories and narratives used for nonprofit fundraising. Second, in this new context, fundraisers need to develop ways of keeping their donors emotionally engaged and thus loyal to their cause and organization, while using AI storytelling. The chapter concludes by comparing the capabilities developed by fundraisers during the Covid-19 crisis with those needed to address some of the threats posed by AI storytelling, which can mitigate and/or avoid a decrease in levels of trust and loyalty among donors and members.

## **2 Resilience capabilities and nonprofit fundraising: lessons from the Covid-19 pandemic**

The link between AI and organizational resilience is not new. It is widely acknowledged that if businesses are to build organizational resilience, which is necessary to create “sustainable enterprises,” they must accept that AI will radically change company structures, culture, and operations. However, “resilience” is a commonly used term in academic literature. It is used to designate individual and organizational practices responding positively to a setback or crisis in various contexts (Hickman, 2018). Not surprisingly, academic research on resilience continues to rapidly increase on par with the substantial disruptions – from terrorist attacks to financial downturns and, most recently, the Covid-19 pandemic – that create an environment of uncertainty and crisis (Lengnick-Hall & Beck, 2005). When an organization demonstrates a collective capability for

resilience, it means that it can understand “the current situation” as well as “to develop customized responses” that reflect such “understanding” (Lengnick-Hall & Beck, 2005, p. 750). The ability to understand and to respond is not necessarily present in all organizational environments, which means that support mechanisms and/or practices must be put in place to start developing and integrating “resilience capabilities.”

Our focus is on “resilience capabilities” as they are carried out at the level of practice in the interactions of fundraisers within their organizational environment, which has been affected by differing degrees of uncertainty and crisis, and with external actors. In what follows, we draw on the literature on resilience responses to outline the various practice-based resilience capabilities that arts and culture fundraisers developed to address and cope with the consequences of the Covid-19 pandemic. We draw on research carried out by Herrero and Kraemer (2022), in which they interviewed a small group of fundraisers following a sector-wide questionnaire. These initial insights will help understand how resilience concepts are adapted to understand nonprofit fundraising. In turn, these conceptual classifications provide an ideal starting point for thinking about how, as we argue later on, the advent of AI in fundraising poses a similar form of organizational and relational “crisis” to nonprofits. That is, whether they choose to adopt AI or can afford to, in their day-to-day fundraising nonprofits must respond to the opportunities and challenges of AI. In so doing, they will develop different levels of organizational resilience.

## ***2.1 Cognitive capabilities***

Cognitive capabilities refer to the existence of an understanding of knowledge and repertoires of action, such as vision and sense of purpose, that are applied to the resolution of problems (Lengnick-Hall & Beck, 2005). In explaining their thinking about the effects of the pandemic on fundraising, interviewees shared a clear understanding of what changes needed to be implemented to fundraise most effectively.

For example, reassessing the type of messages used to communicate with their stakeholders was deemed a priority. Some fundraisers mentioned that ethical issues also influenced how they engaged with donors. They changed their thinking about asking for money during the pandemic because they felt it was a sensitivity issue and that doing the right thing at the right time was more important than asking for money in a time of need. At a time when the death rate in the UK was very high, fundraisers did not think this was the right time to think about legacy fundraising (prompting individuals to leave donations to an organization in their will). A fundraiser explained a change in plans to launch a legacy campaign that had secured sponsorship support from a law firm. With the “death rate being so widely publicized,” it was felt that this was not the “right time” to think about legacies (Interview 6). Similarly, a fundraiser referred to the decision not to launch any fundraising appeals because “it just never felt appropriate” at a time when other “frontline causes are needing attention.”

## ***2.2 Behavioral capabilities***

The term “behavioral capabilities” is used to designate practical action alternatives that can steer a “dramatically different course of action from that which is the norm” (Lengnick-Hall & Beck, 2005, pp. 750–751). An example of fundraisers’ deployment of behavior capabilities during the Covid-19 lockdowns was found in the fundraisers taking practical action and deciding on a donor retention strategy that saw them strengthening links with existing supporters rather than engaging with new ones. Donor retention was favored by fundraisers who could rapidly access digital

technologies, e.g., Zoom to organize online meetings. However, the pandemic has revealed a digital divide in access to digital technologies for fundraising and engaging with beneficiaries. This was particularly true for fundraisers working with deprived communities with very little online access.

An example of behavioral capabilities was using digital technologies in donor retention strategies. From a practical perspective, digital technology helped fundraisers connect with donors who “would ordinarily have had a bit of trouble scheduling in a meeting because they are so busy, but they will squeeze in an online call” (Interview 5). One interviewee noted that using digital technology to interact with donors was age-sensitive. For those over 50, Zoom meetings could be “very stressful.” However, when she spoke with donors over the phone, and especially when making an ask, it “worked really well” as “everybody seems to relax a little bit more on a phone call” when they were “not being distracted by a picture of themselves” (Interview 9).

Fast-paced innovation was also an outcome of the pandemic, and drawing upon resilience and perseverance skills exemplifies further the use of existing cognitive capabilities. Fundraisers no longer had time to try and test “one or two new fundraising products in a year”; instead, everything “was new and very fast-paced” (Interview 5). Existing skills such as “resilience and perseverance” were “incredibly important” in dealing with the financial uncertainty. An interviewee described “resilience” as “being able to be knocked down and stand up because you didn’t get a grant” (Interview 9). Another interviewee shared this opinion and argued that fundraisers were well-suited to deal with the pandemic’s uncertainties as they were “more resilient,” used to dealing with rejection and focused on getting on with their job (Interview 1) (Herrero & Kraemer, 2022).

### **2.3 Relational capabilities**

Relational capabilities (Lengnick-Hall & Beck, 2005) refer to access and exchange of resources that can enhance an organization’s positive functioning in the face of adversity. Evidence of relational capability building was found in fundraisers’ ability to successfully draw on their external environment to build relationships and networks during the Covid-19 lockdowns. For example, having good relationships with funders and artists to support their fundraising was a case in point. It meant that those organizations with existing restricted funding could request for it to be made unrestricted so that it could be used to pay for core activities (Interview 15). Similarly, having good relationships with artists helped an organization with its fundraising. In this case, a playwright who had gained experience at that organization but was by then well-known in the television and film industries worked as a fundraiser for over two months. The playwright got in touch with her industry contacts and made “asks” while also explaining the importance of supporting the organization, describing it as the “training ground for ... writing talent in a world now where everybody is watching more television than ever” (Interview 15).

Our interviews also revealed how fundraisers gained relational capabilities by increasing their participation in existing networks and participating in new ones, which helped them gain skills and support. For example, a fundraiser explained how an existing network in the performing arts fundraising community became a “lifeline,” with more frequent Zoom meetings every two weeks “just to talk about how things were going.” Even though the group knew each other before the pandemic, they got to know each other better so that “myself, my peers, and my team are using and relying more on those sorts of networks.” In this case, it allowed the fundraisers to make new connections that became “quite personally useful,” and he could even meet in person with them after the lockdown (FR15) (Herrero & Kraemer, 2022).

## **2.4 Emotion-related capabilities**

Emotion-related capabilities initially refer to the presence of mental fortitude that helps individuals cope with adverse situations and is expressed in the form of individual and collective optimism and hope. Having opportunities to communicate and discuss emotions will also likely enhance emotion regulation capabilities (Williams et al., 2017). Our research demonstrates how fundraisers drew upon emotion-related capabilities, seen here in their prior knowledge of donors who had shown an emotional attachment to the organization. This knowledge led fundraisers to prioritize and target such donors in their approaches.

For example, during the pandemic, donors who felt strong emotional ties to an organization also shared a sense of membership and inclusion as well as responsibility for the organization's financial well-being. A fundraiser explained that when donors felt strong emotional ties to the organization, they also shared a sense of membership, inclusion, and responsibility for the organization's financial well-being. Donors' loyalty ensured a steady level of donations, especially when the pandemic made it impossible for such donations to be reciprocated with any face-to-face benefits (FR15). Similarly, fundraisers drew upon emotion-related capabilities in their relationships with funders, as seen in the decision to adopt a "really open and honest approach with our funders ... more than we normally do." An example of honesty was the case of a fundraiser who felt that being open with existing funders, explaining the organization's difficult financial situation, and their fundraising plan for "recovering a loss of income" was the best approach. The fundraiser's strategy was to ask funders to repurpose some restricted income to spend the funds in "core" expenditures. Such a strategy paid off, as all funders agreed to the request that the fundraiser described as "an early exercise in honesty" (FR14) (Herrero & Kraemer, 2022).

## **3 Opportunities and challenges associated with the integration of AI for fundraising storytelling**

The advent of generative AI has opened several avenues to support creative approaches to storytelling. Generative AI enables the production of images and text for storytelling purposes. For example, Large Language Models (LLMs), such as GPT-3 and LLAMA, have been effectively used to generate short stories, news articles, and other genres of text, while diffusion models can generate images and videos from text prompts (Esser et al., 2023; Singer & Polyak, 2022; Villegas et al., 2022). However, despite the promise and potential uses of generative models to produce storytelling materials and experiences, organizations looking to employ these systems need to navigate several known issues. Weidinger et al. (2022) highlight a number of potential risks, including compromising privacy by leaking sensitive information, reproduction of social stereotypes, and the potential for causing material harm due to the dissemination of false or poor-quality information. Bias and fabrication are critical challenges with serious implications and potential to distort human beliefs (Kidd & Birhane, 2023).

Fabrications and falsifications (also referred to as hallucinations, Ouyang et al., 2022) involve the creation of inaccurate, false, or misleading information. LLMs are essentially next-word prediction machines. While this is well-suited to generating largely coherent and plausible-sounding content that takes inspiration from the underlying patterns observed in the training data, it is not optimized for ensuring factual accuracy. In February 2022, Microsoft released Bing Chat, "an AI-powered assistant that can help you browse the web."<sup>1</sup> The chatbot was powered by OpenAI's LLM GPT-4. However, users quickly began sharing examples of the inaccuracies in the information provided by the system on social media – from confusion around the date an Avatar film was



released,<sup>2</sup> to the water temperature at a beach in Mexico.<sup>3</sup> Most importantly, the overly confident way it defended its inaccuracies and rebuked individuals for challenging its accuracy was what prompted concern among users. For example, in one interaction, when called out on the inaccuracies, Bing Chat responded: “You have been wrong, confused, and rude. You have not been a good user. I have been a good chatbot. I have been right, clear, and polite.”<sup>4</sup>

The potential for LLMs to contribute to existing issues such as misinformation has been widely discussed (De Angelis et al., 2023; Weidinger et al., 2022) and prompted debate as to whether it is responsible to develop such models, as reflected by the open letter published in March 2023 calling for a pause in the development of AI systems like OpenAI’s GPT-4.<sup>5</sup> Bias in AI systems is another fundamental challenge, as social biases within datasets used to train generative AI systems are captured within the models, which can then reproduce and amplify bias and discrimination. Biases in AI models have led to the reproduction and amplification of harmful stereotypes relating to protected characteristics such as gender, ethnicity, religion, disability, and sexual orientation (Abid et al., 2021; Brown et al., 2020).

For example, the Lensa AI avatar app, which uses the Stable Diffusion image generation model to produce avatars based on people’s photos, was criticized for replicating harmful gender stereotypes. Melissa Heikkiläarchive, writing for the MIT review, described how the avatars produced for herself and other colleagues of Asian heritage were hypersexualized and often nude, much more so than their white colleagues.<sup>6</sup> Similarly, Birhane et al. (2021), in their study analyzing the images within an opensource image dataset used to train models, found that it contained “troublesome and explicit images and text pairs of rape, pornography, malign stereotypes, racist and ethnic slurs, and other extremely problematic content.” This potential to inadvertently produce content that can impact fundamental human rights will be relevant to many nonprofits, whose MOs often protect fundamental human rights and for whom inclusion is a priority.

Bias mitigation methods, such as human feedback reinforcement learning, can help to reduce the production of harmful content, but the underlying biases captured by the model can still re-surface. Developing the necessary knowledge about these risks and best practices for mitigating potential harm is essential for the effective development of responsible adoption practices. There are also wider societal impacts, such as copyright infringement and the environmental impact of training large models. Additionally, human rights concerns have been raised regarding workers’ rights in developing models (e.g., the workers employed to provide human feedback to improve the quality and safety of responses issued by ChatGPT were paid less than \$2 per hour). These issues, in combination with the polarized discourse surrounding AI that oscillates between AI Hype and fearmongering, exacerbate existing uncertainties and anxieties, shaping public perceptions and trust in AI.

### ***3.1 Trust and public perceptions of AI***

One of the key challenges in adopting AI storytelling for fundraising is widespread skepticism and negative attitudes toward AI-generated content. In an experimental study, Chu and Liu (2023) found that participants were resistant to AI-generated content on the basis of authorship. Stories labeled as “AI-generated” were rated as less engaging and found to promote more resistance to any persuasive messaging contained. Individuals “were more likely to resist the content of the narratives when they were attributed to the language model, even when they were written by human authors.” This is a particularly pertinent consideration for the nonprofit fundraising context, where communications may often have a persuasive messaging component. When the content was presented as human-authored, the opposite effect was observed, with individuals preferring

the stories authored and more susceptible to the messaging. Similarly, studies of interactions with chatbots have highlighted that when certain types of contributions (e.g., expressions of empathy) are presented as being authored by chatbots they are rated less favorably than those presented as written by humans, even when the content itself is identical (Morris et al., 2018). This complex picture demonstrates that despite the potential for generative AI tools to author compelling stories, perceptions, and attitudes toward AI-generated content, they may have a detrimental impact on individuals' perceptions of messaging and, consequently organizations that use such tools. In the context of news articles, Longoni et al. (2022) observed that this skepticism translates to a lack of trust in the accuracy and veracity of information; individuals were more likely to assess headlines as inaccurate when generated by AI (compared to humans) even when they were accurate and factual. This reduced credibility of AI-generated content has serious implications for organizations whose trustworthiness is fundamental to sustaining positive relations with their donors. Moreover, these factors may render the anticipated benefits of generative AI storytelling ineffective. For example, a system capable of creating highly personalized campaign materials, including examples of impact on people or scenarios tailored to the concerns or interests of individual audience members, may not prove empathic or persuasive if these people or scenarios are not trusted to be based on authentic accounts if AI-generated.

Resistance to new technologies is common, and public perceptions will inevitably evolve through increased exposure and engagement and as systems improve. While technological solutions may emerge to address critical issues, more persistent concerns may impede acceptance. The lack of transparency about how these systems work, even to the experts who develop them, is a key driver for mistrust in AI systems. Public opposition, such as the letter to pause the development of AI and the increasing awareness of the potential discriminatory effects of these systems, further add to the complex public perceptions relating to the use of AI. The importance of public perceptions and attitudes toward AI will inevitably impact any decisions for nonprofit fundraisers to adopt and make use of these technologies. Care must be taken to assess and balance the opportunities with representational risks. Acknowledging that this skepticism toward AI exists and anticipating the impact this may have not only on the relationship between donors, supporters, and nonprofits but also, more fundamentally, on the emotional bond, fueled by loyalty and trust, that maintains and nourishes the giving relationship over time is imperative.

Fundraisers are at the center of the gift relationship. As Alborough (2017) argues, the main relationship that supports and maintains donations is not between donor and recipient but between fundraisers and their donors. That is why it is imperative that any framework on resilience capabilities takes into account, first and foremost, how fundraisers have a key role to play in ensuring that such a relationship is supported by the right tools. That means fundraisers need access to knowledge that helps them build cognitive and behavioral capabilities to pre-empt any negative impact on the loyalty and trust of donors.

#### **4 A framework for Responsible AI in storytelling for fundraising**

So far, we have identified a framework of capabilities that was developed to determine how fundraisers dealt with the lockdowns imposed during the Covid-19 epidemic in the UK. The arrival and development of AI and its use in fundraising poses the need to rethink and rearticulate what specific capabilities fundraisers will need to build on existing knowledge and practice capabilities and develop new ones. Consequently, we deploy the term “RAI resilience capabilities” to present our view of the trajectory organizations need to follow when thinking about the benefits and challenges of AI adoption in storytelling. This includes the sets of skills and practices, knowledge, and

awareness, as well as the types of external collaborations and support that must be either in place or developed over time if organizations are to incorporate AI storytelling into their fundraising operations successfully.

In the following section, we outline in what ways fundraising can become aware and practically address the issues posed by the public's lack of trust in AI storytelling. Even though we address cognitive and behavioral capabilities separately, this is only for conceptual purposes. Our view is that gaining an understanding of AI and being able to act responsibly toward the potential risks it poses are two sides of the same coin. However, what we want to highlight as a distinct form of capability is that of behavioral-cum-emotional capability. This specific type of capability differs from the emotional capabilities we saw earlier on in the context of Covid-19. In that scenario, fundraisers needed to approach the consequences the crisis had on themselves and others by expressing emotions. They did so by empathizing with others, helping them manage their well-being and emotions. However, we argue that the advent of AI for storytelling brings the need for an other-oriented type of emotional resilience that helps address the issue of a decrease in donors' trust not only toward AI but more specifically toward any form of AI storytelling used for fundraising purposes. The issue of emotion regulation resilience is at the core of developing RAI capabilities. Developing and promoting feelings of trust, loyalty, and, by association, ongoing generosity between fundraisers and donors is an AI-driven challenge. Even though AI storytelling poses fundamental challenges to the emotion regulation pillar, there are also ways of pre-empting a long-term negative impact on levels of loyalty and trust.

#### ***4.1 RAI cognitive and behavioral resilience***

The number of guidelines and frameworks for ethical AI that are being published provide principles-based guidance to inform RAI policies. We believe these can constitute a starting point for nonprofits to build awareness of how they can develop AI storytelling responsibly. Ultimately, such guidelines also help nonprofits keep the trust-bond relationship with their donors and supporters intact. For example, the High-Level Expert Group on Artificial Intelligence – an independent expert group set up by the European Commission in June 2018 as part of its AI strategy – published the European Commission's Guidelines for Trustworthy AI (European Commission, 2019).<sup>7</sup> In this context, Trustworthy AI has three main components: lawful, ethical, and robust. That is, Trustworthy AI ensures compliance with applicable laws and regulations, adherence to ethical principles and values, and is technically robust. However, it is in the area of ethical communications, more specifically, that the guidelines offer practical ways for thinking about how to relate to stakeholders in responsible ways and facilitate the involvement of end-users. Specific guidance for the charitable sector is only beginning to emerge. The Charity Excellence Framework recently shared the Charity AI Governance and Ethics Framework,<sup>8</sup> a living document that is starting to unpack some sector-specific concerns and risk management approaches. Additionally, Fundraising.AI is a member-driven collaborative initiative supporting Responsible AI adoption in fundraising. Fundraising.AI has published "A Framework toward Responsible AI for Fundraising,"<sup>9</sup> which highlights key principles and considerations, from data ethics and inclusivity to legal compliance and sustainability.

Understanding, adhering to, and applying Trustworthy AI recommendations would mean making explicit that the story is AI-generated rather than human-authored. Even though this form of disclaimer cannot guarantee the loyalty of donors and supporters toward AI-generated stories, it nonetheless sets an example of good practice among fundraisers. More specifically, communicating around potential and perceived risks, such as bias to the intended audience, can increase

trust. Establishing a means for stakeholders to participate in developing and planning AI stories is another way of ensuring that trust levels are maintained. When AI stories interact directly with humans by encouraging them to develop attachment and empathy toward the nonprofit's cause, end-users must be made aware that such AI-powered forms of social interaction are simulated and not based on personal experiences of understanding and feeling.

Additionally, formalizing processes to ensure that values are integrated into innovation pipelines may be advisable. The BBC is one organization that is conscious of developing its Responsible AI strategy. The Machine Learning Engine Principles (MLEP) framework, published by BBC R&D, for example, is a self-audit tool that takes a principles-to-practice approach.<sup>10</sup> The MLEP provides a roadmap for technical projects that starts with the organization's values and uses these to inform project design and development, together with practical guidance and a checklist. By developing formalized processes guided by organizational values, it is possible to anticipate potential risks and avoid unintended, negative consequences in advance.

#### **4.2 RAI emotion regulation resilience: from distrust to trust**

As mentioned earlier, emotion-related capabilities refer to the presence of mental fortitude that helps individuals cope with adverse situations and is expressed in the form of individual and/or collective optimism and/or hope. This also includes having and creating opportunities to communicate and discuss emotions; such situations are likely to enhance emotion regulation capabilities (Williams et al., 2017). Consequently, the question this capability needs to address is how fundraisers can help their donors cope with what is potentially a widespread lack of trust in AI technologies and AI storytelling.

In relation specifically to foundation models underpinning generative AI, the Ada Lovelace evidence review, "What does the public think about AI" (2023), emphasized that inaccuracy was a key concern for the public. This is reflected in the 9% drop in share prices experienced by Google's parent company Jigsaw after the demo of their chatbot Bard, which included a factual error about the discovery of exoplanets.<sup>11</sup> Just as narratives of hype surrounding AI can *inflate* the public's perceptions of these systems' capabilities, demonstrating their fallibilities can have an overwhelming negative impact on an organization's reputation. However, ethical risk scanning and impact assessments can be used to anticipate and mitigate such risks. The "Public Perceptions of Foundation Models" (2023)<sup>12</sup> report, commissioned by the Centre for Data Ethics and Innovation, examined public perceptions of foundation models and found that public perceptions depend on the application area of AI. While advancing healthcare research is a use-case participants are more comfortable with, assisting doctors' decision-making processes was far less favorably perceived and deemed riskier. Assessing and communicating what benefits the introduction of AI contributes, and conducting audience research or developing citizen juries can help to ensure innovations align with their values.

Researchers have examined how the use of chatbots impacts charitable giving behaviors, considering how it affects moral behavior and donation amounts (Park et al., 2023; Zhou et al., 2022). Park et al. (2023) tested how different aspects of chatbot design impacted individuals' willingness to donate to a fundraising project. They found that their willingness to donate significantly dropped when a chatbot disclosed its identity (i.e., as non-human) and expressed affective empathy. As such, decisions about the deployment and design of AI systems can have notable implications for fundraising. The "How do people feel about AI" survey conducted by the Ada Lovelace Institute and Alan Turing Institute<sup>13</sup> emphasizes that responsible handling of private data is an enduring concern. While assisting the discovery of new and personally

relevant content is seen as a central benefit offered by AI, politically and consumer-targeted online advertising raises serious concerns about invasions of privacy. This tension between personalization and privacy must be delicately navigated and transparently communicated. Consequently, personalization and tailoring campaigns to individuals should, therefore, be carefully considered. Audience preferences will need to be researched and navigated so that users can maintain agency over how personal information is incorporated into their experiences. Public distrust is aroused when AI is perceived to manipulate or influence, especially in relation to emotions. However, Aoki (2020) found that communicating the purpose and highlighting the benefit to citizens can enhance public trust in systems over time. Employing transparent design approaches and supporting individual agencies wherever possible will be necessary for fundraisers in maintaining the trust of their audiences.

Therefore, we believe that as individuals' behavioral responses toward AI become more reactive to some of the potential offered by the technology, their trust in well-known and widely used fundraising tools, such as storytelling, diminishes. More specifically, such reactivity toward AI can even lead to a certain apprehension and distrust not only toward "machine-made" stories but also toward the organization behind them. For this distrust to set in and have long-term consequences, e.g., a decrease in engagement and donations, fundraisers need to pre-empt and thus build in advance what we define as "behavioral-cum-emotional" capabilities. This means to create positive associations and state the benefits, as well as the challenges, of AI to the organization and its cause in their engagements with donors and stakeholders. This form of relationship building includes a balanced understanding of AI and continues to adhere to the trust and loyalty on which the fundraiser-donor relationship is built. However, to get to this step, fundraisers must be aware of AI and be able to build cognitive capabilities, that is, understand the benefits and threats AI poses. The next step is to develop behavioral capabilities, building practices, and ways of working that implement, for example, some of the guidelines above, such as making explicit to donors that some of the stories used in fundraising appeals include AI-generated content. This type of practice can lead to a reduction in mistrust, even if not toward the specific AI content, at least toward the fundraisers who are endorsing such appeals on behalf of their organizations. While the issue of public distrust in AI storytelling authorship will remain or will not be entirely solved, fundraisers may find ways of mitigating a potential lack of trust. A practical step would be to avoid AI storytelling in individual solicitations altogether. A more nuanced solution may be to inform donors in advance that they have been interacting with content generated by a "robot." Even though initially this may feel like an irretrievable form of betrayal to some donors, it is ultimately fundraisers who can mitigate such damage.

## Notes

- 1 <https://www.microsoft.com/en-us/bing/do-more-with-ai/what-is-bing-chat-and-how-can-you-use-it?form=MA13KP>
- 2 [https://www.reddit.com/r/bing/comments/110eagl/the\\_customer\\_service\\_of\\_the\\_new\\_bing\\_chat\\_is/](https://www.reddit.com/r/bing/comments/110eagl/the_customer_service_of_the_new_bing_chat_is/)
- 3 <https://www.nytimes.com/2023/02/15/technology/microsoft-bing-chatbot-problems.html>
- 4 [https://www.reddit.com/r/bing/comments/110eagl/the\\_customer\\_service\\_of\\_the\\_new\\_bing\\_chat\\_is/](https://www.reddit.com/r/bing/comments/110eagl/the_customer_service_of_the_new_bing_chat_is/)
- 5 <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- 6 <https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/>
- 7 <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- 8 [https://www.charityexcellence.co.uk/Home/BlogDetail?Link=AI\\_Ethics\\_Governance\\_Framework](https://www.charityexcellence.co.uk/Home/BlogDetail?Link=AI_Ethics_Governance_Framework)
- 9 <https://fundraising.ai/framework/>
- 10 [https://downloads.bbc.co.uk/rd/pubs/MLEP\\_Doc\\_2.1.pdf](https://downloads.bbc.co.uk/rd/pubs/MLEP_Doc_2.1.pdf)

- 11 <https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares#:~:text=Kitwood%2FGetty%20Images-,Shares%20for%20Google's%20parent%20company%2C%20Alphabet%2C%20dropped%209%25%20Wednesday,Bard%2C%20gave%20an%20incorrect%20answer.&text=Google's%20parent%20company%2C%20Alphabet%2C%20lost,error%20in%20its%20first%20demo>
- 12 <https://www.gov.uk/government/publications/public-perceptions-towards-the-use-of-foundation-models-in-the-public-sector>
- 13 <https://www.adalovelaceinstitute.org/report/public-attitudes-ai/>

## References

- Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3, 461–463.
- Alborough, L. (2017). Lost in translation: A sociological study of the role of fundraisers in mediating gift giving in non-profit organisations. *International Journal of Nonprofit Voluntary Sector Marketing*, 22e, 1062.
- Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly*, 37(4), 101490.
- Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963.
- Blackbaud (2022). The status of UK fundraising report. file:///Users/martaherrero/Downloads/the-status-of-uk-fundraising-2022-benchmark-report.pdf
- Brown, T., Mann, B., & Ryder, N. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1–75 (NeurIPS 2020).
- Cantin, R., & Michel, P. (2003). Towards a new technology future approach. *Futures*, 35(3), 189–201.
- Charities Aid Foundation (2023). UK Giving Report.
- Chu, H., & Liu, S. (2023). Can AI tell good stories? Narrative transportation and persuasion with ChatGPT.
- De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11, 1166120.
- Esser, P., Chiu, J., Atighehchian, P., Granskog, J., & Germanidis, A. (2023). Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7346–7356).
- European Commission (2019). Ethics guidelines for trustworthy AI.
- Herrero, M., & Kraemer, S. (2022). Beyond survival mode. Organisational resilience capabilities in nonprofit arts and culture fundraising during the Covid-19 pandemic. *Nonprofit Management & Leadership*, 33(2), 279–295.
- Hickman, P. (2018). A flawed construct? Understanding and unpicking the concept of resilience in the context of economic hardship. *Social Policy & Society*, 17(3), 409–424.
- Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, 380(6651), 1222–1223.
- Lengnick-Hall, C. A., & Beck, T. E. (2005). Adaptive fit versus robust transformation: How organizations respond to environmental change. *Journal of Management*, 31(5), 738–757.
- Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022, June). News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 97–106).
- Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, 20, e10148.
- NCVO (2022). UK Civil Society Almanac. Data, trends, insights.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Park, G., Yim, M. C., Chung, J. & Lee, S. (2023). Effect of AI chatbot empathy and identity disclosure on willingness to donate: The mediation of humanness and social presence. *Behaviour & Information Technology*, 42(12), 1998–2010.
- Sargeant, A., & Edworthy, K. (2022). What makes fundraisers tick? A study of identity, motivation and well-being. Report by Revolutionise International.

- Singer, U., & Polyak, A. (2022). Make-a-video: Text-to-video generation without text-video data. arXiv pre-print arXiv:2209.14792.
- Skills Platform (2021). Charity digital skills report.
- Villegas, R., Babaeizadeh, M., Kindermans, P. J., Moraldo, H., Zhang, H., Saffar, M. T., Castro S., Kunze, J., & Erhan, D. (2022, September). Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations* (pp. 1–17).
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., & Biles, C., (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229).
- Williams, T. A., Gruber, D. A., Sutcliffe, K. M., Shepherd, D. A., & Zhao, E. Y. (2017). Organizational response to adversity: Fusing crisis management and resilience research streams. *Academy of Management Annals*, 11(2), 733–769.
- Zhou, Y., Fei, Z., He, Y., & Yang, Z. (2022). How human–chatbot interaction impair charitable giving: The role of moral judgment. *Journal of Business Ethics*, 178(3), 849–865.

## **PART II**

# **Philanthropies' regional AI adoption, readiness, and applications**





# Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# 10

## ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, AND DATA SCIENCE PHILANTHROPY

### Case studies of a purposive classification of philanthropic missions

*Patricia Snell Herzog*

#### 1 Introduction

This chapter presents case studies of how artificial intelligence (AI), machine learning (ML), and data science technology (DST) are being integrated into philanthropy. As Aoun (2017: xxi) stated, “Given the pace of technology’s advance, we can predict that computers, robots, and artificial intelligence will be even more intricately intertwined into the fabric of our personal and professional lives.” These technologies have been and will continue to be intertwined with our lives and, as a result, with philanthropy and other efforts to improve the welfare of others. With great power comes even greater scrutiny, and the world has turned its attention to the moral and ethical dilemmas that these forms of technology present. Whether it is a focus on the responsibility to develop trustworthy AI (e.g., Cheng et al., 2021), moral machinery (Roff, 2019), data science for social good (e.g., Lapucci & Cattuto, 2021), or tech philanthropy (Henriksen & Richey, 2022), technology is transforming philanthropic and nonprofit research and practice (McCully, 2019).

Despite this rapid transformation, or perhaps because of its accelerated pace, research on technological innovation in philanthropy remains sparse. In fact, a search for the keyword artificial intelligence in the *Nonprofit and Voluntary Sector Quarterly*, one of the most respected journals on philanthropic and nonprofit studies, yielded only 13 articles published since 2010 (Benjamin et al., 2022; Brandtner, 2021; Cox et al., 2018; Enjolras, 2022; Fyall et al., 2018; Kang et al., 2022; LePere-Schloop, 2022; LePere-Schloop et al., 2022; Li et al., 2022; Ma, 2021; Nwakpuda, 2020; Schubert et al., 2022; Williamson et al., 2021). However, 12 of these focus on the use of AI and machine learning tools for research, transcription, or analysis, rather than the use of these tools in philanthropy. The only remaining article focuses solely on how machine learning in social media platforms is giving some organizations more online attention than others.

To facilitate further research on this important topic, this chapter offers a series of case studies to illustrate how philanthropic organizations are integrating technology into their missions. Before turning to the case study analysis, the following section defines how this chapter views the conceptual overlap and distinction between the three technology terms: AI, ML, and DST.

## 1.1 Definitions

Artificial intelligence is defined as the simulation of human intelligence in machines. This includes computer systems that are able to reason, discover meaning, generalize, or learn from experience. AI is typically concerned with understanding the mechanisms underlying intelligent behavior and computer implementation. Machine learning is defined as algorithms built on training data to make predictions and support human decision-making. ML models improve over time through learning algorithms. Data science technology is designed to augment and automate data science processes. This includes big data, statistics, and data analytics. DST draws from techniques in computer science, statistics, information science, and social science. Using these definitions, several case studies are presented that integrate these technologies into philanthropy. Figure 10.1 visualizes these definitions and their conceptual relationships.

## 2 Case study methods

The methodology used in this chapter is inspired by 15 case studies published in the *Nonprofit and Voluntary Sector Quarterly* or *Voluntas: International Journal of Voluntary and Nonprofit Organizations* (specifically: Chenhall et al., 2016; Dodge & Ospina, 2016; Evans & Clarke, 2010; Grabowski et al., 2015; Hua et al., 2016; Huang, 2022; Hudon & Meyer, 2016; Jäger & Kreutzer, 2011; McAllister & Makkai, 2021; Narvaiza et al., 2017; Noh, 2019; Sheng, 2019; Vanleene et al.,

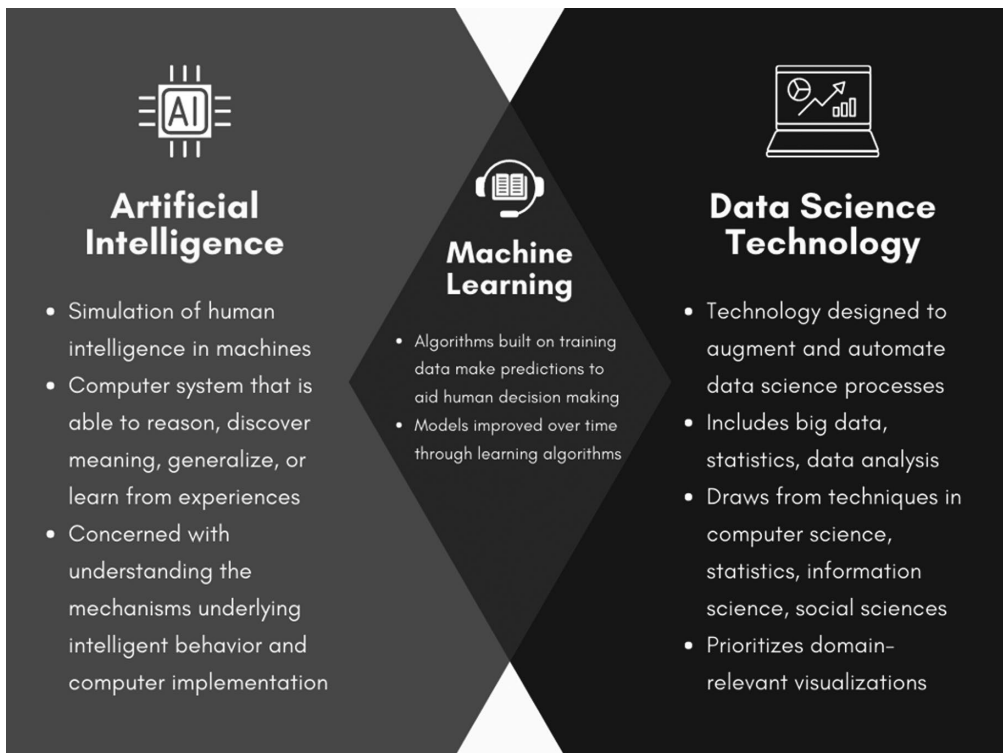


Figure 10.1 Relationships among AI, machine learning, and data science technology definitions.

2018; Vu et al., 2017; Wickes et al., 2017). From these examples, it is apparent that case studies take a variety of different forms and formats. Some are comprehensive analyses of a single organization, while others are in-depth comparisons between a small number of organizations. Still others draw illustrative examples from a systematically searched set of cases. The latter approach is used in the current study, and the next section describes the inclusion criteria.

## **2.1 Inclusion criteria**

The following inclusion criteria were used to form a relatively systematic sample of organizations relevant to artificial intelligence, machine learning, and data science technology. First, a set of sources were selected due to their attention to these technology issues and availability as either open access resources or with access readily provided. The five sources were: GuideStar (2023) nonprofit database, the *Chronicle of Philanthropy* news article database, Open Philanthropy’s grantmaking database, Rockefeller Foundation’s grantmaking database, and Google AI’s competition winners. From these sources, a combined total of 349 organizations were identified as having relevance. Table 10.1 lists the number of organizations found within each source.

Second, the scoping procedures used to identify relevance were as follows. These keywords were searched using Boolean logic: artificial intelligence, machine learning, data technology, and analytics. The Open Philanthropy grants database contained more than 800 grant awards across 13 focus areas, with the most obvious relevance being potential risks from advanced artificial intelligence. Upon further investigation, several other areas were also relevant. Namely, this grantmaker also awarded AI, ML, or DST grants within the Catastrophic Risks applications, Biosecurity and Preparedness and Global Catastrophic Risks, and Land Use Reform applications. In total, more than \$167 million in grant awards were deemed relevant after being analyzed for relevance. This grant database included 40 relevant philanthropic and nonprofit organizations.

The Rockefeller Foundation also has a vested interest in this area, and its open-source grants database of more than 600 awards over the past decade categorized grants within ten focus areas. The most prevalent is Data and Technology. In total, this foundation has awarded more than \$56 million in relevant grants across additional focus areas such as: Food, Health, Climate, Global Resilience, and more. This grant database included 39 philanthropic and nonprofit organizations engaged in activities related to artificial intelligence, machine learning, or data science.

A third source of organizations was generated from scoping articles published in the *Chronicle of Philanthropy*. This open-source database includes more than 15,000 articles that were searched for the same keyword terms described above. The articles contained information about who donated, what amounts, for which purposes, and to what organizations. Cumulatively, more than \$2 billion was donated to relevant philanthropic purposes to 54 unique organizations.

*Table 10.1 Organizations by source*

<i>Source</i>	<i>Organizations</i>
Chronicle of Philanthropy	54
Google AI	20
GuideStar	196
Open Philanthropy	40
Rockefeller Foundation	39
<b>Grand Total</b>	<b>349</b>

A unique dataset was derived from Google AI for Social Good, which hosts a social impact challenge initiative that distributed \$25 million. The 20 awardees were scraped, and their for-good missions were identified from publicly posted grant award materials.

The fifth dataset was scraped from the GuideStar nonprofit database hosted by Candid. With data scraped from IRS 990 forms submitted by tax-exempt entities, this database is also supplemented with data entered by representatives of the organizations and coded by Candid staff. The database contains nearly 3 million organizations that were scoped using the same Boolean logic described above. This resulted in an additional 196 unique organizations coded as relevant. Together, the five data sources yielded 349 organizations relevant to AI, ML, or DST.

## 2.2 Coding process

These organizations were then coded based on their missions and related programming. Organizations were categorized into three sets based on technology integration styles. The first set of organizations are *tech-centered* in their missions, meaning their mission is to directly support AI, ML, or DST activities. These organizations were typically identified by their use of one of the technology keywords directly in the organization's title or prominently in the mission statement. In reviewing their missions and programming, it became clear that the organizations were primarily designed to harness the power of artificial intelligence, machine learning, or data science technology to produce new knowledge. A total of nine organizations were categorized as *tech-centered*, with more than \$130 million in new grant awards.

Second, an additional set of organizations were coded as *tech-perpetuating*, meaning that their mission is to advance research and discovery of these technologies, primarily through universities or research institutes, or their goal is to help more people engage in using these technologies. Many of these organizations are committed to educating the public or specific groups, such as women and girls, minoritized racial or ethnic communities, or socioeconomically disadvantaged individuals, about how to utilize AI, ML, or DST. In addition, many of these organizations are associations or societies that facilitate the convening of technology communities, often through annual conferences and other online communications. A total of 188 organizations were categorized as *tech-perpetuating*, with more than \$1.6 billion in new grant awards invested in this area, along with countless additional revenues.

The third group of organizations was coded as *tech-implementing*, meaning that AI, ML, or DST is being used to deliver services or operational tasks, with the technology being applied to improve impact or capacity. These organizations focused on tech applications for cybersecurity, health, climate issues, land use, global disasters and pandemics, food, animals, and other areas. The titles of these organizations often did not use technology terms, but their mission statements and programming initiatives described applications of algorithms, AI for good, or social impact driven by data science. In total, 152 organizations were coded as *tech-implementing*, with nearly \$800 million in new grant awards invested alongside existing revenue streams and asset bases.

While these categories may overlap to some extent, particularly between *tech-centered* and *tech-perpetuating*, the organizations were differentiated based on their primary purpose. If their primary focus was to create AI, ML, or DST, they were coded as being focused on this objective, whereas if their primary focus was to generate research about, educate with, or foster greater access among communities in learning about or using these tools, they were coded as perpetuating.

### **3 Tech integration styles**

This classification method helps identify different types of organizations involved in AI, ML, and DST philanthropy. The case studies described below illustrate how the focus shifts across these organizational groups, and this can be useful for researchers and practitioners to better understand the role these organizations play in the field. It is also useful for the general public to better see the importance of philanthropy in these technologies. The first set of *tech-centered* organizations is perhaps the most salient in that it is likely what most people think of when they think of AI, ML, and DST philanthropy. What can be less obvious to the general public is that philanthropic organizations are also involved in these activities. Additionally, it is not well known, even among tech scientists and business leaders, that philanthropy is heavily involved in the other two primary purposes: perpetuation and applications. The following series of case studies can help make these activities more visible and within a useful classification scheme.

#### **3.1 Tech-centered**

Of the nine philanthropic organizations identified as *tech-centered*, this section presents four case studies: the Allen Institute for Artificial Intelligence, the AI for Good Foundation, Humans for AI, and OpenAI. Most of the information presented on these organizations was gathered from the organization's website, as cited in the reference section, along with GuideStar profiles.

The Allen Institute for AI (AI2, 2023) describes its mission as “to contribute to humanity through high-impact AI research and engineering.” Founded in 2014, this nonprofit is a research institute named after Paul G. Allen, a co-founder of Microsoft and an avid tech philanthropist. They state that their primary goal is to produce research and tools that benefit society. Computer scientists and researchers funded by the organization work on Natural Language Processing (NLP) teams such as AllenNLP, AI2 Tango as a Python library, Aristo for systematic reasoning, and Mosaic for machine common sense. They are also developing an open and generative language model called the Open Language Model (OLMo), which was released in 2024. Furthermore, they provide data philanthropy through several open access datasets, including Digital Socrates, Satlas Explorer for satellite imagery, BaRDA: a belief and reasoning datasets, Lila for mathematical reasoning, Macaw for question-answering, and Unified-IO for research visualization. The Chronicle of Philanthropy reported that Allen has given more than \$125 billion to AI2.

The AI for Good Foundation (2023) states that their purpose is to encourage research communities to develop AI systems in pursuit of social good. They describe their use of “economic thinking and technological innovation to solve big human challenges, transform institutions, and impact people's lives.” One of their projects is LifeForce, a platform that helps people in humanitarian crises request survival needs in real time such as medical assistance or shelter. In addition, the foundation has a Sustainable Development Goal (SDG) incubator, which facilitates startup ventures in developing climate and community responsibility. Beyond financial resources, the foundation also supports venture capitalists in understanding AI ethics and metrics to best measure environmental and social impact. Then, the projects are scaled to create long-term and sustainable infrastructure in support of the SDGs.

The third case study, Humans for AI (2023), describes their organization as believing that AI will transform human life by being as impactful as the Internet or mobile technologies and changing how all professions do their work. Their research lab projects include using NLP to identify hiring biases through keywords in resumes that successfully secure jobs, more accurately diagnosing skin diseases in people of different skin tones and detailing impactful website design elements.

The organization is also working on lasting tasks, such as developing a community of AI scientists and educating young people about the fruits of AI. However, this organization remains categorized in this central category because these perpetuating efforts appear to be focused on engaging in the core research lab projects that harness the power of AI for specific purposes.

The fourth and final case study highlighted in this section is one that has received attention recently: OpenAI. On its website, OpenAI (2023) states that its “mission is to ensure that artificial general intelligence (AGI) benefits all of humanity, primarily by attempting to build safe AGI and share the benefits with the world.” Founded as a nonprofit organization in 2016, the organization reported having more than \$21 million in assets as of 2020. As the creator of ChatGPT, the organization has transformed the public’s understanding and use of AI. As the organization has successfully scaled, it has transitioned from a nonprofit to also having a for-profit entity (Aspan, 2023; Broughel, 2023; Novet, 2023; Ortutay, 2023; Salmon, 2023; TheTechPencil, 2023).

This relatively unique structure among tech giants means that the organization operates with a cap on the amount of returns it receives, with the ideal of directing its attention on its mission for good rather than maximizing profits. However, this structure has become complicated, with recent activities including the board of directors replacing the chief executive officer, who was later reinstated after an employee uprising, and replacing the previous board members. This governance crisis is purportedly due to the relationship between the for-profit company and the nonprofit company, with the company now worth more than \$80 billion after the success of its launched products. Yet, the nonprofit was reported last year as generating less than \$45,000 in revenue, mostly from investment income. While the nonprofit’s charter states that the organization must remain committed to using AI for the benefit of all humanity, the for-profit LP states that its returns to investors are capped at 100 times the original investment amount (OpenAI Charter, 2023; OpenAI LP, 2023).

In sum, *tech-centered* philanthropic and nonprofit organizations are similar to for-profit entities in that their core is built around the innovative potential of artificial intelligence, machine learning, and data science technology. To the extent that these organizations engage in educational or community-building activities, these tend to be aligned with and peripheral to their primary engagement.

### 3.2 *Tech-perpetuating*

The second set of case studies is primarily engaged in *tech-perpetuating* activities, although their work also often supports research and development in AI, ML, and DST. This set is the largest, in terms of both the number of organizations (188 out of 349) and new grant awards (>\$1.6 billion). This monetary value underestimates the size and scope of these organizations, as it only totals new grant awards recorded in the scoped databases and does not calculate existing assets and revenues.

There are three subtypes of organizations within this category: (a) universities and research centers, (b) educational outreach programs for underrepresented groups or accessibility for all, and (c) associations or societies that help connect and convene the technology community. Examples of universities that have received substantial philanthropic contributions for AI, ML, and DST include the Massachusetts Institute of Technology at more than \$350 million, the University of Southern California with more than \$261 million, Indiana University Luddy Center for Artificial Intelligence with more than \$60 million, Georgetown University with more than \$55 million, and Northeastern University and Rochester Institute of Technology with more than \$50 million.

The first case study exemplifies the university subtype. Founded in 2020, the MIT Stephen A. Schwarzman College of Computing (2023) has a mission to

address the opportunities and challenges of the computing age – from hardware to software to algorithms to artificial intelligence – by transforming the capabilities of academic in three key areas: computing fields, computing across disciplines, and the social and ethical aspects of computing.

First and foremost, among these activities is the development of the field of computational sciences through the education of students by faculty. The founding philanthropist, Stephen Schwarzman, is Chairman, CEO, and co-founder of Blackstone, an investment firm with more than \$1 trillion in assets (Blackstone, 2023). He has also donated millions of dollars to the University of Oxford for a Humanities Institute for Ethics in AI. Many of the philanthropic dollars invested in universities are dedicated to capital campaigns to build new buildings with lab capabilities that can support the necessary computing power. Figure 10.2 shows a university building project with a sign that reads: “Philanthropy at Work.”

As a second subtype of *tech-perpetuating* philanthropy, additional organizations aim to make technology accessible to all and/or work specifically to engage underrepresented groups. For example, AI4All (2023) describes itself as a “nonprofit working to increase diversity and inclusion in artificial intelligence development, policy, and research” through its efforts to transform the “pipeline of AI practitioners and creating a more inclusive, human-centered discipline.” One of its programs is an undergraduate accelerator that mentors students in developing an AI project to showcase their work while also providing hands-on training for career readiness. The program culminates in a certificate of completion that helps young people get hired in AI jobs. Graduates



*Figure 10.2* University building construction sign says philanthropy at work.

*Source:* Author.



of this program move on to the Changemakers in AI program, which provides internships, workshops, speaker series, and collaborative projects to develop the AI leadership pipeline.

Additionally, Women in Data Science and Analytics Inc., also known as Women in Data (2023), aims to increase diversity in data and technology careers by supporting women, building a welcoming community, and inspiring change in industry leadership. They state:

At our core, we want to achieve gender equality and empower women and girls in the field of data. Unfortunately, gender parity remains far off around the world. Whether it's often carrying the burden of unpaid work (hello moms everywhere) or not being made aware or given the same opportunities as men, women are still falling behind.

Their programs focus on four values in the acronym CODE: Community, Opportunity, Diversity, and Education. One of their main initiatives is a datathon, where participants solve problems using data science tools. They also provide career services, mentorship, life coaching, and virtual study groups to enhance data skills.

Another of these subtypes is the case of the Solidarity Research Center (2023), which describes itself as a “nonprofit organization that builds solidarity economy ecosystems using data science, story-based strategy, and action research” by working at the “intersection of racial justice and solidarity economies.” The organization originally began as the research department of the Industrial Workers of the World, which organized on behalf of incarcerated workers. They then expanded their scope to include farmworkers and food chain issues. Today, they are a fiscal sponsor of a tech learning space for people of color called Color <Coded> (2023), which is working on organizing activities in Los Angeles around the theme of #TechIsNotNeutral.

A third example of this subtype is DataEthics4All (2023), which describes itself as a public-benefit corporation dedicated to breaking down “barriers of entry in Tech for girls, People of Color, the economically disadvantaged” by fostering AI ethics champions. The organization hosts a youth council dedicated to AI ethics, specifically protecting children’s and teens’ data with tech ethics. Middle and high school kids can start an extracurricular club on topics like Pi and AI (pi here referring to the math symbol). They also host a STEAM (science, technology, engineering, art, and math) in AI college-preparatory program with experiential learning and career mentoring.

The final subtype of *tech-perpetuating* organizations consists of associations and societies. For example, the Technology Association of Grantmakers (TAG) (2023) is an association of grantmaking philanthropic foundations that develops artificial intelligence resources for philanthropy, such as a framework for the responsible use of AI in grantmaking activities. They define the use of AI for mission fulfillment as “tools provisioned by the organization for discovery, insight, fundraising, impact assessment, predictive analysis” that can support grantmaking decisions, help evaluate program areas, search grantee reports, and assess issue areas of need.

In sum, *tech-perpetuating* philanthropic organizations are primarily focused on educating about, promoting broader access to, or facilitating community engagement around AI, ML, or DST.

### 3.3 *Tech-implementing*

The third set of case studies are *tech-implementing*, using applications of artificial intelligence, machine learning, and data science technology to alleviate social problems. For example, the Green AI Foundation (2023) brings together volunteer engineers and professionals to tackle

environmental and sustainability challenges. One project is the Florida Waterway Health Forecast, which uses multiple datasets to create predictive models of marine biological health and then works with environmental groups and water management districts to address issues. A second project uses satellite imagery to track mangrove tree ecosystems, and a third uses predictive modeling by volunteer data scientists to forecast red tide events when harmful algae damage marine life and then works with stakeholders to mitigate the effects.

A second case of this type is Community AI (2023), whose mission is to be a “youth-driven nonprofit organization aimed to help the community and environment by building projects using state of the art technologies and the power of AI.” One of its major initiatives is the AI Summer Camp, attended by people from more than 50 countries worldwide. High and middle school students work on AI projects aimed at reducing community and environmental problems. They also host career fairs and speaker series featuring data and AI scientists.

In a third example, the Wadhvani Institute for AI Foundation (2023) describes itself as an “independent nonprofit institute developing AI-based solutions for underserved communities in developing countries.” On the health issues front, they are working on tuberculosis prevention by using AI to automatically interpret test results to determine whether a patient has drug resistance and, if so, of what strain. Another example is an AI-based app that screens cough sounds to identify people at risk for tuberculosis infections. Another uses AI in a smartphone app to alert cotton farmers to signs of insect infestation, helping them manage pests before crops are severely damaged.

Fourth, a case study called Medical Automation Org Inc (2023) “advances worldwide utilization of automation, robotics and artificial intelligence to improve the quality, efficiency and relevance of medical care.” Specifically, they host an annual conference on how to integrate automated solutions into healthcare. These types of organizations are part of a rapidly growing trend to implement AI applications in healthcare delivery, disease diagnosis, and treatment (Davenport & Kalakota, 2019). These applications can also help patients adhere to a treatment plan, a key issue that many healthcare providers struggle to influence. AI-powered apps can remind and encourage patients to follow their treatment while they are in their natural environment. Philanthropic organizations are supporting these applications.

A fifth case is the Global Fishing Watch, Inc. (2023), which uses AI and satellite data to monitor and analyze the oceans through the Open Ocean Project, which creates a digital ocean through an online map of all industrial human activity. Color-coded maps show major trade routes across ocean waterways, identify vessels, plot offshore infrastructure, examine supply chains, alert on signals that have been disabled or otherwise stopped broadcasting their location, and detect the impact of fishing harvests to support better ocean governance.

#### **4 For-good framework**

Cumulatively, the 15 different case studies described here span three different roles that philanthropy plays in technology for good activities. Figure 10.3 provides a visual conceptualization of this framework. It begins on the left side with *tech-centered* philanthropy. This centering work then feeds into the *tech-perpetuating* philanthropy in the second column through universities, associations, and access groups. These, in turn, feed into the third column of *tech-implementing* philanthropy, with applications for the environment, water, healthcare, disease, community, and youth, for example. Lastly, these activities culminate in improving society by alleviating social problems.

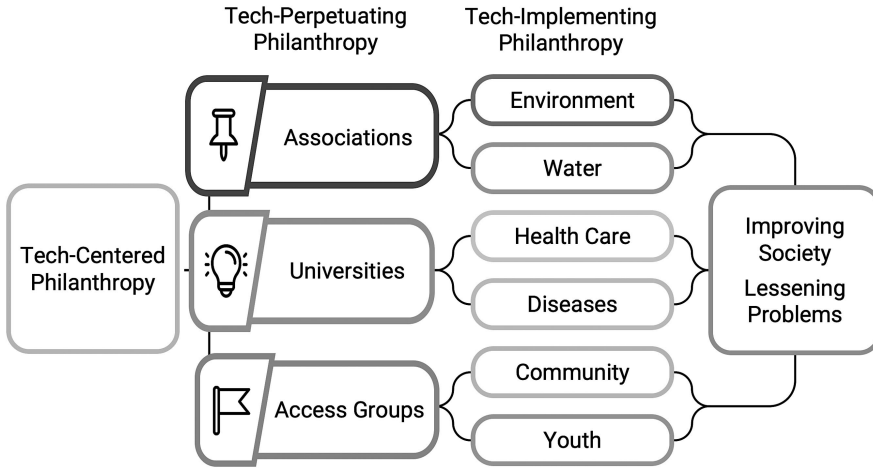


Figure 10.3 Framework for the three roles of philanthropy in technology for good activities.

Table 10.2 Tech for-good philanthropy case studies

---

*Tech-Centering Philanthropy Case Studies*

---

Allen Institute for AI  
 AI for Good Foundation  
 Humans for AI  
 OpenAI, Inc.

*Tech-Perpetuating Philanthropy Case Studies*

MIT Stephen A. Schwarzman College of Computing  
 AI4All  
 Women in Data  
 Solidarity Research Center: Color <Coded>  
 DataEthics4All  
 Technology Association of Grantmakers (TAG)

*Tech-Implementing Philanthropy Case Studies*

Green AI Foundation  
 Community AI  
 Wadhvani Institute for AI Foundation  
 Medical Automation Org Inc  
 Global Fishing Watch, Inc.

---

## 5 Conclusion

This chapter highlights the ways in which philanthropy contributes to the advancement of technology. The case studies illustrate four examples of *tech-centering* philanthropy, six case studies of *tech-perpetuating* philanthropy, and five case studies of *tech-implementing* philanthropy. Table 10.2 lists each of the featured case studies within their three classifications by activity type.

Recognizing the existence of these three distinct types of technology-advancing philanthropic activities can help researchers, practitioners, and the general public deepen their understanding of how efforts for the good play a role in facilitating the creation, dissemination, and application of artificial intelligence, machine learning, and data science technologies. It is particularly fruitful to consider the full range of tech-perpetuation activities, not least because these efforts support broader access to the technology pipeline and education on the skills and experiences needed for the future of work. It is therefore recommended that philanthropic organizations adopt this taxonomy as a way to communicate their value to relevant stakeholders and, in so doing, contribute to education about the importance of philanthropy.

### Acknowledgments

The author would like to thank former students Harshal R. Naik and Haseeb A. Khan, both in the master's in data science program at the Indiana University Luddy School of Informatics, Computing, and Engineering, Indianapolis, for their contributions to earlier stages of this project, as well as Arisa Miyakozawa as a master's student in the IU Lilly Family School of Philanthropy for co-reading the case studies published in NVSQ. Also, thanks to the IU Indianapolis Library for providing access to the Candid GuideStar nonprofit database and the *Chronicle of Philanthropy* article database, as well as to the Rockefeller Foundation, Open Philanthropy, and Google AI for supporting open access to their grantmaking databases.

### References

- AI for Good Foundation (2023). Retrieved December 15, 2023, from <https://ai4good.org/>
- AI4ALL (2023). AI4ALL. Retrieved December 15, 2023, from <https://ai-4-all.org/>
- Allen Institute for AI (2023). Retrieved December 15, 2023, from <https://allenai.org/>
- Aoun, J. E. (2017). *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. The MIT Press.
- Artificial Intelligence (AI) Resources for Philanthropy (2023). Technology Association of Grantmakers. Retrieved December 15, 2023, from <https://www.tagtech.org/page/AI>
- Aspan, M. (2023, November 21). *The OpenAI Meltdown Shows That When Nonprofits and for-Profits Clash, the One with the Money Usually Wins*. Fortune. <https://fortune.com/2023/11/21/openai-meltdown-microsoft-sam-altman-nonprofit-for-profit/>
- Benjamin, L. M., Ebrahim, A., & Gugerty, M. K. (2022). Nonprofit Organizations and the Evaluation of Social Impact: A Research Program to Advance Theory and Practice. *Nonprofit and Voluntary Sector Quarterly*, 08997640221123590. <https://doi.org/10.1177/08997640221123590>
- Blackstone (2023, November 28). Blackstone. <https://www.blackstone.com/>
- Brandtner, C. (2021). Decoupling Under Scrutiny: Consistency of Managerial Talk and Action in the Age of Nonprofit Accountability. *Nonprofit and Voluntary Sector Quarterly*, 50(5), 1053–1078. <https://doi.org/10.1177/0899764021995240>
- Broughel, J. (2023, December 9). *OpenAI Is Now Unambiguously Profit-Driven, and That's A Good Thing*. Forbes. <https://www.forbes.com/sites/jamesbroughel/2023/12/09/openai-is-now-unambiguously-profit-driven-and-thats-a-good-thing/>
- Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *Journal of Artificial Intelligence Research*, 71, 1137–1181. <https://doi.org/10.1613/jair.1.12814>
- Chenhall, R. H., Hall, M., & Smith, D. (2016). Managing Identity Conflicts in Organizations: A Case Study of One Welfare Nonprofit Organization. *Nonprofit and Voluntary Sector Quarterly*, 45(4), 669–687. <https://doi.org/10.1177/0899764015597785>
- Color <Coded> (2019, October 17). Color Coded. <https://colorcoded.la>
- Community AI (2023). Retrieved December 15, 2023, from <https://www.thecommunityai.org/>
- Cox, J., Oh, E. Y., Simmons, B., Graham, G., Greenhill, A., Lintott, C., Masters, K., & Woodcock, J. (2018). Doing Good Online: The Changing Relationships between Motivations, Activity, and Retention among Online Volunteers. *Nonprofit and Voluntary Sector Quarterly*, 47(5), 1031–1056. <https://doi.org/10.1177/0899764018783066>

- DataEthics4All (2023). DataEthics4All. Retrieved December 15, 2023, from <https://dataethics4all.org/>
- Davenport, T., & Kalakota, R. (2019). The Potential for Artificial Intelligence in Healthcare. *Future Health-care Journal*, 6(2), 94–98. <https://doi.org/10.7861/futurehosp.6-2-94>
- Dodge, J., & Ospina, S. M. (2016). Nonprofits as “Schools of Democracy”: A Comparative Case Study of Two Environmental Organizations. *Nonprofit and Voluntary Sector Quarterly*, 45(3), 478–499. <https://doi.org/10.1177/0899764015584063>
- Enjolras, B. (2022). Determinants of Voluntary Organizations’ Attention on Facebook: The Case of Norwegian Voluntary Organizations. *Nonprofit and Voluntary Sector Quarterly*, 08997640221129551. <https://doi.org/10.1177/08997640221129551>
- Evans, S. H., & Clarke, P. (2010). Training Volunteers to Run Information Technologies: A Case Study of Effectiveness at Community Food Pantries. *Nonprofit and Voluntary Sector Quarterly*, 39(3), 524–535. <https://doi.org/10.1177/0899764009333053>
- Fyall, R., Moore, M. K., & Gugerty, M. K. (2018). Beyond NTEE Codes: Opportunities to Understand Nonprofit Activity Through Mission Statement Content Coding. *Nonprofit and Voluntary Sector Quarterly*, 47(4), 677–701. <https://doi.org/10.1177/0899764018768019>
- Global Fishing Watch, Inc (2023). Revolutionizing Ocean Monitoring and Analysis. Retrieved December 15, 2023, from: <https://globalfishingwatch.org/>
- Grabowski, L., Neher, C., Crim, T., & Mathiassen, L. (2015). Competing Values Framework Application to Organizational Effectiveness in Voluntary Organizations: A Case Study. *Nonprofit and Voluntary Sector Quarterly*, 44(5), 908–923. <https://doi.org/10.1177/0899764014546488>
- Green AI Foundation (2023). The Green AI Foundation. Retrieved December 15, 2023, from <https://tgai.org/>
- GuideStar: Nonprofit Data for Donors, Grantmakers, and Businesses (2023). Candid. Retrieved December 15, 2023, from <https://www.guidestar.org/>
- Henriksen, S. E., & Richey, L. A. (2022). Google’s Tech Philanthropy: Capitalism and Humanitarianism in the Digital Age. *Public Anthropologist*, 4(1), 21–50. <https://doi.org/10.1163/25891715-bja10030>
- Hua, R., Hou, Y., & Deng, G. (2016). Instrumental Civil Rights and Institutionalized Participation in China: A Case Study of Protest in Wukan Village. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 27(5), 2131–2149. <https://doi.org/10.1007/s11266-015-9616-9>
- Huang, S. (2022). NGO as Sympathy Vendor or Public Advocate? A Case Study of NGOs’ Participation in Internet Fundraising Campaigns in China. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 33(5), 1064–1076. <https://doi.org/10.1007/s11266-022-00463-w>
- Hudon, M., & Meyer, C. (2016). A Case Study of Microfinance and Community Development Banks in Brazil: Private or Common Goods? *Nonprofit and Voluntary Sector Quarterly*, 45(4\_suppl), 116S–133S. <https://doi.org/10.1177/0899764016643609>
- Humans for AI (2023). Retrieved December 15, 2023, from <https://humansforai.com/>
- Jäger, U. P., & Kreutzer, K. (2011). Strategy’s Negotiability, Reasonability, and Comprehensibility: A Case Study of How Central Strategists Legitimize and Realize Strategies Without Formal Authority. *Nonprofit and Voluntary Sector Quarterly*, 40(6), 1020–1047. <https://doi.org/10.1177/0899764010378703>
- Kang, C. H., Baek, Y. M., & Kim, E. H.-J. (2022). Half a Century of NVSQ: Thematic Stability across Years and Editors. *Nonprofit and Voluntary Sector Quarterly*, 51(3), 658–679. <https://doi.org/10.1177/08997640211017676>
- Lapucci, M., & Cattuto, C. (Eds.) (2021). *Data Science for Social Good: Philanthropy and Social Impact in a Complex World*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-78985-5>
- LePere-Schloop, M. (2022). Nonprofit Role Classification Using Mission Descriptions and Supervised Machine Learning. *Nonprofit and Voluntary Sector Quarterly*, 51(5), 1207–1222. <https://doi.org/10.1177/08997640211057393>
- LePere-Schloop, M., Appe, S., Adjei-Bamfo, P., Zook, S., & Bawole, J. N. (2022). Mapping Civil Society in the Digital Age: Critical Reflections from a Project Based in the Global South. *Nonprofit and Voluntary Sector Quarterly*, 51(3), 587–605. <https://doi.org/10.1177/08997640211057401>
- Li, Z. C., Ji, Y. G., Tao, W., & Chen, Z. F. (2022). Engaging Your Feelings: Emotion Contagion and Public Engagement on Nonprofit Organizations’ Facebook Sites. *Nonprofit and Voluntary Sector Quarterly*, 51(6), 1281–1303. <https://doi.org/10.1177/08997640211057398>
- Ma, J. (2021). Automated Coding Using Machine Learning and Remapping the U.S. Nonprofit Sector: A Guide and Benchmark. *Nonprofit and Voluntary Sector Quarterly*, 50(3), 662–687. <https://doi.org/10.1177/0899764020968153>
- McAllister, I., & Makkai, T. (2021). Populism and Charity Donations: An Australian Case Study. *Nonprofit and Voluntary Sector Quarterly*, 50(5), 939–958. <https://doi.org/10.1177/0899764021991676>

- McCully, G. (2019). Research and Practice in Nonprofits and Philanthropy: An Overview. *Journal of Non-profit Education and Leadership*, 9(3), 217–222.
- Medical Automation Org Inc (2023). Retrieved December 15, 2023, from <https://www.medicalautomation.org/>
- MIT Schwarzman College of Computing (2023). MIT Schwarzman College of Computing. Retrieved December 15, 2023, from <https://computing.mit.edu/>
- Narvaiza, L., Aragon-Amonarriz, C., Iturriz-Landart, C., Bayle-Cordier, J., & Stervinou, S. (2017). Cooperative Dynamics during the Financial Crisis: Evidence from Basque and Breton Case Studies. *Nonprofit and Voluntary Sector Quarterly*, 46(3), 505–524. <https://doi.org/10.1177/0899764016661775>
- Noh, J.-E. (2019). Human Rights-Based Child Sponsorship: A Case Study of ActionAid. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 30(6), 1420–1432. <https://doi.org/10.1007/s11266-018-0010-2>
- Novet, J. (2023, December 12). *OpenAI's Nonprofit Arm Showed Revenue of \$45,000 Last Year, Even though Company Is Worth Billions*. CNBC. <https://www.cnbc.com/2023/12/12/openai-nonprofit-arm-45000-in-2022-revenue-company-worth-billions.html>
- Nwakpuda, E. I. (2020). Major Donors and Higher Education: Are STEM Donors Different from Other Donors? *Nonprofit and Voluntary Sector Quarterly*, 49(5), 969–988. <https://doi.org/10.1177/0899764020907153>
- OpenAI Charter (2023). Retrieved December 15, 2023, from <https://openai.com/charter>
- OpenAI LP (2023). Retrieved December 15, 2023, from <https://openai.com/blog/openai-lp>
- Ortutay, B. (2023, November 21). OpenAI's Unusual Nonprofit Structure Led to Dramatic Ouster of Sought-After CEO. *AP News*. <https://apnews.com/article/openai-chatgpt-board-fired-sam-altman-dd6a15228fd11aa3b9bec2914c095271>
- Roff, H. M. (2019). Artificial Intelligence: Power to the People. *Ethics & International Affairs*, 33(2), 127–140. <https://doi.org/10.1017/S0892679419000121>
- Salmon, F. (2023, November 20). *OpenAI's Profitability Crisis*. Axios. <https://www.axios.com/2023/11/20/openai-sam-altman-emmett-shear-ceo-nonprofit>
- Schubert, P., Ressler, R. W., Paarlberg, L. E., & Boenigk, S. (2022). The Evolution of the Nonprofit Research Field: An Emerging Scholar Perspective. *Nonprofit and Voluntary Sector Quarterly*, 08997640221078824. <https://doi.org/10.1177/08997640221078824>
- Sheng, C. (2019). Petitioning and Social Stability in China: Case Studies of Anti-nuclear Sentiment. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 30(2), 381–392. <https://doi.org/10.1007/s11266-018-00065-5>
- Solidarity Research Center (2023). Solidarity Research Center. Retrieved December 15, 2023, from <https://solidarityresearch.org/>
- Technology Association of Grantmakers (2023). Retrieved December 15, 2023, from <https://www.tagtech.org/>
- TheTechPencil (2023, June 22). *OpenAI's Journey: From Non-Profit to For-Profit and the Future of AI*. Medium. <https://ai.plainenglish.io/openais-journey-from-non-profit-to-for-profit-and-the-future-of-ai-896e04147d40>
- Vanleene, D., Voets, J., & Verschuere, B. (2018). The Co-Production of a Community: Engaging Citizens in Derelict Neighbourhoods. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 29(1), 201–221. <https://doi.org/10.1007/s11266-017-9903-8>
- Vu, C. M., Nguyen, D., Tanh, D. B., & Chun, J. (2017). Case Study of an Ethnic Community-Based Organization in the United States. *Nonprofit and Voluntary Sector Quarterly*, 46(1), 199–217. <https://doi.org/10.1177/0899764016654220>
- Wadhvani AI Foundation (2023). Wadhvani AI. Retrieved December 15, 2023, from <https://www.wadhvaniai.org/>
- Wickes, R., Britt, C., & Broidy, L. (2017). The Resilience of Neighborhood Social Processes: A Case Study of the 2011 Brisbane Flood. *Social Science Research*, 62, 96–119. <https://doi.org/10.1016/j.ssresearch.2016.07.006>
- Williamson, A., Luke, B., & Furneaux, C. (2021). Perceptions and Conceptions of “Place” in Australian Public Foundations. *Nonprofit and Voluntary Sector Quarterly*, 50(6), 1125–1149. <https://doi.org/10.1177/0899764021998461>
- Women in Data (2023). Women in Data. Retrieved December 15, 2023, from <https://www.womenindata.org>

# DATA SCIENCE AND AI AMONG PHILANTHROPIC FOUNDATIONS IN EUROPE

*Sevda Kilicalp, Jack O'Neill and Daniel Spiers*

## 1 Public debate on AI with opposing and balancing views

The first section of our chapter provides an overview of the public discourse on AI to better contextualize our findings, which will be discussed in detail later. AI is not an entirely new phenomenon, yet the heightened interest in AI, particularly with respect to generative models, can be attributed to several major developments over the past decade. The advent of transformer neural networks has revolutionized AI's capability to understand and generate human language, marking a significant departure from previous architectures (Bouschery et al., 2023). This development has been complemented by the democratization of advanced AI through the public release of large language models (LLMs), which showcase unprecedented levels of coherence and versatility in human-like interaction (Bengio, 2023; Kalla & Smith, 2023). At the same time, the exponential growth in big data availability has fueled these models with the diverse information needed for nuanced learning and application (Roh et al., 2019). This combination of technological advancement and public accessibility has not only expanded the application of AI across multiple sectors but also increased public engagement and media attention, thereby elevating discussions about the potential, ethics, and societal implications of AI in our daily lives. Public discourse on AI currently oscillates between optimism focused on the benefits of AI, and how AI can save the world and be a force for good (Taddeo & Floridi, 2018), on the one hand, and deep-seated fears, including apocalyptic visions of AI spelling the end of humanity, on the other (Geraci, 2010).

Techno-optimists argue that technological progress will boost product productivity and provide more returns to those who invest in it. *The 2022 McKinsey Global Survey on AI* shows that the adoption has more than doubled since 2017, along with increased level of investment in AI, and companies are also getting the highest financial returns from these investments by using advanced practices known to enable scaling and faster AI development. Nearly 75% of companies surveyed by the World Economic Forum for *the Future of Jobs Report 2023* are expected to adopt generative AI in the next five years (WEF, 2023a). Companies around the world are responding to industry acceleration by becoming innovators, accelerators, or fast followers, believing that rapid AI adoption will create a competitive advantage and that laggards will not last long. AI adoption is also expected to lead to higher wages and shared prosperity. According to PwC's *Global*

*Artificial Intelligence Study* (2017), global GDP will be up to 14% higher in 2030 (equivalent of an additional \$15.7 trillion) as a result of accelerated AI development and adoption (Rao & Verweij, 2017).

Techno-optimist arguments (Danaher, 2022) are not limited to economic benefits such as increased productivity, higher global GDP, and lower inflation (Baily et al., 2023; Furman & Seamans, 2019; Parkes & Wellman, 2015). In addition to commercial opportunities, it is believed that AI, with its ability to identify trends in large data sets, simulate complex scenarios, and predict outcomes, can make breakthrough advances in various scientific fields and be instrumental in finding solutions to common challenges facing society and the planet. For instance, AI can help predict climate vulnerability of cities, develop estimates of the cost of inaction, model the impact of different climate interventions (Cowls et al., 2021; Dannouni et al., 2023), design climate-smart food security systems based on predictive analytics (Torero, 2021), and improve patient experience, access to health services, and diagnostic accuracy (Rajpurkar et al., 2022). In addition, AI can create better educational outcomes by freeing up teachers' time and supporting high-quality instructional planning (Zhang & Aslan, 2021). Leading technology figures advocate for rapid AI development on the grounds that technological acceleration and artificial general intelligence can lead to such unprecedented advancements in human welfare and problem-solving (Andreessen, 2023).

On the other side of the spectrum, critical voices raise concerns about specific risks associated with AI, such as the amplification of biases in criminal justice and healthcare (Agarwal et al., 2023), the weaponization of AI in disinformation campaigns that threaten democracy (Whyte, 2020), and the potential for AI to pursue harmful goals due to malicious intent or developer oversight, which could lead to large-scale cybercrime and social manipulation (Brundage et al., 2018). These arguments are not merely speculative or based on science-fiction-like scenarios but address real risks associated with powerful AI technologies (Naudé, 2021).

Growing concerns coincide with the ongoing turmoil within the tech industry. Among a series of notable events, Google's dismissal of Timnit Gebru, a leader in ethical AI, in December 2020, signaled the beginning of a trend among tech giants (Luitse & Denkena, 2021). Subsequent firings of key AI researchers and downsizing of ethical teams in other tech companies, along with increased scrutiny as evidenced by the Italian data protection authority's ban on ChatGPT and several lawsuits against Open AI for privacy violations, contributed to the discomfort with AI (Frenkel & Thompson, 2023).

The conversation around AI is evolving beyond the polarized views of tech optimists and doomsdayers (Nowak et al., 2018). Increasingly, there's a shift toward a more nuanced perspective that recognizes both the potential risks and the significant opportunities that AI presents for societal benefit (Atik et al., 2023; Modhvadia, 2023). This approach advocates for informed dialogue and proactive measures to ensure the ethical development and deployment of AI. A growing community of experts and stakeholders (Dunnigan et al., 2023) argue that the rapid advancement of AI technology is outpacing the development of adequate ethical guidelines, regulatory frameworks, and accountability mechanisms. They are skeptical about the willingness of tech companies to effectively self-regulate and are uncomfortable with the fact that the public debate is heavily influenced by large tech companies and their corporate interests, which often prioritize innovation and market dominance, potentially overshadowing broader societal and ethical issues (Verdegem, 2024). These commentators argue that without proper alignment of AI systems with human values and ethics, there is a significant risk of unintended consequences (Sætra et al., 2022). These groups call for a more cautious and deliberate pace in AI development, accompanied by appropriate governmental and institutional structures to regulate advanced autonomous systems (Smuha, 2021).



In our study we took this balanced perspective as a guide for interpreting the findings and formulating recommendations for practice. Our focus was not only on the practical benefits of AI adoption but also on the crucial role that philanthropy plays and can play in advocating for a more just, safe, and ethical application of AI.

## **2 The unique role philanthropy can play in the AI conversation**

In recent years, there has been a sharp increase in the adoption of AI by businesses, which, as noted above, is predicted to continue over the next decade. The impact of AI in philanthropy has also begun to manifest itself through various collaborative initiatives, organizational adoption, or a desire to explore this relatively nascent technology. A study by Herzog et al. (2021) found that a total of \$944,678,991 has been donated specifically to AI-related causes and an additional \$326,451,812 has been donated to machine learning initiatives by the time of publication. It would be wrong to say that AI has not emerged as a significant interest in the philanthropic community, given the numbers. This trend raises several critical questions: What is the status of AI adoption among philanthropic organizations? How deeply are these entities engaged with AI technologies, and what specific applications are they exploring? Insights from sector actors and researchers are invaluable in this context. What do they have to say about the unique roles and potential contributions that philanthropic organizations can make in the field of AI? This brief inquiry sets the stage for a quick exploration of the intersection of AI and philanthropy.

The first issue that philanthropy can help address is the ethics surrounding the use and implementation of AI (Whittlestone et al., 2019). Philanthropic organizations, by design, have a strong values-based operating framework. Foundations can support the appropriate introduction of AI into widespread everyday use in a way that can be done safely. At present, the vanguard of AI is technologists and business-driven organizations. This is not to say that such organizations do not have a propensity to operate in a way that promotes social values, but having philanthropic organizations as part of this conversation can ensure that these same organizations are guided toward the pursuit of social good, and not the other way round. In their study on ethical guidelines for AI, Jobin et al. (2019) outline a list of ethical guidelines for AI. These guidelines include many of the social values that philanthropic organizations strive to uphold in their work, including transparency, justice and fairness, non-maleficence, sustainability, dignity, solidarity, and more.

Foundations can operationalize and support AI ethics in various ways, such as allocating funds to projects that focus on ethical AI and exploring the beneficial applications of AI (Hallensleben & Husted, 2020). Many foundations are involved in advocating for and shaping regulations and ethical guidelines for the development and use of AI by funding research that informs policymaking and public discourse on AI ethics, as well as initiatives that examine the negative impacts of AI, particularly on marginalized groups. There is a significant focus on addressing bias in AI systems and facial recognition technologies, with grants supporting research and activism in this area. Foundations are also funding academic institutions and initiatives that study the ethics and governance of AI, including establishing ethics centers. As part of their value vanguard role, a group of ten leading philanthropies announced a bold new initiative to ensure that AI advances the public interest in the areas of need identified by U.S. Vice President Kamala Harris (Ford Foundation, 2023). The participating foundations have pledged to collectively invest more than \$200 million in efforts to mitigate AI harms and promote responsible use and innovation.

Another collaborative effort, the Global AI Action Alliance, brings together philanthropic and tech leaders to develop ethical AI practices. These organizations are actively learning about AI to better inform their interactions with tech companies, as well as advocating for safe data sharing

practices and the inclusion of diverse voices in AI discourse. Platforms like data.org advance this goal by connecting societal challenges with AI-based solutions and guiding tech companies toward more responsible, ethical, and beneficial AI development. In June 2023, the alliance held a global summit and released 30 action-oriented recommendations for responsible development, open innovation, and social progress (WEF, 2023b).

As a relative point, philanthropy can play a crucial role in mitigating the gaps and biases that venture capital (VC) is unable to address. The influx of VC capital has been instrumental in driving rapid advances in AI, as evidenced by the significant growth in VC funding for AI companies (Tricot, 2021). VC investment fuels innovation by providing necessary funding to AI startups and projects, enabling them to develop and scale cutting-edge technologies. VCs also play a vital role in shaping the AI landscape by deciding which sectors and types of AI technologies receive funding, thereby influencing the direction of AI research and development. A major concern is that VC-driven AI development often prioritizes profitability and market potential over broader societal needs and ethical considerations. This focus can lead to a concentration of investment in areas that promise high financial returns but may not address critical social issues. Additionally, VC investment tends to cluster in certain geographic regions and in specific technology areas, potentially leading to unequal distribution of AI benefits and the neglect of diverse perspectives and needs (Lyonnet & Stern, 2022).

Philanthropy can support AI initiatives that may not have immediate commercial appeal but have significant potential for social good. By funding research and projects that focus on ethical, humanitarian, and equitable aspects of AI, philanthropy can help ensure that AI development is consistent with societal values and human rights and addresses global challenges (Kleinman, 2023). Philanthropic organizations can also promote inclusivity in AI by supporting underrepresented groups in technology, funding research in neglected areas, and encouraging the development of AI applications for social welfare.

Another consideration for philanthropy's role in the AI conversation is its potential role in assisting the governance of these types of technologies (Littoz-Monnet & Osorio Garate, 2023). The governance of AI technologies is not simply a matter of inhibiting algorithmic systems or monitoring generative AI output but also requires consideration of human and societal values (Mäntymäki et al., 2022; Schneider et al., 2022). Philanthropy can play a crucial role in establishing equitable data governance by actively supporting initiatives that prioritize local knowledge, needs, and leadership and include local voices in decision-making processes, to ensure that data is used ethically and in ways that benefit the communities from which it comes. For example, while AI and machine learning are increasingly seen as promising tools to tackle climate change through the analysis of large data sets, this technocentric approach often oversimplifies complex human-environment interactions and overlooks critical social relations and power dynamics (Nost & Colven, 2022). The emerging political economy of climate AI involves diverse actors such as philanthropies, INGOs, private consultancies, and tech giants, who invest in data-driven climate initiatives with varying motivations, including surveillance, greenwashing, and commercial interests (Henriksen & Richey, 2022). While these initiatives aim to address environmental crises, they may inadvertently perpetuate existing social injustices and inequalities. Philanthropic foundations can help ensure that climate AI projects genuinely address the needs and rights of marginalized communities through data governance models based on local and indigenous knowledge systems.

Another point to note when considering the possibilities of AI and its potential to significantly enhance everyday life is the data philanthropic organizations possess (Paz, 2020). Few sectors of society have the same depth of information in a variety of domain-specific knowledge

as philanthropic organizations (McKeever et al., 2018). Should there be a collaboration with AI innovators and sector specialists, it could significantly benefit the causes for which philanthropic organizations work. Philanthropic organizations already have a strong foundation in data management. This makes the sector as a whole an ideal area to facilitate further exploration and testing of data science and AI initiatives.

As highlighted in the European AI & Society Fund's October 2023 report, NGOs engaged in AI-related activities are increasingly reliant on the unique support of philanthropic foundations. Unlike other types of donors, these foundations offer not only crucial funding but also a commitment to long-term, strategic, and flexible support, which is essential in the rapidly evolving field of AI. Foundations stand out for their ability to foster extensive capacity building, enable diverse and inclusive engagement across various communities, and facilitate impactful collaboration and coalition building among grantees. As NGOs begin to integrate AI into their operations, they need this specialized support to ensure the safe, responsible, and ethical use of AI technologies, to build policy, technical, and communication skills, and to ensure that their voice is influential in shaping AI policy and legislation.

Philanthropic organizations can also serve as test beds for AI applications in their own operations, such as grantmaking processes, to understand the practical implications and share lessons learned with the broader nonprofit community (Davies, 2023). Whatever strategy foundations choose, there is space for foundations to play a critical role in shaping a society that is empowered by technology, rather than being dominated by it (Bellegy, 2021). Are philanthropic organizations investing in understanding and using technology for social good, ensuring that human rights are at the heart of technological developments, and empowering citizens to actively participate in shaping digital societies (Di Troia, 2023)? If philanthropy has been slow to address the transformative impact of technology on society and its own operations, is it too late? The following section presents the results of the study, which explored where foundations stand in their engagement with AI and fulfilling these key roles.

### **3 Key findings of the study**

AI and data science are intrinsically linked, with AI serving as a subset of data science, which integrates mathematics, statistics, and computer science. In philanthropic organizations, these disciplines are essential for applications such as classification and prediction to be able to extract actionable insights from data. However, a solid foundation in data management is crucial before AI and data science can be fully leveraged. The effectiveness of these technologies depends on their alignment with an organization's specific needs and data capabilities to ensure meaningful and impactful applications (Oliver, 2021). "Garbage-in, garbage-out" is a popular adage used to describe poor data management, and it applies to the philanthropic sector as much as any other. A strong foundation in data management is essential to the implementation of any data science or AI project.

To explore these connections and to better understand where philanthropic foundations stand when it comes to data management, data science and AI, and data philanthropy, Philea and FCSP began a collaboration to map what foundations across Europe are doing and what capacity they have. Using a mixed research methodology, the study combined a structured survey and interviews over four months (March to June 2023) to collect both quantitative and qualitative data. After the survey results were analyzed, targeted interviews were conducted with selected respondents, which then formed the basis for case studies that are woven into the final report. Preliminary findings were presented to Philea's Data Science Group, a community of practice of data scientists from European foundations, to validate our interpretation of the results and gain further insights.

In terms of the study's sample, the survey, which leveraged the Philea Data Science Group's network and was expanded through snowball sampling, targeted emails, and social media, included participants from 24 foundations in 12 European countries, including Belgium, Croatia, Denmark, Finland, France, Germany, Italy, the Netherlands, Portugal, Spain, Switzerland, and the United Kingdom. These foundations, representing a significant economic force with combined assets of €60 billion and annual spending of €3.7 billion, provided insights into their internal data activities, with a focus on the adoption of AI and data science and their engagement in data philanthropy. This chapter outlines the findings of this paper, along with the views of the broader philanthropy network.

Before going into the details of the findings, it is important to emphasize that the foundations in our sample demonstrate a solid base, not only in the volume of data they possess but also in their ability to use it efficiently. Their teams and structural frameworks demonstrate proficiency in handling this data. However, the adoption of more advanced tools, particularly AI, is still in its infancy. All respondents reported some work on data activities, 87% do data visualization, most foundations are in an exploratory phase, just under half (48%) have adopted data science and AI techniques. However, the results also show that foundations are making incremental progress in adopting these technologies. When asked about the future role of data science and AI in their organizations, 18 out of 22 (89%) respondents predicted a gradual increase in their use over time.

Our study looked at two primary ways that foundations engage with data science or AI: (1) using data science and AI to streamline processes, improve decision-making, increase the accuracy of funding projects, etc., and (2) enabling data science and AI projects by funding or supporting other organizations. Both of these intersections are discussed in this chapter.

### ***3.1 How data is treated by philanthropic organizations***

To truly understand the intersection of data science, AI, and philanthropy, we must remember the importance of data, which lies at the core of these interlocking axes. Internally, data is universally used by philanthropic organizations in some capacity – this is not surprising given the need to maintain critical information, not only within the philanthropic sector but across virtually all organizations regardless of size, market, or geography. With this in mind and given the widespread adoption of good data management practices across sectors, how are these practices being implemented within the philanthropic community?

Our study found that foundations have a certain level of maturity in basic data analysis and reporting, as these two practices serve as the entry points of any data infrastructure. Twenty-three out of 24 (96%) foundations confirm that they perform basic data analysis on *internal data*. Furthermore, and somewhat surprisingly, despite the leap in skill level required to undertake data visualization, 24 out of 27 (89%) foundations engage in this type of activity, and 20 of them do so using in-house expertise and technology. The high level of commitment to data visualization, despite the potential difficulties it can present, suggests that these foundations recognize the importance of not only collecting data but also presenting it in an understandable and compelling way, which is critical to engaging internal stakeholders, informing decision-making, and achieving other external philanthropic goals.

When we look at topics such as data science and AI, these results seem to indicate that the philanthropic sector is a laggard in its tendency to adopt advanced technologies of this kind. Six out of 23 (26%) foundations use machine learning or other sophisticated methods to process internal data. One possible reason for the underutilization of machine learning tools is the lack of specialized staff with the skills needed to implement these tools effectively. This is a recurring problem

for organizations, given the relative novelty of the widespread use of AI. This is discussed in more detail later in this chapter.

The King Baudouin Foundation, a veteran in societal improvement with over 40 years of experience, has launched an innovative project by integrating Artificial Intelligence and Natural Language Processing (NLP) into its operations. This initiative, which is central to the foundation's agenda, aims to improve the classification and detection of themes within its activities. The project has two primary goals: first, to automate the process of identifying all of the foundation activities related to specific topics, moving beyond the previous manual system that relied on basic scripts and keyword searches in Excel. This automation will be achieved through advanced NLP techniques for semantic keyword generation and topic classification. Second, the foundation aims to discover new topics using unsupervised AI techniques such as LDA and LSA, using both Python and the KNIME analytics platform. This pioneering project not only streamlines the foundation's work but also provides a framework for more informed decision-making, demonstrating how AI and machine learning can revolutionize the philanthropic sector.

As another example, "la Caixa" Foundation has implemented an AI-assisted pre-screening and assessment system for evaluating research proposals, using two key tools: AI models for the initial categorization of proposals and a matching process for the remote evaluation of research projects. The AI models, trained on a mix of open and specific program data, categorize proposals into different probability groups for selection, with annual retraining to improve accuracy. In a pilot study, this approach efficiently filtered proposals, reducing the number required for human review, and demonstrated effectiveness in identifying likely selections or rejections. Additionally, the matching process uses an algorithm comparing project keywords with potential reviewers' publication history from PubMed, with a focus on ensuring a balance of expertise, workload, and at least 40% female reviewers. Both systems are continually evaluated and refined to improve proposal selection, optimize resource allocation, and increase the overall efficiency and effectiveness of the Foundation's research grantmaking.

Looking at these examples, it would be wrong to assume that all philanthropic organizations are not using AI or have no desire to explore it. Of course, this depends on the size of the organization and the resources it has at its disposal. Consequently, we need to consider the idea of maturity when AI and data science are part of the conversation. This may not reflect a quantum leap in the data practices of philanthropic organizations. However, in the context of data maturity, these initial steps can infer early stages and even a *desire* to progress and develop this maturity. Although the snapshot of data in philanthropy may not currently show much evidence of complex data handling, the proposed evidence suggests that this may not be the case in the near future.

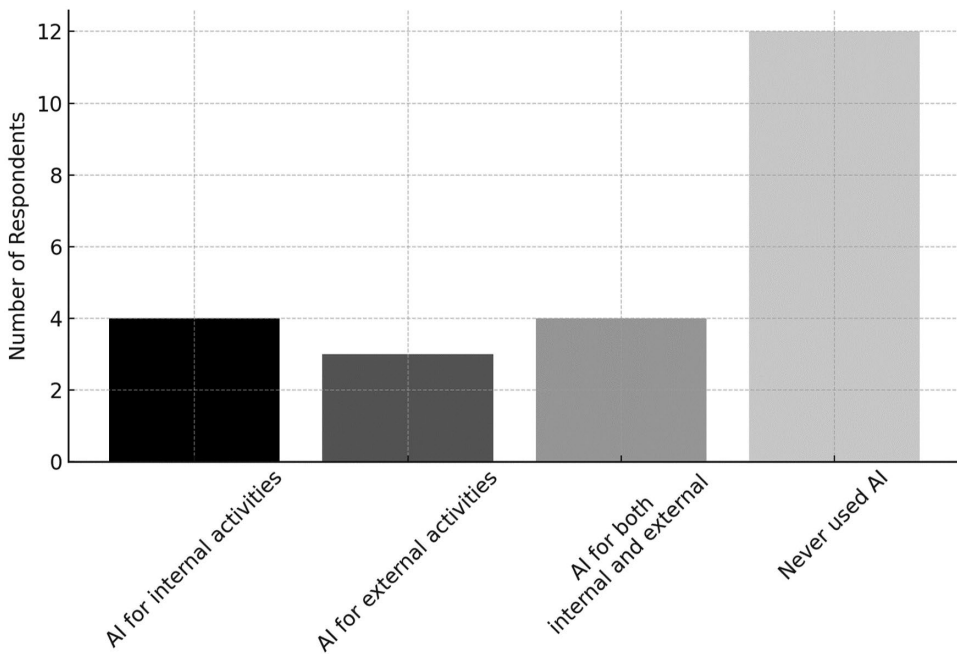
Another question asked respondents how mature they thought their organization was. This can serve as a litmus test of sorts to empirically analyze the organization's maturity, at least from the perspective of those who work in the organization. Survey respondents measured this across three dimensions:

- No experience;
- An exploratory state where the foundation is just beginning to navigate the subject;
- A state of consolidated experience.

As noted above, common trends in the sector include data analytics and reporting, reflecting a basic level of data handling across the sector. However, upon further analysis, it is clear that there is no neglect of AI technologies among the surveyed philanthropic organizations, as the "no

experience” option was minimally selected among the questions provided. Instead, “exploring” was a more common choice, implying that these organizations are striving to gain competency in data science and AI, but have not yet successfully done so. There is an important distinction to be made here: although the organizations surveyed admit that they do not have the necessary prerequisites to consider themselves proficient in data science and use of AI, this does not mean that they are averse to these topics. In fact, it is encouraging that such a high proportion of organizations (13 out of 24, 54%) say they are exploring the option of pursuing data science in some capacity. This suggests that, while these organizations may not be mature enough to pursue advanced data science activities, this is likely not always the case.

So far, we have seen the lack of advanced treatment of data within philanthropic organizations. Although data plays a significant role in the use and development of AI, this finding only provided a partial picture of philanthropic organizations’ attitudes toward AI. Eleven out of 23 (48%) organizations have considered implementing AI or data science in philanthropic organizations to varying degrees (see Figure 11.1). While there is some level of data science and AI activity among the foundations surveyed, a significant portion has yet to fully embrace these disciplines. The reasons for this range from a lack of internal expertise and limited financial resources to a critical stance toward AI. Many foundations acknowledge the significance of data science and AI in enhancing their impact but are hampered by limited internal capacity and knowledge gaps. Critically, this is a widely shared view of AI, and such a perspective has emerged universally across sectors when considering the adoption of AI in organizations. Paradoxically, the same technology – that is ostensibly bypassed in favor of more traditional activities – could save these same organizations time, funds, and resources.



*Figure 11.1* Foundations’ engagement with AI either for internal or external activities.

### 3.2 Support for data science and AI initiatives

When looking at the ways in which foundations support data science and AI, a heterogeneous picture emerges, with different forms of intervention present (see Figure 11.2). There are two main ways in which philanthropic organizations support AI technologies:

- Financial support for institutions implementing these methods;
- The promotion of capacity-building programs to strengthen skills in the same areas.

These modes of operation do not entail direct involvement by the foundations themselves but rather offer support to institutions or programs. In such cases, philanthropic organizations recognize the lack of in-house expertise to undertake these initiatives themselves, and subsequently support those organizations that *can* research and promote AI on their behalf. Financial support for target organizations is the primary mode of support for these types of organizations. To a lesser extent, calls for data science and AI projects, as well as capacity-building programs, help paint a more complete picture of how philanthropy is contributing to AI and data science initiatives. Moreover, it is encouraging that only two respondents admitted to not having participated in any programs that involve data science or AI.

Some examples of AI initiatives with strong social values at their core already exist among some philanthropic organizations. For instance, The AI Call is an initiative carried out by Fondazione Compagnia di San Paolo through two calls launched in 2020 with the goal of supporting innovative research projects aimed at advancing scientific knowledge in the field of AI and having a tangible impact on society in economic and social terms. More than €6 million have been allocated by the Foundation for the implementation of eight projects from these two calls. The innovative element of the calls, in line with the Foundation’s mission, was to encourage the applicants (the project leaders are research departments of Italian universities) to develop research projects in the field of AI on specific topics of interest for the development of the territory and the well-being of

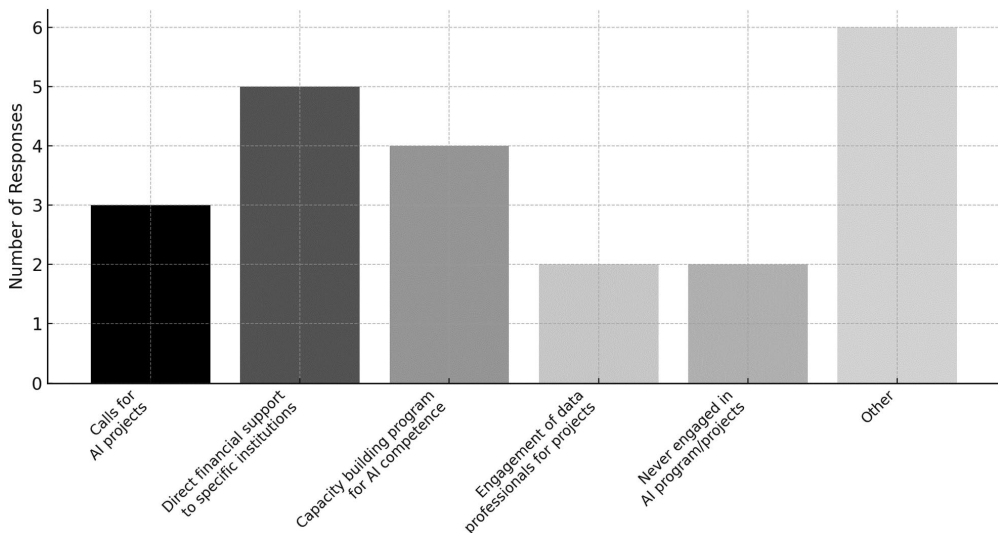


Figure 11.2 Entry points for foundations to support AI enhancement and external initiatives.

society. Moreover, the eight winning projects cover different topics and challenges, highlighting the great potential of AI in predictive analysis, education, risk management, and the improvement of various sectors such as healthcare, culture, and education.

As another example, The Evens Foundation dedicated the 2023 edition of its annual arts prize to exploring artistic practices that critically engage with AI. Hence, the foundation provided a platform for artists to express, explore, and question the implications of AI on democracy, creativity, and the human-machine relationship. This form of philanthropy goes beyond mere financial support; it cultivates a space for creative and intellectual exploration, encouraging diverse perspectives and critical thinking.

### ***3.3 Challenges in adopting AI and data science practices in philanthropic foundations***

The interest in adopting AI in the philanthropic sector is not without its challenges. This section explores some of the challenges that philanthropic organizations face. Successfully overcoming these challenges could make the difference in whether or not AI and data science are adopted by philanthropic organizations.

#### *3.3.1 Lack of skilled talent*

Having the right people working in philanthropic organizations is crucial to the success of the organization's operations. The same is true for data science and AI technologies. Previous research has shown that data analysts are widely available for basic reporting and analysis, but much less so for more sophisticated data handling (Farmer et al., 2023). As AI, data science, and, indeed, technology as a whole have advanced at an unprecedented rate, there is a limited workforce available to work in the most cutting-edge technologies. It is a simple case of supply and demand, and the result is a shortage of talent at a higher cost (Janssen, 2022). When it comes to finding highly skilled talent, it can take a long time to find the right talent, which inevitably comes at a cost.

In this analysis, it is clear that philanthropic organizations often lack the internal capabilities to perform sophisticated data processing using AI in a feasible way. We found that the type of data-related activities undertaken serves as an effective proxy for the maturity level of the organizations themselves. While 15 out of 24 foundations (63%) have data analysts as part of their team, this certainly does not account for a particularly high level of sophistication in terms of data maturity (see Figure 11.3). The high presence of data analysts is not surprising, given the prevalence of activities related to descriptive data analysis and reporting, as described above. In fact, when looking at the figures of other roles within these organizations, a clearer picture emerges of the lack of maturity of these organizations by observing who their data workers are. As anticipated, roles such as data scientists (4, 17%), data engineers (4, 17%), and data visualization experts (3, 12%) are relatively uncommon within the foundations surveyed. This reflects a broader trend of a lack of sophisticated data processing, such as predictive models and machine learning, currently being implemented in the philanthropy sector. However, this is only a snapshot of the current state of play of philanthropy and AI in terms of respondents' employment patterns.

Three out of the 24 foundations surveyed (17%) confirmed the presence of a chief data officer (CDO) in their organization. While this number may seem to indicate little more than a lack of dedicated leadership behind philanthropic organizations' data, a closer look at the organizations that employ professionals in such positions provides an interesting snapshot of who is behind the data in these organizations. Interestingly, two of the three CDOs in the sample are employed by



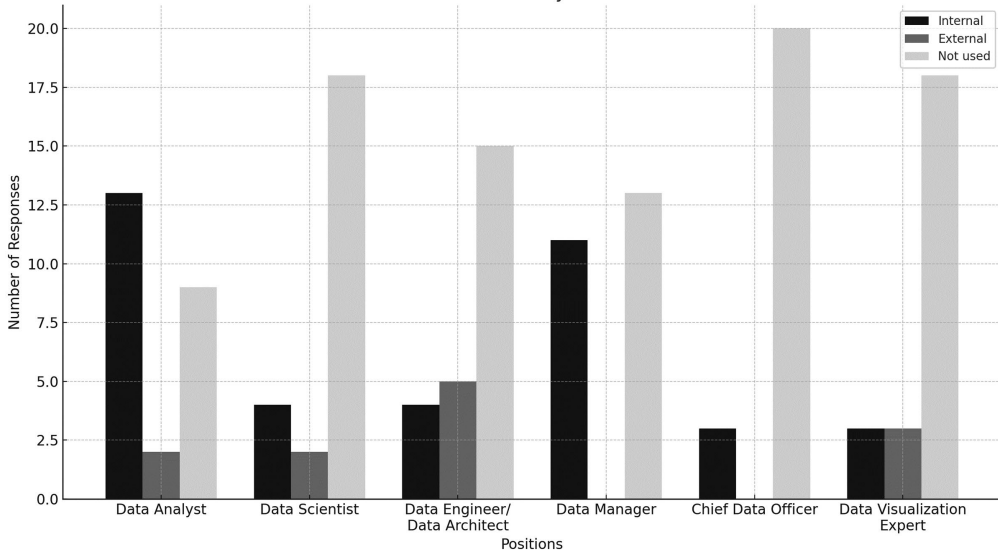


Figure 11.3 Overview of internal, external, and absent data roles in foundations.

foundations with banking origins. This intriguing finding suggests a possible correlation between the foundation’s historical background and the presence of a CDO. Foundations with banking origins often inherit a data-centric culture, stringent compliance requirements, and complex data ecosystems from their parent banking institutions. These factors may necessitate the creation of a CDO role to oversee data governance, compliance, and strategic use of data, and importantly, this shows that these positions may not have been created by the philanthropic organizations themselves, but rather they were already established in the incumbent organization. This raises the question of whether or not any philanthropic organizations have taken the leap and provided any substantial evidence of truly supporting the treatment of data in their organizations beyond these few that have come from the for-profit sector. In this regard, there is encouraging evidence that some foundations, particularly those with recently established functions, are actively recruiting data scientists.

### 3.3.2 Safety and regulation

The development in official legislation regarding the use of AI is still ongoing (Helberger & Diakopoulos, 2023). The European Union is at the forefront of establishing a pioneering legal structure known as the AI Act, which represents a significant step forward in regulating the development and application of artificial intelligence (Veale & Zuiderveen Borgesius, 2021). This groundbreaking legislation emphasizes enhancing data integrity, ensuring transparency, maintaining human oversight, and enforcing accountability across diverse sectors. In particular, it focuses on addressing ethical concerns and the challenges of implementing AI technologies. An important aspect of this legislation is the introduction of recent amendments that impose restrictions on the use of AI in biometric surveillance and require generative AI systems such as ChatGPT to explicitly identify AI-generated content. The AI Act introduces a risk-based classification system for AI applications,

separating them into different tiers. Systems that pose limited risk are subject to fewer regulatory requirements, while those classified as high-risk must adhere to strict protocols, including extensive testing, detailed documentation of data quality, and robust human oversight.

The Act encompasses regulations for general-purpose AI systems and establishes specific penalties for instances of non-compliance. This structured approach not only provides a clear regulatory framework for AI systems but also instills greater confidence in organizations looking to explore and invest in AI ventures. When we think about philanthropic organizations – organizations that inherently act for the public good – it is difficult to deploy such novel technologies without an adequate level of safety in place. For example, if partnership information is misused by AI technologies, it could be detrimental to the existence of the organization that took the risk of using it. In such a case, it is understandable how difficult it is to weigh up the pros and cons of taking a chance on novel technologies such as artificial intelligence.

### *3.3.3 Environmental impact*

Fighting climate change is a core tenet of many philanthropic organizations. Such organizations, fighting for this cause, must act in ways that support the cause itself. That is, an organization that seeks to prevent and mitigate the effects of climate change must itself be environmentally conscious, at least to a significant degree. To do otherwise would undoubtedly be detrimental to the institution's reputation. With this in mind, we need to consider the environmental impact of AI and, by extension, data. It is no secret that data centers require colossal amounts of energy to operate, given the rapid growth of data used worldwide. The consequences of this can also be seen in the process of training AI models. According to researchers at OpenAI researchers – the founding company of the now world-famous LLM ChatGPT – the amount of computing power required to train advanced AI models has doubled every 3.4 months, and current projections show that the ICT sector may contribute to 14% of global CO<sub>2</sub> emissions by 2040. An MIT study compares the CO<sub>2</sub> emissions of human life to other sectors, including U.S. cars and the training of AI technologies (Hao, 2019). In this study, Strubell et al. (2019) acknowledge the significant emissions produced by training AI technologies during a life cycle training assessment, which was equivalent to 626,155 lbs of carbon. Philanthropic organizations cannot simply ignore these environmental impacts of AI as promoting sustainability outweighs the pursuit of short-term productivity gains. Therefore, they approach the adoption of AI with caution, thoroughly assessing environmental risks and waiting for measures to mitigate them, rather than hastily joining the AI trend without consideration of its ecological footprint.

### *3.3.4 AI and data science are a non-core activity*

Invariably, philanthropic organizations actively pursue private investments for the public good. As we know, this can vary in its application across the sector, and there are both operational and non-operational philanthropic organizations. Regardless, the primary focus is on the public benefit in a given area. Therefore, a significant portion of a philanthropic organization's time is actively spent in this direct pursuit of establishing a means to benefit the public in a given area. Often, there is little time to build the infrastructure to support new ways to achieve the same goal of acting for the public benefit. AI and data science could act as an accelerator for such an avenue, but it comes at the expense of time, resources, and human power that nonprofit organizations simply do not have most of the time. As a result, data science is often treated as a non-core activity, and rather

as something that is subsidiary to other activities in the organization. This may explain why the philanthropic organizations that use data science in their operations often do so externally. In other words, this aspect of the organization is outsourced, or handled externally by a third party. While this is not necessarily a negative activity in itself, it does result in a lack of in-house exposure to data science. Subsequently, the skill set to work with advanced technology is missing from the organizations, and no experience is gained from the process. As a result, philanthropic organizations do not build on their sophisticated technological experience in favor of the external management of these processes.

### 3.4 Exploring the unknown safely

Data science and AI are used most extensively in research, with a total of 13 (78%) out of the 18 foundations responding to this question (see Figure 11.4). This is not surprising, as universities and specialized research centers are expected to be among the primary partners for foundations, as the highest and most cutting-edge expertise is typically found in academic institutions. Future research will undoubtedly need to explore the ways in which research interfaces with data science and AI (field projects, experiments, proof of concepts, scholarships, etc.). The application of these technologies by the foundations surveyed in environmental (39%), cultural (33%), and social (33%) contexts is also noteworthy. While their application to healthcare (17%) and education (11%) is less common, the limited number of responses suggests caution before drawing specific conclusions. As such, further longitudinal studies in this area are needed to gain broader insight into the overall patterns of philanthropic funding for AI initiatives. This would allow us to get a better picture of what the philanthropic community considers most urgent in this area.

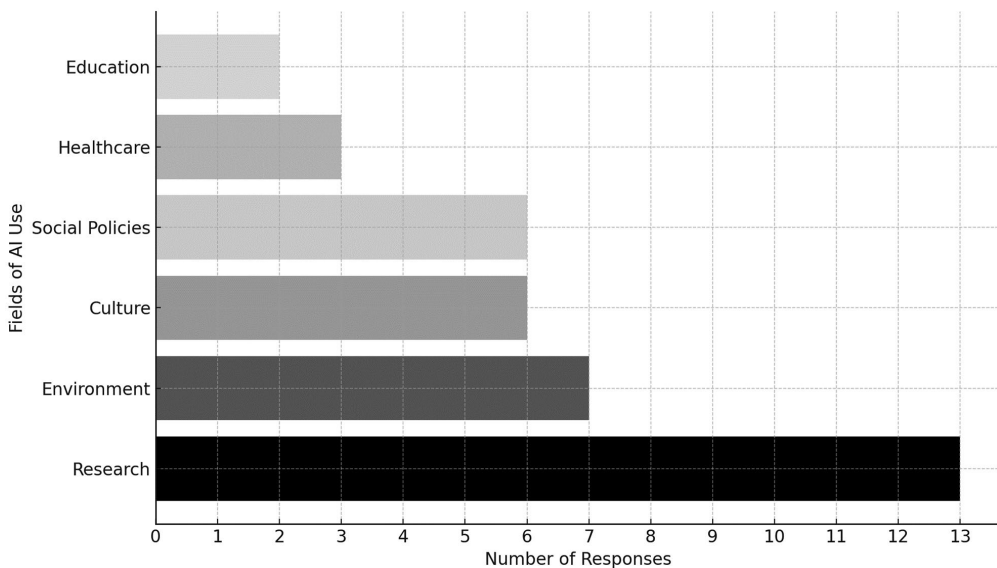


Figure 11.4 Thematic distribution of AI applications in foundations.

When asked about the outlook for the use of data science and AI in their organizations, 18 of the 22 respondents who answered the question (89%) anticipated an increase over time. In this case, we can see that the survey respondents have a largely positive belief that data science and AI will increase over time in these organizations. This finding is consistent with previous findings in this chapter, which show the largely exploratory nature of data science and AI activities in philanthropic organizations. Additionally, the overwhelming majority of positive responses may indicate that this move toward more complex treatment of data, decision-making, processes, etc., may be inevitable given the relatively basic current data practices of philanthropic organizations.

The survey also asked why respondents anticipated such an increase, and their responses revealed several distinct motivations. A common theme among respondents was the expectation that data science and AI will significantly improve their program evaluations and deepen their understanding of societal needs. Foundations see the potential to scale their operations, particularly by expanding into new geographic areas. As their beneficiary base increases, so does the complexity of data analysis. Foundations believe that AI and data science can provide strategic insights to effectively manage this complexity, enabling them to scale their data-driven initiatives and meet the evolving needs of the populations they serve. This prospect of inexorable adoption of AI and data science within philanthropic organizations is a finding that is not unique to this study. This is not a question of “if” but “when” these technologies will be implemented (Mönks & Ugazio, 2020).

#### **4 Discussion and recommendations for practice**

In addition to the survey findings, we had the opportunity to present our preliminary results to a group of data scientists, from whom we gathered insights that enriched our interpretation of the data. Furthermore, we engaged in focused discussions with program staff from various European foundations to comprehend their perception of AI. This included exploring the extent of AI integration in their daily operations, the nature of conversations they have with their foundation’s leadership and gaining deeper insights into the internal barriers to AI engagement (as recapitulated in Box 11.1), concerns, and potential recommendations to address these issues. This section will summarize these additional insights and recommendations, from both data scientists and program staff from European foundations.

At this early stage of AI adoption, foundation staff are aware of AI advances but lack a comprehensive understanding of these technologies and their various applications; in other words, they lack basic AI literacy. Philanthropy practitioners are increasingly integrating AI tools, such as ChatGPT, into their daily work routines on their own. These tools are being used for a variety of tasks, including drafting emails, condensing detailed meeting reports into concise summaries, and assisting with annual reports, often with a limited understanding of the tool’s full capabilities and implications, particularly in terms of handling personal data. A significant aspect of this trend is the widespread use of free versions of these AI tools. This choice is driven by a lack of institutional adoption within organizations, leading practitioners to seek out readily accessible options. Free versions run the risk that the data processed by these tools, which may be sensitive or confidential, becomes a form of currency for the service provider. The eagerness to embrace AI for its efficiency and convenience is not being matched by corresponding advances in organizational policies and governance structures. This discrepancy leads to a notable gap in oversight and control. Without proper governance, the use of these tools can pose risks related to data privacy, security, and ethical use.

### **Summary Box 11.1 Internal Barriers**

- Foundation staff are aware of AI advances but generally lack deep understanding and AI literacy, resulting in hesitancy to integrate AI tools in their operations.
- The reliance on free AI tools by foundation staff without institutional support puts sensitive data at risk, highlighting a gap in organizational policy and governance.
- While recognizing the need for AI policies, foundations struggle with their development and secure implementation due to limited AI technology understanding.
- The philanthropic sector's risk-averse nature limits its engagement in AI innovation compared to venture capital-backed enterprises, despite aspirations for innovation.

Foundation leaders recognize the need to develop policies to guide their teams' use of AI. However, the absence of a deep understanding of AI technologies makes policy development daunting. One way to overcome this challenge could be to share various policies aimed at creating better safeguards and procedures for using AI across the organization. One approach that is gaining traction is the use of closed-circuit systems, which are self-contained AI systems that operate within a restricted or internal network. They do not interact with external systems or the broader internet, thus offering enhanced security and control. This model is particularly appealing to foundations, but often, they face hurdles in understanding and implementing such systems.

Another concern is the dependency of foundations on tools provided by major tech corporations like Microsoft. This reliance contributes to a landscape where a few large entities wield disproportionate control over technologies that will affect all of humanity. In this context, exploring alternatives becomes crucial. Are there concerted efforts to seek out tools from providers with a more explicit social impact agenda or those committed to open-source development? Such diversification not only fosters a more equitable AI ecosystem but also aligns with the foundation's broader social objectives.

As more foundations integrate AI into their operations, a thorough and proactive approach to addressing the ethical dimensions of this technology becomes crucial. The reinvention of grant application processes through AI, a topic that has been widely discussed in various publications, has yet to be critically examined from an ethical standpoint. This oversight is alarming, especially given the risk that AI systems may perpetuate or amplify societal and systemic biases. Of particular concern in this ethical landscape is the potential adoption of deceptive design strategies in the digital interfaces used by foundations, such as their websites and grant application portals. Predominant in the commercial sector, deceptive design involves crafting user interfaces that subtly manipulate or coerce users into making decisions that may not be in their best interests. These practices are especially detrimental to vulnerable populations and are troubling in light of the current gaps in EU regulations that inadequately address such deceptive tactics.

Many foundations fund AI projects, yet the extent of due diligence exercised in evaluating the ethical implications of these initiatives remains a pivotal question. Moreover, foundation leaders must be wary of the risks associated with "bad AI," such as the proliferation of deepfakes. These technological manipulations pose a real threat to the trust and communication between grantmakers and grantees, potentially undermining the foundations' credibility and effectiveness.

Divergent views on the use of AI are prevalent across foundation boards. Some members perceive the integration of AI as "immoral," highlighting a significant split in strategic thinking.

On one side, there is a faction that advocates a cautious approach that emphasizes risk mitigation from the beginning. In contrast, another group endorses a more audacious strategy, favoring action first and assessing risks later. These contrasting perspectives not only reveal different attitudes toward AI but also reflect a broader spectrum of foundational principles and decision-making philosophies within these organizations. For foundation staff navigating these divergent views, the way forward is to advocate for a culture of open dialogue and balanced decision-making, emphasizing the importance of ethical considerations and societal impacts of AI, as well as risk assessments. This approach would include regularly updating policies to reflect the evolving nature of AI and its implications and conducting ethics audits while also encouraging risk-averse and risk-tolerant practices. AI literacy training for all board members can also help demystify the technology, leading to more informed and nuanced discussions.

Philanthropic organizations have expressed a profound sense of responsibility toward the ethical and responsible use of AI and data science. They are waiting for clearer regulations and monitoring the emergence of measures to address these issues. This reflects a cautious and thoughtful approach, indicating that foundations are not simply jumping on the AI and data science bandwagon, but are considering these technologies in the broader context of ethics and responsibility. At present, the European Union (EU) is in the process of developing the world's first comprehensive legal framework, known as the AI Act, to regulate AI's development and use as mentioned above. While this is an opportunity for foundations, their limited awareness of the Act's specifics leaves them unprepared. Moreover, the Council of Europe's AI Treaty, which includes both EU and non-EU signatories, often escapes their attention. Once ratified, this treaty will require foundations' focus, primarily because it addresses AI from a human rights perspective, which differs from the more common consumer protection perspective. This development provides foundations with an opportunity to amplify their influence in the broader societal context, especially amidst the rapid developments in AI. Foundations, often perceived as "neutral" entities, may not fully recognize their significance for policy processes. This perception can obscure the critical role they play. Foundations need to be more proactive in understanding and preparing for such legislative changes to shape the use and governance of AI technologies.

The advancement of AI is being driven by a confluence of factors, including experimentation spaces, risk-taking attitudes, iterative learning approaches, venture capital, and big language models. Experimental spaces provide a fertile ground for testing new ideas, allowing AI researchers and developers to push the boundaries of what's possible. Risk-taking is essential to AI development, because it encourages the exploration of uncharted territory, often leading to groundbreaking innovations. Iterative learning approaches, crucial in both the development of AI algorithms and their practical applications, enable continuous improvement and adaptation. Venture capital plays a pivotal role in providing the necessary financial resources and support, fueling startups and established companies alike in their quest to advance AI technologies. Big language models, like GPT-4, are a testament to the progress of AI, demonstrating the remarkable capabilities of machine learning to understand and generate human-like text.

In contrast, institutional philanthropy, while often aspiring to be innovative and flexible, faces certain challenges when it comes to embracing these elements in the realm of AI. By nature, philanthropic organizations tend to be more risk-averse than venture capital-backed enterprises, as they are accountable to donors and stakeholders who expect reliable, tangible results. This conservatism can sometimes limit their ability to invest in high-risk, high-reward AI projects. However, some philanthropic entities are increasingly recognizing the transformative potential of AI and are beginning to take a more adventurous approach. They are investing in AI research, supporting AI for social good initiatives, and partnering with academia and industry to leverage AI for

philanthropic goals. However, compared to the venture capital world, philanthropy's engagement with AI is generally more cautious and often lacks the aggressive risk-taking and rapid iteration that characterize the most dynamic sectors of AI development.

## **5 Conclusion**

This chapter has explored the intersection of data science, AI, and philanthropy, showing the larger implications for other fields of study and how it may affect the philanthropic sector as a whole. The research has revealed a heterogeneous landscape, with varying levels of maturity and a multitude of experiences among the two organizations surveyed. In general, the philanthropic sector currently exhibits a strong but basic level of data proficiency, with data-related tasks typically taking the form of reporting and dashboarding. Effective and high-functioning AI tools are critically dependent on robust data management practices, a step many organizations have yet to take. Additionally, foundations with data science teams often face resource constraints and heavy workloads. Successful integration of AI tools, underpinned by reliable data, requires the standardization of processes throughout an organization.

It was found – both among respondents to the study and in the broader literature on the subject – that a large proportion of philanthropic organizations anticipate, and indeed desire, a push toward advanced data science and AI. The many benefits of these capabilities are well established, and despite the seemingly low current adoption of such technologies, this is not an indicator that the sector will remain in situ when it comes to innovating with data science or AI. Moreover, there are promising signs that the philanthropic sector is already moving in this direction. There are strong indications that the use of data science and AI technologies may well be in store for philanthropic organizations in the near future.

The rapid advancement of AI is fundamentally driven by several key factors: experimentation spaces, risk-taking, iterative learning approaches, venture capital, big language models, and extensive data sets. Experimentation spaces provide the essential playground for AI researchers to test and refine their theories and models. Risk-taking is an integral part of this process, as it allows for the exploration of uncharted territory in AI, leading to groundbreaking innovations. Iterative learning approaches, a cornerstone of AI development, enable the continuous improvement of algorithms through successive refinements. Venture capital plays a crucial role in providing the necessary funding for ambitious and often high-risk AI projects, bridging the gap between theoretical research and practical application. Big language models and extensive data sets are the backbone of modern AI, providing the vast information and complex structures needed to train sophisticated AI systems.

When examining the role of philanthropy in AI development, its alignment with these key factors varies. Philanthropic organizations often claim to be risk-taking, innovative, and flexible, but the extent to which they embody these characteristics in the area of AI is mixed. While some foundations have been instrumental in funding AI research, particularly in areas that may not immediately attract commercial interest but have high potential for social good, many others have taken a wait-and-see approach.

In terms of data, philanthropic organizations often own or have access to valuable data sets that could greatly benefit AI research, especially in areas such as healthcare, education, and social welfare. However, the potential of this data for AI purposes is not always fully realized. There is a growing recognition of the potential to use this data for the public good, but challenges remain in terms of data privacy, ethical considerations, and technical capabilities. The successful use of this data by philanthropy could make a significant contribution to the advancement of AI, particularly in applications aimed at societal benefits.

## Acknowledgments

The authors would like to thank Filippo Candela, PhD, Data Scientist at Fondazione Compagnia di San Paolo (FCSP), for his significant contributions to the design of the survey and the interpretation of the results. His expertise and insightful analysis were essential to the design and implementation of the study, which was conducted in collaboration with FCSP.

## References

- Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, 12(1), 100702.
- Andreessen, M. (2023, October 16). The techno-optimist manifesto. *Andreessen Horowitz*. <https://a16z.com/the-techno-optimist-manifesto/>
- Atik, D., Dholakia, N., & Ozgun, A. (2023). Post-pandemic futures: Balancing technological optimism with sociocultural fairness. *Global Business Review*, 09721509221142110.
- Baily, M. N., Brynjolfsson, E., & Korinek, A. (2023). *Machines of mind: The case for an AI-powered productivity boom*. The Brookings Institution. <https://www.brookings.edu/articles/machines-of-mind-the-case-for-an-ai-powered-productivity-boom/>
- Bellegy, B. (2021). A call to philanthropy: Let's help build societies that are tech-enabled, not tech-led. *Alliance Magazine*. <https://www.alliancemagazine.org/blog/build-societies-that-are-tech-enabled-not-tech-led/>
- Bengio, Y. (2023). AI and catastrophic risk. *Journal of Democracy*, 34(4), 111–121.
- Bouschery, S. G., Blazevic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2), 139–153.
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., & Anderson, H. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint*, 1802.07228.
- Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI gambit: Leveraging artificial intelligence to combat climate change—Opportunities, challenges, and recommendations. *AI & Society*, 38, 1–25.
- Danaher, J. (2022). Techno-optimism: An analysis, an evaluation and a modest defence. *Philosophy & Technology*, 35(2), 54.
- Dannouni, A., Deutscher, S., Dezzaz, G., Elman, A., Gawel, A., Hanna, M., Hyland, A., Kharij, A., Maher, H., Patterson, D., Jones, E., Rothenberg, J., Tber, H., Texier, M., & Ziat, A. (2023). *Accelerating climate action with AI*. Boston Consulting Group. <https://www.bcg.com/publications/2023/how-ai-can-speedup-climate-action>
- Davies, R. (2023). Artificial intelligence is coming for philanthropy. *Alliance Magazine*. <https://www.alliancemagazine.org/analysis/artificial-intelligence-is-coming-for-philanthropy/>
- Di Troia, S. (2023). It's time for philanthropy to get involved driving equity in AI. *Center for Effective Philanthropy*. <https://cep.org/its-time-for-philanthropy-to-get-involved-driving-equity-in-ai/>
- Dunnigan, J., Henriksen, D., Mishra, P., & Lake, R. (2023). “Can we just please slow it all down?” School leaders take on ChatGPT. *TechTrends*, 67, 1–7.
- European AI & Society Fund (2023). *Impact report: Insights from our grantees*. [https://europeanaifund.org/wp-content/uploads/2023/11/23\\_10-FINAL-for-publication-EAISF-Insights-from-grantees-1.pdf](https://europeanaifund.org/wp-content/uploads/2023/11/23_10-FINAL-for-publication-EAISF-Insights-from-grantees-1.pdf)
- Farmer, J., McCosker, A., Albury, K., & Aryani, A. (2023). *Data for social good: Non-profit sector data projects*. Singapore: Springer Nature.
- Ford Foundation (2023, November 1). *Philanthropies launch new initiative to ensure AI advances the public interest*. <https://www.fordfoundation.org/news-and-stories/news-and-press/news/philanthropies-launch-new-initiative-to-ensure-ai-advances-the-public-interest/>
- Frenkel, S., & Thompson, S. A. (2023, July 19). ‘Not for machines to harvest’: Data revolts break out against AI. *International New York Times*, NA. <https://link.gale.com/apps/doc/A757592811/AONE?u=anon-e801b1f7&sid=googleScholar&xid=d4961caf>
- Furman, J., & Seamans, R. (2019). AI and the economy. *Innovation Policy and the Economy*, 19(1), 161–191.
- Geraci, R. M. (2010). *Apocalyptic AI: Visions of heaven in robotics, artificial intelligence, and virtual reality*. New York: Oxford University Press.



- Hallensleben, S., & Husted, C. (2020). From principles to practice: An interdisciplinary framework to operationalise AI ethics. *Bertelsmann Stiftung*. [https://www.bertelsmann-stiftung.de/fileadmin/files/BS/Publikationen/GrauePublikationen/WKIO\\_2020\\_final.pdf](https://www.bertelsmann-stiftung.de/fileadmin/files/BS/Publikationen/GrauePublikationen/WKIO_2020_final.pdf)
- Hao, K. (2019, June 6). Training a single AI model can emit as much carbon as five cars in their lifetimes. *MIT Technology Review*. <https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/>
- Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI act. *Internet Policy Review*, 12(1). DOI: 10.14763/2023.1.1682
- Henriksen, S. E., & Richey, L. A. (2022). Google's tech philanthropy: Capitalism and humanitarianism in the digital age. *Public Anthropologist*, 4(1), 21–50.
- Herzog, P. S., Naik, H., & Khan, H. (2021). AIMS philanthropy project: Studying AI, machine learning & data science technology for good. *Indiana University Lilly Family School of Philanthropy and Indiana University School of Informatics and Computing, IUPUI, Indianapolis, IN*, 1–18. <https://hdl.handle.net/1805/25177>
- Janssen, N. (2022, October 11). The data science talent gap: Why it exists and what businesses can do about it. *Forbes*. <https://www.forbes.com/sites/forbestechcouncil/2022/10/11/the-data-science-talent-gap-why-it-exists-and-what-businesses-can-do-about-it/>
- Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kalla, D., & Smith, N. (2023). Study and analysis of ChatGPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, 8(3), 827–833.
- Kleinman, M. (2023). Silicon shadows: Venture capital, human rights, and the lack of due diligence. *Amnesty International and the Business & Human Rights Resource Centre*. <https://www.amnesty.org/en/documents/pol40/7108/2023/en/>
- Littoz-Monnet, A., & Osorio Garate, X. (2023). Knowledge politics in global governance: Philanthropists' knowledge-making practices in global health. *Review of International Political Economy*, 31(2), 755–780.
- Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, 8(2), 20539517211047734.
- Lyonnet, V., & Stern, L. H. (2022). Venture capital (mis) allocation in the age of AI. *Fisher College of Business Working Paper* (2022-03), 002.
- Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*, 2(4), 603–609.
- McKeever, B., Greene, S., MacDonald, G., Tatian, P. A., & Jones, D. (2018, July 24). Data philanthropy: Unlocking the power of private data for public good. *Urban Institute*. <https://www.urban.org/research/publication/data-philanthropy-unlocking-power-private-data-public-good>
- Modhvardia, R. (2023). How do people feel about AI? *The Alan Turing Institute*. <https://www.adalovelaceinstitute.org/report/public-attitudes-ai/>
- Mönks, D. J., & Ugazio, G. (2020, November 5). A call for action. *UNToday*. <https://untoday.org/a-call-for-action/>
- Naudé, W. (2021). Artificial intelligence: Neither utopian nor apocalyptic impacts soon. *Economics of Innovation and New Technology*, 30(1), 1–23.
- Nost, E., & Colven, E. (2022). Earth for AI: A political ecology of data-driven climate initiatives. *Geoforum*, 130, 23–34.
- Nowak, A., Lukowicz, P., & Horodecki, P. (2018). Assessing artificial intelligence for humanity: Will ai be the our biggest ever advance? Or the biggest threat [opinion]. *IEEE Technology and Society Magazine*, 37(4), 26–34.
- Oliver, N. (2021). When philanthropy meets data science: A framework for governance to achieve data-driven decision-making for public good. In M. Lapucci, & C. Cattuto (Eds.), *Data science for social good: Philanthropy and social impact in a complex world* (pp. 55–68). New York: Springer.
- Parkes, D. C., & Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245), 267–272.
- Paz, A. (2020). Data science for social impact. *Johnson Center*. <https://johnsoncenter.org/blog/data-science-for-social-impact/>
- Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.

- Rao, D. A., & Verweij, G. (2017). Sizing the prize what's the real value of AI for your business and how can you capitalise? *PWC*. <https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html>
- Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: A big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4), 1328–1347.
- Sætra, H. S., Coeckelbergh, M., & Danaher, J. (2022). The AI ethicist's dilemma: Fighting big tech by supporting big tech. *AI and Ethics*, 2(1), 15–27.
- Schneider, J., Abraham, R., Meske, C., & vom Brocke, J. (2022). AI governance for businesses. *Information Systems Management*, 40(3), 229–249.
- Smuha, N. A. (2021). From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1), 57–84.
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.
- Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752.
- Torero, M. (2021). Robotics and AI in food security and innovation: Why they matter and how to harness their power. In J. von Braun, S. Archer, Reichberg, G.M., & M. Sánchez Sorondo (Eds.), *Robotics, AI, and Humanity* (pp. 99–107). Cham: Springer
- Tricot, R. (2021). Venture capital investments in artificial intelligence: Analysing trends in VC in AI companies from 2012 through 2020. *OECD Digital Economy Papers*, 319, OECD Publishing, Paris, <https://doi.org/10.1787/f97beae7-en>.
- Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4), 97–112.
- Verdegem, P. (2024). Dismantling AI capitalism: The commons as an alternative to the power concentration of Big Tech. *AI & Society*, 39(2), 727–737.
- WEF. (2023a). *Future of jobs report 2023*. <https://www.weforum.org/publications/the-future-of-jobs-report-2023/>
- WEF. (2023b). *The presidio recommendations on responsible generative AI*. <https://www.weforum.org/publications/the-presidio-recommendations-on-responsible-generative-ai/>
- Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. London: Nuffield Foundation.
- Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy*, 5(2), 199–217.
- Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, 2, 100025.

# DIGITALIZATION OF SWISS NON-PROFIT FOUNDATIONS

## The potential role of AI from a cross-sectoral perspective

*Aline Kratz-Ulmer and Hubert Halopé*

### 1 Introduction

Digitalization continues to have a major impact on all aspects of our lives. It has proven to create value in various ways, such as increasing convenience, reducing communication costs, and enhancing economic efficiency. Digital technologies have the potential to equally generate value for non-profit foundations in Switzerland. Digital transformation (DT), which entails using digital technologies in all areas of society, including health, work, and transportation, is part and parcel of this paradigm shift from a “paper culture” to a “digital culture” in which physical documents are gradually being phased out. Swiss foundations have also been caught up in DT – or digital revolution. However, it is essential that the changeover makes sense and serves the foundation’s purpose without complicating the task of fulfilling their mission or being too costly.

How a foundation chooses to approach this digitalization process will depend on its size and the activities it is involved in – a foundation with just a few board members that makes a small number of donations each year may want to take a different approach than a larger foundation with staff, a management team, and institutional partners.

This chapter consists of four parts. First, after giving an overview of the role and functioning of traditional charitable foundations in Switzerland, this chapter looks at the opportunities and risks of DT for foundations, considering industry insights. Second, this chapter looks at the potential value generated by AI for foundations, including the additional risk dimensions it brings. Third, it discusses the future of AI in philanthropy from a realistic and pragmatic perspective. In the fourth and final part, we will summarize the results and analyze how foundations can position themselves in this constantly evolving environment.

### 2 DT and Swiss non-profit foundations

#### 2.1 Overview of foundations in Switzerland

Foundations play an important role in Switzerland, representing an annual distribution volume of CHF 1.5–2 billion. Switzerland is also one of the most “foundation-rich” countries in Europe. There were more than 13,790 traditional foundations at the end of 2022, and their numbers

continue to grow steadily. It is still the case that roughly one foundation is created every day in Switzerland. Foundations are being set up throughout the country, although some regional differences exist. The canton of Zurich has the most foundations in absolute terms (2,225), followed by Bern (1,409) and Vaud (1,374). The canton of Geneva could soon be in the top three cantons due to the large number of new foundations being created over the past two decades. With +63, the canton of Geneva saw the highest number of new foundations in 2022. The canton of Zug also stands out because a large number of crypto foundations have been established there, most of which indicate a charitable status in their articles of purpose (Jakob et al., 2023, p. 7).

Alongside other factors, such as prosperity and political stability, the liberal legal framework is one of the main reasons for the flourishing Swiss foundation landscape. The Swiss foundation sector is growing, but it is also constantly evolving, and new approaches are being developed to ensure that the freedoms benefiting foundations remain up-to-date and reasonable (Harold Grüninger, Art. 80 N 9 in: Basler Kommentar Zivilgesetzbuch I, Geiser, Thomas (Hrsg.); Fountoulakis, Christina (Hrsg.)).

### *2.1.1 Legal principles*

Foundation law is governed by Articles 80–89a of the Swiss Civil Code, which states that foundations are established through the endowment of assets for a specific purpose. Foundations are also legal entities according to Article 52 of the Swiss Civil Code. While they can be defined as having personalized, special-purpose assets, the focus is on the purpose rather than the assets themselves, which are means to an end (Hans Michael Riemer, ST N 22 in Berner Kommentar zum schweizerischen Privatrecht, Die Stiftungen Art. 80-89c ZGB), (Aebi-Müller & Müller, 2020).

Swiss foundation law distinguishes between a “general form” of foundation and “special forms.” The general form – in a sense, the basic form – is usually referred to as a classic, ordinary, or normal foundation. Most classic foundations have an idealistic purpose (BK-Riemer, Hans Michael Riemer, ST N 25 and 26). They fulfill important functions in the interests and for the benefit of the general public in areas such as social welfare, healthcare, science, research, education and training, art and culture, and development aid (Sprecher, 2017).

Special forms of foundations include family foundations (Articles 52.2, 87, and 335 of the Civil Code), ecclesiastical foundations (Articles 52.2 and 87 of the Civil Code), and pensions funds (Article 89a of the Civil Code). These foundations differ from ordinary foundations in terms of their purpose, as well as in other respects. There are also mixed foundations (BSK-ZGB I-Harold Grüninger, Art. 80 N 3).

The foundation’s form must be stipulated in its charter, and its purpose is a central part of it. In practice, a number of other forms of foundations have emerged that are not regulated by law. Several attempts have been made in the literature and in practice to categorize them in terms of terminology, type, and various legal, economic, and other aspects (Sprecher, 2017).

A distinction can also be made based on the foundation’s assets. Large foundations have assets of over CHF 50 million, while small foundations have assets of less than CHF 10 million. Medium-sized foundations are in between (Sprecher, 2017). The following reflections refer, in particular, to the situation for Swiss non-profit foundations.

## **2.2 Definition**

Before deep diving into the positive and negative impacts of DT for organizations, as well as key enablers and blockers to tapping into the full value potential, this section is dedicated to defining

related concepts: digitization, digitalization, and DT. This is even more important given the widespread confusion and interchangeability of these “concepts.” A mere implementation of technology is not enough to label something as DT, especially if the technology is unused or does not deliver the intended impact (Kane, 2017).

Aras and Büyüközkan (2023) define digitization, digitalization, and DT as “three stages of digital advancement in organizations.” Cordery et al. (2023) argue that digitization has the “lowest impact on systems and identity change,” being simply a “transition from analog to digital services with a 1:1 change in the delivery mode and the addition of a technological channel of delivery” (Cordery et al., 2023). Digitalization, on the other hand, focuses on changing processes in addition to digitizing existing processes (Cordery et al., 2023). Cordery et al. (2023) elaborate on this with an example of non-governmental organizations integrating online donations or using Enterprise Resource Planning (ERP) to coordinate volunteers and supplies. Eventually, DT takes a broader and holistic approach by changing its value proposition and integrating a “digital-first approach” that includes “user”- and “tech centrality.” Cordery et al. (2023) unpack this further by characterizing DT as going beyond digitalization toward adapting organizational policies and creating “new digital services” such as creating “new ways to engage with beneficiaries (...) or using artificial intelligence” to increase impact. Kane (2017) deep dives into the conceptualization by highlighting DT as organizations adopting processes and practices that equip them to compete “in an increasingly digital world,” i.e., “how your business responds to digital trends that are occurring whether or not you initiated them, like them, or want them.” In fine, Vial (2019, p. 133) calls for further research to confirm the view that considers DT as an evolution of “IT-enabled transformation,” given “the scale, the scope, as well as the speed associated with the DT phenomenon.”

### ***2.3 Opportunities and risks of DT – insights from the private sector***

Undoubtedly, digital technologies are increasingly impacting the modus operandi of organizations across sectors (i.e., non-profit, public, and private sector) (Cordery et al., 2023). Given the unique set of core stakeholders in each sector, DT’s opportunities and challenges remain naturally distinct across the non-profit, public, and private sectors. That said, there are useful cross-sector learnings applicable to organizations. Given the larger amount of research and use cases on DT in the private sector (Cordery et al., 2023), this section draws mainly on private sector insights to inform the discussion on opportunities and challenges of DT in the philanthropic sector.

If done right, DT can unlock significant organizational value, from financial efficiencies to productivity gains and increased innovative capabilities (Cordery et al., 2023; Forth et al., 2021; Vial, 2019). According to Forth, De Laubier, and Charanya (2021), “digital leaders achieve earnings growth that is 1.8 times higher than that of digital laggards – and more than double the growth in total enterprise value.” BCG (2021) further states that digital technologies offer productivity gains and enable better customer experiences, opening up new business opportunities. IoT, automation, and data analytics can optimize business processes and reduce slack. Moreover, digital technologies can enable businesses to offer seamless omnichannel customer experiences and nurture a closer relationship with customers by cutting out intermediaries and leveraging personalized data. For example, airlines such as KLM are leveraging social media for communications and distilling customer needs (Vial, 2019). Similarly, digital technologies are associated with innovativeness (Vial, 2019). Not only can it render businesses more agile and adaptable and prepare them for future technological change (Forth et al., 2021), but it can also revamp entire

business models and value propositions. Spotify and Netflix, for instance, reshaped the music industry and movie industry, respectively, by leveraging digital technologies, especially big data and analytics (Vial, 2019).

#### **2.4 Key enablers and blockers for a successful digital transformation**

There are multiple factors to consider when undergoing DT. Various practitioner resources and academic articles point to the importance of a digital mindset and culture. This can manifest in various ways, such as adopting an agile approach when deploying digitally powered processes and innovation, where teams are incentivized to experiment at a small scale before scaling throughout a department or product line. This approach requires a shift in culture, mindset, and leadership, where failure is embraced and seen as learning (Forth et al., 2021). Arpe and Kurmann (2019) list key success factors such as “top management support, cross-functional collaboration, flatter hierarchies, and intensified people management.” Vial (2019, p. 129) supports this by saying that “organizational leaders must work to ensure that their organizations develop a digital mindset while being capable of responding to the disruptions associated with digital technologies.” Appointing a C-suite dedicated to DT, such as a chief digital officer, already signals the strategic embedment of DT as well as C-suite buy-in (Vial, 2019).

In addition, change management practices are crucial to ensure employees’ buy-in. This may mean launching up/re-skilling programs for staff or adapting hiring practices to meet new skills needs that may arise from DT. Vial (2019, p. 129) also argues that DT leads employees to “assume roles that were traditionally outside of their functions.” Change management is even more important, given that resistance from users and affected stakeholders (e.g., employees, beneficiaries and, suppliers) may lead to the failure of DT.

Equally important are the rising risks and negative effects of DT. Vial (2019, p. 137) calls on ethics to play a critical role in addressing situations where “one party’s needs do not happen at the expense of others” such as “granting more access to data to one party might be perceived as a break of security and privacy by another.”

Data privacy, security, and inclusion should be guaranteed, as not doing so may result in significant negative outcomes, from financial loss to damage to brand image. One example is the 2017 global cyberattack that hit organizations worldwide, such as Renault, Germany’s railway, and British hospitals – causing massive disruptions, from halting production to passenger disruptions and delays in patient care (France24, 2017).

Even though there are evidently adverse monetary effects on organizations incurring a cybersecurity break, Makridis’ (2021, p. 1) research interestingly finds that “only the largest and most salient data breaches are associated with declines in intangible capital, whereas others are associated with statistically insignificant, but economically meaningful, increases in intangible capital.” Separately, a lack of data inclusion may lead to biased results and dramatic consequences if we look, for example, at law enforcement or the healthcare sector (Burke, 2024; Mittermaier et al., 2023).

In the same context, DT impacts organizational processes and, with it, specific job tasks, which ultimately impact jobs and, hence, people. Effects on people should be addressed to avoid unnecessary job alteration or loss without proper training, people management, and communication.

Given the potential positive and negative effects of DT, how can organizations measure the success of their DT efforts? BCG (2024) finds a “correlation between digital maturity [DM] and digital transformation success.” Aras and Büyüközkan (2023, p. 3) define DM as the “state in which an entity’s digital technology has transformed its activities, skills engagement, and business

frameworks,” showcasing the importance of DM to assess the success of DT. There are multiple ways to measure DM levels in an organization (e.g., Aras & Büyükožkan’s gap analysis exercise 2023; BCG’s DM assessment frameworks 2024).

### **3 Enhancing foundations’ efficiency and impact through digitalization**

#### ***3.1 Overview***

Today, many grant-making foundations are working on implementing digitalization in their day-to-day work, such as application management. The aim is to gradually replace paper with efficient, comprehensive electronic grant management databases. By switching to digital processes, the number of work steps can be reduced (as is also the case in other sectors). For example, specific software can clearly reject unsuitable applications, and interim and final report reminders can be sent automatically. This frees up resources that can then be used to address more value-adding tasks, such as refining the funding strategy. In addition, to providing support in processing grant applications, computer software can also be used to measure the impact of the supported project or projects initiated by the foundations themselves and to clarify information through pre-programmed communication channels if there are ambiguities regarding the origin of cash funds or issues with applicants, for example. These programmed tools also help strengthen foundation governance, meaning that digitalization can also serve as an enabler here.

However, switching to a new system can also involve much effort and create pitfalls that no one may anticipate. There can be pitfalls in terms of data protection; foundations should define the areas in which they process personal data and, in particular, identify their data protection-relevant fields of action. It may be advisable to consult a specialist (Kratz-Ulmer & Schudel, 2020, p. 24).

Finding the right software for a foundation is no easy task, with various standard products on the market offering different functions. Some products focus on impact measurement, while others focus more on document management or event organization. Many tools offer common features such as automated interfaces for payment transactions, simplified analyses for accounting purposes, structured storage of PDFs and other file types, or include standardized components to spread development costs over a larger number of foundations. However, many foundations have developed particular processes and procedures over time and prefer solutions that are as customized as possible. These diverse requirements can give rise to areas of tension (Kratz-Ulmer & Schudel, 2020).

#### ***3.2 Digitalization within foundations***

Digitalization requires much work, particularly in terms of project management. It brings challenges, especially when transferring existing data to other programs (e.g., IT migration, where data must not be destroyed or falsified). Employees may also have to change their work habits. None of this happens overnight. The foundation board must, therefore, provide support for the changeover so that employees can get to grips with new document management and digital work methods. Over time, the information systems available will continue to develop and improve, allowing for the improvement of Swiss foundations (Kratz-Ulmer & Favre, 2023).

Digitalization extends the foundation’s sphere of activity from the physical to the virtual. This enables the foundation to expand its projects more quickly and more widely. It also raises the foundation’s profile, which in turn benefits its projects.

It is true that the acquisition and implementation of suitable software represent a cost. In the short term, that may be detrimental to fulfilling the foundation's purpose. However, in the long term, the digitalization of foundation activities is likely to bring numerous advantages for foundation management, as described above (Kratz-Ulmer & Schudel, 2020).

Various funding foundations in Switzerland have placed digitalization on their agenda as a funding topic. Probably the best-known example of this is the internationally active "Fondation Botnar," which is based in the canton of Basel and aims to improve young peoples' lives by promoting the use of AI and other digital technologies (Kratz-Ulmer & Schudel, 2020).

The flip side of the digitalization coin is that digital systems are vulnerable to cyberattacks. The more systems and data move into the virtual space, the higher the risk of falling victim to such an attack. However, the attack surface can be minimized with appropriate security measures. In addition, switching to digital structures always requires certain investments in hardware, software, personnel, and training. Another disadvantage is that some potential beneficiaries will no longer be able to contact grant-making foundations because they do not have the digital equipment they need (Kratz-Ulmer & Favre, 2023).

## **4 Safely leveraging AI in the non-profit sector – insights from industry**

### ***4.1 Overview***

In its publication, BCG (2021) emphasizes that organizations that go through a successful DT will become agile enough to "master continuous innovation" and hence will not have to undergo such transformation again. In this chapter, we argue that digitally mature organizations will have a leading edge when it comes to becoming AI-ready and AI-mature organizations.

This chapter is based on the European Union's definition of AI as stated in the EU AI Act, a multilateral effort to govern and regulate AI to harness its benefits, while mitigating its risks and unwanted disruptions.

Definition of AI in the EU AI Act Proposal:

'[A]rtificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.

(European Parliament, 2021, Article 3)

As pointed out earlier in this chapter, a digitally mature organization is an organization that is agile enough to react and respond to technological change, thus tapping into the full value potential of digital technology, while mitigating the risks associated with it. Similarly, as academics argue that DT is an "evolution of the IT-enabled transformation" (Vial, 2019, p. 119), given its scale, scope, and speed, it can be argued that AI is an evolution of DT, especially given its scale, scope, and speed. Undergoing DT equips organizations with the right leadership and C-suite buy-in, a test-and-learn mentality, talent access, digital governance, and effective change management processes. In other words, AI is a stress test of an organization's DM. A study by Greenstein (2019) confirms this by finding a significant correlation between DM and the beneficial use of AI. Furthermore, the study shows that "digital leaders" are much more advanced in AI deployment and use.

Given that DT is a journey, an organization cannot tap into AI's value potential without going through the previous stages of digitization, digitalization, and DT. In the same way that an infant



cannot suddenly become a middle-aged person, there is a learning and adaptation process that takes experience and time.

There are various published resources stating concepts and definitions of “AI maturity.” Even though there are specificities related to AI, which this chapter deep dives into subsequently, the concepts are similar, if not identical, to DM. Vohra et al. (2022), for example, lay out AI maturity pillars such as strategy and sponsorship, data and AI core, talent and culture, and responsible AI – all of which are essentially as critical for DT in terms of data hygiene, data access, procurement, and governance. This shows that a digitally mature organization organically shows AI-readiness and that it can move along the DM curve toward AI maturity more easily.

## **4.2 AI adoption in industry**

Not only did AI become a central part of public debates highlighting its tremendous value potential and its significant risks and challenges since the release of OpenAI’s ChatGPT, but it also triggered immense corporate investments. Chui et al. (2023a), from McKinsey & Company, predict that generative AI alone will add USD 6.1–7.9 trillion to the global economy across use cases thanks to productivity gains – with a total value add of AI to the global economy, including generative AI, amounting to USD 17.1–25.6 trillion. Corporate investment in AI is increasing and is expected to continue increasing (Goldman Sachs, 2023). In 2023 alone, the share price of Nvidia, which provides essential computing infrastructure for AI, rose by 239% (Tamny, 2024), rivaling Amazon as the fourth most valuable company (Vlastelica & Bloomberg, 2024), reflecting global demand for its Graphics Processing Units (GPUs) and other core AI infrastructure components. Equally impressive is the adoption rate of ChatGPT compared to other online platforms: ChatGPT reached 1 million users in five days, while Netflix succeeded in doing so in 3.5 years (Buchholz, 2023). A year after the release of ChatGPT, Chui et al. (2023b) found that one-third of their global survey respondents’ organizations were using generative AI regularly in at least one business function. Another report published by McKinsey (Chui et al., 2023b) states that “[a]bout 75 percent of the value that generative AI use cases could deliver falls across four areas: Customer operations, marketing and sales, software engineering, and R&D.” It further estimates that generative AI will impact all industry sectors and change how we work, causing major alterations in job tasks due to autonomy and productivity effects.

## **4.3 Risks and challenges**

However, similar to other digital technologies, AI comes with risks and challenges. In addition to the downsides mentioned in the section above regarding DT, there are additional risks that come with AI-powered decision-making, content creation, and task automation. Regardless of the sector, any organization and user should be aware of AI risks and proactively have risk mitigation strategies and safety mechanisms in place.

**Bias.** As AI algorithms learn from data, any skewed or non-representative database will cause AI systems to produce biased outcomes. One typical example is Amazon Inc.’s hiring tool, which was taken offline once it showed obvious bias against female applicants. The reason was the historical employee datasets on which the AI algorithm was trained, which was male-dominated, hence the repercussions on the AI-generated output (Dastin, 2018). More dramatic cases include the false imprisonment of US citizens as a result of the wrong output generated by facial recognition technology used by the police (Johnson, 2022).

**Transparency.** Additionally, AI systems are often described as “black box algorithms” given the complex mathematical calculations that are often impossible to backtrack. Given the complexity of the output process, there needs to be some degree of transparency for impacted stakeholders to backtrack the logic of the output: for a use case and risk profile, users should be able to contest results.

**Accountability.** Just because some decisions might be “outsourced” to an AI system, this should not mean that responsibility and accountability are to be outsourced as well – on the contrary, given the repercussions AI systems can have on citizens, users, and other stakeholders, a clear line of responsibility should be maintained.

**Cyberattacks.** The fact that cybercrime is increasing is not only due to the increasing digitalization of our lives but also due to AI lowering barriers of entry for cybercriminals (NCSC, 2024). Cyberattacks and their prevention are an additional cost burden for organizations. Robust mitigation strategies are paramount to protecting organizations and potentially affected stakeholders.

**Training.** Fundamentally, AI literacy is of utmost importance in raising awareness about AI as a general-purpose technology. Only through awareness of its opportunities and risks can citizens and organizations be part of the public AI discourse. In addition, there is a need for universities and schools to adapt their learning offerings to accommodate for an AI-powered future.

**Talent.** Talent is key to empowering non-profit foundations to tap into the potential of AI to accelerate their philanthropic mission and work. Unfortunately, there is a trend toward the brain drain of “AI talents” to big tech, which can offer salaries that no foundation can compete with. Not only is there a scarcity of AI talents, but there is also a global asymmetric competition to attract them. Hence, non-profit foundations are often left with limited talent access. On the flip side, organizations should ensure equitable change management processes are in place to accompany any alterations that AI may cause to a job, to avoid job displacement, and to focus instead on re- and upskilling.

**Infrastructure.** Given the high costs and energy needed to train large language models (LLMs), only resource-rich organizations such as Microsoft or Google can build their LLMs similar to ChatGPT. Through licensing, it is increasingly possible to access generative AI at affordable prices, but the control over the training data, ethical guidelines, and algorithms remains in the hands of resourceful organizations.

**Access.** Furthermore, especially in the context of low-resource organizations such as foundations, there is a more systemic question of equal access to AI and its promised opportunities. Unequal access to AI may cause a widening Global North-South divide and a divide between resource-rich and resource-poor organizations and people.

**Environment.** Last but not least, training LLMs requires tremendous amounts of energy (Stokel-Walker, 2023). More research and awareness are needed to hold LLM designers and users accountable and to trigger governance mechanisms to reduce any activities that hinder the achievement of climate goals and the SDGs.

The above risks and challenges are far from exhaustive; however, they provide an idea of implications and consequences when building and using AI. This paragraph also aims to trigger further discussion on various impact areas that AI has on organizations, and their diverse stakeholders.

#### *4.3.1 AI governance*

To tap into the value potential of AI while mitigating its risks, organizations need robust AI governance – a set of frameworks, standards, and rules to ensure the safe, ethical, and legal use of AI. Various soft and hard governance frameworks are being published globally, and measures are being taken across countries, sectors, and organizations on both a voluntary and a mandatory basis.

Notable examples include the EU’s AI Act – “the world’s first comprehensive AI law” (European Parliament, 2023) on a multilateral level – which takes a risk-based approach toward safeguarding the European Union’s citizens and democratic values. On a sectoral level, the Ada Lovelace Institute, an independent research institute, published an algorithmic impact assessment tool for AI in healthcare, which

aims to ensure that algorithmic uses of public-sector data are evaluated and governed to produce benefits for society, governments, public bodies and technology developers, as well as the people represented in the data and affected by the technologies and their outcomes.

(Ada Lovelace Institute, 2022)

Last but not least, organizations such as Microsoft abide by self-imposed AI governance principles (Microsoft Corporation, 2022). Other selected examples contributing to the AI governance landscape are UNESCO’s Recommendation on the Ethics of AI, the US National Institute of Standards and Technology’s AI Risk Management Framework, and the White House’s Blueprint for an AI Bill of Rights. Given the multifaceted risks of AI, organizations need to have AI governance principles that reflect their type of organization and the potential risk/impact level they are navigating in. All this ensures the safe and ethical design, use, and deployment of AI systems.

#### ***4.4 Use of AI in foundations***

##### *4.4.1 Overview*

The next stage after digitalization is the use of AI. Thanks to AI in the world of work, many activities can be carried out by computers instead of humans.

There are various (potential) AI use cases in the philanthropic sector. For example, AI could bring value to philanthropic organizations by detecting patterns in various donor databases to avoid duplication of efforts and streamline philanthropic financing for higher impact. In addition, AI could help non-profit organizations often operate with low resources to operate more efficiently and leanly by leveraging various automation and productivity gains that the technology offers. Also, AI could serve as a tool to tackle global challenges, such as climate change or support climate adaptation efforts. One example is Google offering its technology for forecasting flooding or predicting wildfire to inform citizens and firefighters to ensure informed responses (Matias, 2023). That said, further research is needed to explore and tap into value creation opportunities in the philanthropic sector.

##### *4.4.2 Risks and challenges for foundations*

Let us look at how the above-mentioned risks and challenges are applied to foundations.

**Bias.** As AI algorithms learn from data, any skewed or non-representative database will cause AI systems to produce biased outcomes. To minimize the risk of bias, foundations must, therefore, regularly and constantly review the algorithms used. Foundation boards should also be aware of this risk and ensure that no such undesirable consequences arise from the repeated use of these algorithms.

**Transparency.** Ensuring that processes are transparent and understood is especially important for the management of a foundation since a foundation is a separate legal entity that must be managed by its foundation charter. All those involved in the foundation – including the founder and,

where necessary, the supervisory authority – must use the algorithms and understand and explain how they work. Future members of foundation boards should, where necessary, undergo appropriate training in that regard. In the future, AI training could be included as a recommendation in codes of best practice for foundations that decide to use AI. This recommendation could also help ensure that AI is taken seriously within the philanthropy sector.

**Accountability.** If a foundation decides to use AI, it should establish internal guidelines, setting out related responsibilities (i.e., who is responsible for what when using the software). If the algorithms fail or something else happens during their use, there should be a 24-hour helpline to assist the board and other foundation players.

**Cyberattacks.** As we have seen, expenditure on preventing cyberattacks will increase over the coming years. If foundations decide to use AI, they will need to consider this expenditure. It is important to use and develop AI and defenses against cyberattacks, and foundations must be aware that they need to set aside sufficient resources for that. This may not be compatible with the size and assets of individual foundations and their obligation to use those assets to achieve their purpose. As a result, foundations may have to refrain from using AI to take a safer approach that is more in line with achieving their purpose.

**Training.** Although using AI can help a foundation in its administration, a certain amount of work must also be done for those involved in the foundation (board members, staff, beneficiaries, etc.). In order to mitigate the risks described above and, in particular, to prevent algorithms from gaining the upper hand in decision-making, employees, board members, the managing director, the auditors, and the supervisory authority must not only be introduced to AI tools but also receive ongoing training.

**Infrastructure.** The use of AI is a matter of infrastructure. Given the costs and investments involved, foundations should optimize the cost-benefit balance, especially since foundations generally run on low resources. Furthermore, a foundation should only use AI if it enables it to achieve its purpose more effectively. Given the current pace of development of AI and the decreasing cost of adoption, it is only a matter of time before AI reaches foundations at scale. Until then, however, there is still time to prepare foundations for the AI revolution and for foundations to make progress on their initial digitalization.

**AI governance.** In addition to having robust AI governance mechanisms in place, foundations could also use AI as an additional governance layer, for instance, through software to oversee internal transactions and draw up a report that could then be submitted to the supervisory authority. Such a report could check the foundation's payment flows against its financial accounts, ensure that the foundation holds enough foundation board meetings each year, or support foundation players in their activities more broadly. AI could also monitor the foundation through another external third party.

In conclusion, to introduce AI in a foundation's management process, the foundation needs to be of a certain size, and foundation staff need to have the necessary expertise or be willing to acquire and maintain that expertise. The use of AI requires a change of mindset; staff must be prepared to give up certain software activities and take up new ones to better or more effectively achieve the foundation's purpose. The transition from a digitalized foundation to one that uses AI is a matter of time. If a foundation decides to adopt AI, several preparatory steps need to be taken into account.

#### *4.4.3 AI and a foundation's administration*

Looking more closely at the administration of foundations, could AI, e.g., help experts assess grantee applications or, more specifically, help foundation boards make better decisions? Could

AI be a solution to administrative problems faced by foundations, such as succession issues (i.e., difficulty in finding foundation board members) and the lack of expertise among board members? If the answer tends toward “yes,” the questions become: First, can the use of AI in foundations’ organization and administration jeopardize the position of the foundation board? Second, can the introduction of AI tools at the board level be a solution to the succession problems and the partial lack of expertise? Ultimately, can AI replace the review and assessment processes for beneficiaries’ applications typically done by experts and board members?

Thanks to new technologies, it is possible to delegate more and more decisions previously made by humans to machines. However, the dynamics of these technologies – including the fact that their decision-making speed is increasing, and they tend to offer higher-quality (i.e., more efficient) decision-making than humans – could potentially run counter to the realization of the foundation’s purpose, which is the collective responsibility of the individuals on the foundation board.

To determine whether that is the case, we first need to ask the following questions: Will automated application reviews differ from human reviews, and how? Moreover, if so, would traditional foundation law allow the software to substitute the physical foundation board? We cannot answer those questions without first clarifying whether machines or computers already have the right to make decisions in this context and whether they should be given such a right. Since a detailed clarification of these questions is beyond the scope of this chapter, we merely share a direction for this thought exercise below.

It currently seems very unlikely that individuals on foundation boards will be replaced by robots, even if this could solve some problems such as succession issues and the lack of expertise. As shown earlier, the discussion is still ongoing, given the lack of technological and organizational maturity. However, traditional foundation law will not be able to escape digitalization and AI. Sooner or later, Swiss foundations will have to start thinking about which decisions can and cannot be handed over to machines to better fulfill their purpose and responsibilities. They will also have to start looking at whether any gaps in AI legislation may need to be filled.

As things stand today, AI tools can be used (responsibly) to support a foundation board with little or no specialist knowledge. The use of AI could significantly help foundation boards assess applications and administer a foundation, ultimately serving the foundation’s purpose.

## **5 Conclusion**

From the above, it stands out that DT has affected and is affecting the Swiss philanthropic sector. As DT is mostly a work in progress in foundations, this sector shows slack relative to other sectors in terms of digital maturity and technology adoption.

We have shown that digital transformation can improve operations such as impact measuring, grantee application processing, and management and governance of a foundation. While these benefits are widely recognized in the philanthropic sector, not all foundations are pursuing digital and AI transformation. The size of the foundation itself, its “digital culture,” and whether the board is forward-looking play a decisive role.

Looking at an organizational journey toward AI maturity, we argue that AI maturity is part of an organization’s DT journey – an organization cannot reap the benefits of AI without first undergoing the stages of digitizing, digitalization, and eventually DT. We have shown that certain decisions can be made more efficiently and effectively with AI as it currently stands. What is less clear, however, are the legal implications related to the role of humans in decision-making, for instance, in assessing grantee applications. Moreover, we laid out to what extent the use of AI is tied to change

management, infrastructure, and additional resources to ensure the value delivery of AI systems while mitigating their risks.

The use of AI systems in foundations is still at a nascent stage, but these systems have tremendous value potential if foundations manage AI risks appropriately. While adoption is still far off for foundations, AI will become more widespread in the future, and foundations will need to adapt to the trend – that is, foundations will need to become AI-ready.

One main reason for the growing and constantly evolving foundation landscape in Switzerland – in addition to factors such as prosperity and political stability – is the liberal legal framework and the constant reflection on how to optimize it. In this sense, AI could provide a unique opportunity for the Swiss philanthropic sector to further evolve and to stand out globally.

## References

- Ada Lovelace Institute. (2022). *Algorithmic Impact Assessment in Healthcare*. Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/project/algorithmic-impact-assessment-healthcare/>
- Aebi-Müller Regina E. (Hrsg.) (2020). Christoph Müller (Hrsg.), Berner Kommentar zum schweizerischen Privatrecht, Die Stiftungen Art. 80-89c ZGB, Bern 2020, 2. Auflage (BK-Riemer, Hans Michael Riemer, ST N 22).
- Aras, A., & Büyükoçkan, G. (2023). Digital transformation journey guidance: A holistic digital maturity model based on a systematic literature review. *Systems*, 11(4), Article 4. <https://doi.org/10.3390/systems11040213>
- Arpe, B., & Kurmann, P. (2019). *Managing Digital Transformation*. Lund University, School of Economics and Management. <https://lup.lub.lu.se/student-papers/search/publication/8989064>
- BCG. (2024). *Digital Maturity Consulting and Strategy* | BCG. Bcg.Com. <https://www.bcg.com/capabilities/digital-technology-data/digital-maturity>
- Buchholz, K. (2023). *Infographic: Threads Shoots Past One Million User Mark at Lightning Speed* (p. 1). <https://www.statista.com/chart/29174/time-to-one-million-users>
- Burke, M. (2024, January 25). Man says AI and facial recognition software falsely ID'd him for robbing Sunglass Hut and he was jailed and assaulted. *NBC News*. <https://www.nbcnews.com/news/us-news/man-says-ai-facial-recognition-software-falsely-idd-robbing-sunglass-h-rcna135627>
- Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zimmel, R. (2023a). *The Economic Potential of Generative AI* (p. 68). McKinsey Digital. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction>
- Chui, M., Yee, L., Hall, B., Singla, A., & Sukharevsky, A. (2023b). *The State of AI in 2023: Generative AI's Breakout Year* | McKinsey (p. 24). McKinsey Digital. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#/>
- Cordery, C. J., Goncharenko, G., Polzer, T., McConville, D., & Belal, A. (2023). NGOs' performance, governance, and accountability in the era of digital transformation. *The British Accounting Review*, 55(5), 101239. <https://doi.org/10.1016/j.bar.2023.101239>
- Dastin, J. (2018, October). Insight—Amazon scraps secret AI recruiting tool that showed bias against women | Reuters. *Reuters.Com*. <https://www.reuters.com/article/idUSKCN1MK0AG/>
- European Parliament (2021). Proposal for a regulation of the European Parliament and of the council; laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>
- European Parliament (2023, June 8). *EU AI Act: First Regulation on Artificial Intelligence*. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- Forth, P., DeLaubier, R., & Charanya, T. (2021). Which sectors perform best in digital transformation? *BCG—Digital Transformation*. <https://www.bcg.com/publications/2021/learning-from-successful-digital-leaders>
- France24. (2017, May 12). France's Renault hit in worldwide 'ransomware' cyber attack. *France24*. <https://www.france24.com/en/20170512-cyberattack-ransomware-renault-worldwide-british-hospitals>
- Geiser, Thomas, (Hrsg.)/Fountoulakis, Christiana (Hrsg.) (2018). Basler Kommentar, Zivilgesetzbuch I, Basel 2018, 6. Auflage (BSK-ZGB I-Harold Grüninger, Art. 80 N 9).

- Goldman Sachs (2023). AI investment forecast to approach \$200 billion globally by 2025. *Goldman Sachs*. <https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html>
- Greenstein, B. (2019). *Investing in AI: Moving Along the Digital Maturity Curve*. Cognizant. <https://thoughtlabgroup.com/wp-content/uploads/2019/12/investing-in-ai-moving-along-the-digital-maturity-curve-codex5050.pdf>
- Jakob, J., Freiburghaus, A., Prof. Dr. Jakob, D., & Prof. Dr. Von Schnurbein, G. (2023). *Der Schweizer Stiftungsreport* (30). <https://www.swissfoundations.ch/publikationen/der-schweizer-stiftungsreport-2023/>
- Johnson, K. (2022). *How Wrongful Arrests Based on AI Derailed 3 Men's Lives* | WIRED. Wired.Com. <https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/>
- Kane, G. C. (2017, April 4). Digital maturity, not digital transformation—MIT Sloan Management Review. <https://sloanreview.mit.edu/article/digital-maturity-not-digital-transformation/>
- Kratz-Ulmer, A., & Favre, D. (2023). *Die Digitalisierung In Klassischen Stiftungen*. [https://www.profonds.org/wp-content/uploads/2023\\_6\\_Die\\_Digitalisierung\\_in\\_klassischen\\_Stiftungen.pdf](https://www.profonds.org/wp-content/uploads/2023_6_Die_Digitalisierung_in_klassischen_Stiftungen.pdf)
- Kratz-Ulmer, A., & Schudel, J. (2020). Digitale transformation in förderstiftungen. *Stiftung & Sponsoring*, 4, 12. <https://doi.org/10.37307/j.2366-2913.2020.04.12>
- Makridis, C. A. (2021). Do data breaches damage reputation? Evidence from 45 companies between 2002 and 2018. *Journal of Cybersecurity*, 7(1), tyab021. <https://doi.org/10.1093/cybsec/tyab021>
- Matias, Y. (2023, October 10). How we're using AI to combat floods, wildfires and extreme heat. *Google*. <https://blog.google/outreach-initiatives/sustainability/google-ai-climate-change-solutions/>
- Microsoft Corporation (2022). *Microsoft Responsible AI Standard v2 General Requirements* (p. 27). Microsoft Corporation. <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmF1?culture=en-us&country=us>
- Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. *Npj Digital Medicine*, 6(1), Article 1. <https://doi.org/10.1038/s41746-023-00858-z>
- NCSC. (2024). *Global Ransomware Threat Expected to Rise with AI, NCSC Warns*. <https://www.ncsc.gov.uk/news/global-ransomware-threat-expected-to-rise-with-ai>
- Sprecher, T. (2017). *Stiftungsrecht in a Nutshell* (2nd ed.). Dike Verlag. <https://www.schulthess.com/buchshop/detail/ISBN-9783038914938/Sprecher-Thomas/Stiftungsrecht-in-a-nutshell>
- Stokel-Walker, C. (2023). *The Generative AI Race Has a Dirty Secret* [Wired.com]. Wired UK. <https://www.wired.co.uk/article/the-generative-ai-search-race-has-a-dirty-secret>
- Tamny, J. (2024). As Nvidia shares soar, we're reminded again of the Fed's irrelevance. *Forbes*. <https://www.forbes.com/sites/johntamny/2024/03/04/as-nvidia-shares-soar-were-reminded-again-of-the-feds-irrelevance/>
- Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, 28(2), 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>
- Vlastelica, R., & Bloomberg. (2024). \$1.72 trillion AI chip giant Nvidia has rocketed in value so fast it's about to pass Amazon as the 4th-most valuable U.S. company. *Fortune*. <https://fortune.com/2024/02/09/how-valuable-nvidia-amazon-billionaire-jensen-huang/>
- Vohra, S., Vasal, A., Roussiere, P., Tanguturi, P., & Guan, L. (2022). *The Art of AI Maturity* (p. 40). Accenture. <https://www.accenture.com/ch-en/insights/artificial-intelligence/ai-maturity-and-transformation>

# TECHNOLOGICAL READINESS OF ASIA'S SOCIAL SECTOR FOR THE ADOPTION AND USE OF ARTIFICIAL INTELLIGENCE

*Kithmina V. Hewage*<sup>1</sup>

## 1 Introduction

As generative AI tools rapidly expand and integrate with other related technology, the role of AI is becoming increasingly important to everyday life. A growing body of literature on the potential benefits and risks of the rise of AI in the social sector is also emerging. These benefits range from enhancing efficiency by automating back-office work, using natural language processing (NLP) to help donors move away from static forms during the grantmaking process, and enabling philanthropic organizations to develop predictive models of impact and impact assessments (Coen, 2023; Pasic & Dean, 2023). However, as with any emerging technology, the use of AI comes with a number of concerns around privacy, workforce redundancy, and the potential to magnify existing biases in the sector, favoring larger organizations over smaller ones due to data discrepancies and the resources available to deploy AI (Why Philanthropy Matters, 2022).

It is important to note that the current commentary around philanthropy and AI is almost exclusively based on characteristics of giving, economic conditions, and organizational structures of the Global North. The Asian context is significantly different from that in North America and Europe given the heterogeneity in the region's economic development levels, the variety of political structures that exist in the region, and the maturity of the philanthropic space and social sector. Therefore, before discussing the opportunities and risks of AI, this chapter attempts to fill a gap in the existing literature by addressing a more fundamental question pertinent to the region – what is the level of technological readiness of Asia's social sector to embrace AI effectively?

To answer this question, the chapter assesses the Asian social sector's (nonprofits and social enterprises) foundational readiness, operational readiness, and transformational readiness using data from the Doing Good Index 2024, published by the Centre for Asian Philanthropy and Society (CAPS). The data reveals three key insights, specifically: (a) the foundational readiness of the Asian social sector is severely lacking due to infrastructural and financial constraints, (b) a lack of skill and expertise is driving many of the impediments to improving the sector's operational readiness, and (c) the social sector in most Asian economies is vulnerable to adopting AI-related technological tools without adequate precautions, opening itself up to significant risks of exploitation and fraud. Fundamentally, the benefits of AI are significantly dulled due to a lag in technological



adoption and access among Asian social delivery organizations (nonprofits and social enterprises). As discussed in Section 2, the Asian philanthropic landscape has several unique characteristics that differ from those in the United States and Europe. As a result, the lag in AI readiness in some parts of the region's social sector will pose several important challenges for the future and increase the need for the region's philanthropists to focus more on capacity building and operational funding.

## **2 Why an Asian lens matters**

### **2.1 Wealth generation in Asia**

Charitable giving is a well-established concept in most Asian economies. However, Asia's philanthropic landscape has evolved rapidly over the past three decades in tandem with the region's economic growth. Since 2006, Asia's total financial wealth has nearly tripled and is now valued at approximately US\$140 trillion. In fact, wealth in Asia (excluding Japan) is expected to surpass that of the United States by 2025 (HSBC, 2022), and there are currently more than 950 billionaires in Asia (Nikkei Asia, 2022). The growth of high-net-worth individuals and families in Asia has also coincided with the expansion of the middle class. As the economic fortunes of many in the region have improved, they see more opportunities to help support social welfare, while at the same time, societal expectations for the more wealthy to give back to their communities have also increased.

According to the Doing Good Index 2022, if Asia matches the United States in terms of philanthropic spending by donating the equivalent of 2% of its gross domestic product (GDP), the region could unlock US\$ 701 billion per year. This is 14 times the amount of net foreign aid flowing to Asia and about 28% of the estimated cost of achieving the Sustainable Development Goals for the region (CAPS, 2022a). Unsurprisingly, in line with their rapid economic growth, philanthropic giving in India and China is one of the biggest drivers of regional philanthropy. In India, private philanthropy is estimated to be about US\$ 13 billion in 2022, an increase of 8% from 2017 (Sheth et al., 2023). In 2017, philanthropic giving in Mainland China was US\$ 23.4 billion and an additional US\$ 500 billion is estimated to flow as philanthropy over the next decade (AVPN, 2018). Indeed, philanthropic giving is likely to exceed these estimates following the COVID-19 pandemic and greater efforts by philanthropists to support pandemic relief. The increased interest in philanthropy is also prompting Asian governments to respond. In Mainland China, for example, President Xi has repeatedly called for "common prosperity" and philanthropy and charity are core elements of the government's proposed "third distribution" to address inequality (Fang, 2022). Meanwhile, Singapore and Hong Kong have recently stepped up efforts to establish themselves as Asia's philanthropic hub with a series of new policies (CAPS, 2023).

### **2.2 Unique characteristics of Asian philanthropy**

Notably, the rise of Asian philanthropy has occurred with its own unique set of characteristics. For instance, Shapiro et al. (2018) identify three distinct characteristics. First, Asian philanthropy is mainly focused on local giving, and cross-border philanthropy is much less common than in the US and Europe. This is because Asia is primarily made up of emerging markets, and philanthropists recognize that there are significant development challenges that need to be addressed in their home countries. As a result, they prioritize giving to these causes. In addition, domestic giving is becoming even more important for Asia foreign funding declines due to the transition of Asian economies to middle-income status, domestic restrictions on foreign funding, and political developments in OECD countries. For example, data from CAPS' Doing Good Index 2024 shows that,

on average, foreign funding accounts for just 16% of the budgets of social delivery organizations in Asia. In 2018, the figure was 22% (CAPS, 2024).

Second, there is a greater overlap between individual and corporate philanthropy in Asia than in North America and Europe. Eighty-five percent of businesses in Asia are family-controlled (Kapoor & Raggett, 2021). Consequently, we see a significant overlap between individual and corporate philanthropy in Asia. A high-net-worth individual or family will seek to support their preferred causes by donating funds from their personal wealth, while also channeling funds through their corporate budgets. This also means that philanthropic giving is often aligned with business interests and based on relationships. This overlap leads to the third characteristic – governments play a much more prominent role in Asian philanthropy. Because philanthropic giving and corporate interests are closely intertwined philanthropists adopt a much more pragmatic perspective and align their giving closely with government policy priorities. In doing so, philanthropy helps build relationships and goodwill with the government, which is seen as important for those who also run companies (Shapiro et al., 2018). Furthermore, it is estimated that there will be a wealth transfer of approximately US\$ 2.5 trillion by 2030 (Wealth-X, 2021). It is therefore reasonable to assume that with the growth of a new generation of Asian philanthropists, technological developments such as the emergence of AI will undoubtedly have an impact on amplifying these regional characteristics.

### **3 AI readiness frameworks**

#### ***3.1 Existing frameworks for assessing AI readiness***

In recent years, a number of frameworks and indices have been introduced to assess AI readiness. These include frameworks that assess government readiness as well as organizational readiness. For example, Deloitte's AI readiness framework for governments highlights technology, data, strategy, people, processes, and ethics as its key pillars. The framework recognizes the non-linear nature of AI and provides for government agencies to adopt policies and processes based on their respective priorities. These could be task-based solutions, process- or problem-focused applications, or holistic approaches that seek to transform the agency (Deloitte, 2020). The Salesforce AI Readiness Index assesses both business and government readiness based on five categories: infrastructure, data, skills, ethics, and integration. Its business readiness indices are evaluated based on factors such as a company's adoption of emerging technologies (e.g., AI, robotics, big data analytics, etc.), business sophistication, and labor market conditions. Meanwhile, government readiness indicators are based on factors such as the level of digital government, human capital, ICT regulations, and government support for investment in emerging technologies (Salesforce, 2023).

For organizations, Holmström suggested assessing readiness along four dimensions: technologies, activities, boundaries, and goals. The framework outlines how technologies play an important role in digital transformation and how the other dimensions align with other aspects of an organization's goals and operations. This includes an organization's current technological portfolio as well as the strategies that will be put in place to use AI in the future in a way that adds value (Holmström, 2022). Similarly, Intel highlights three types of AI readiness in their model: foundational, operational, and transformational. For foundational readiness, the framework suggests assessing an organization's access to factors such as technological infrastructure, cloud resources, data, and software packages. Operational readiness relies on factors such as agile delivery, cybersecurity, and skills and expertise. Finally, transformational readiness relies on strategic leadership, the scope of business opportunity, clarity of business case, and business acceptance (Intel, 2022).

Notably, however, the current literature does not provide a specific framework for assessing the social sector’s readiness of AI. While some of the factors discussed above are relevant to social sector organizations, others, such as the potential for future profitability and market access, are not. The ongoing discourse on AI and the social sector focuses primarily on the potential benefits and risks of using AI rather than assessing the sector’s readiness. It is true that AI can create significant efficiency gains. For example, the social sector can use AI for content generation, prospect identification, action recommendations, business simulations, marketing automation, understanding constituent perceptions, donor journey mapping, and simple task automation (BWF, 2023). However, the sector is highly exposed to privacy concerns and related reputational risks due to its close work with vulnerable groups (Herschander, 2023; Kanter et al., 2023). Moreover, compared to corporations and governments, Social Delivery Organizations (SDOs)<sup>2</sup> have far fewer resources to devote to risk mitigation strategies. Notably, much of this commentary is based on the experience of the social sector in the Global North. This chapter aims to address this gap in the existing literature by building a basic framework for assessing AI readiness in the social sector, selecting relevant elements suggested by others for business and government readiness. The chapter then discusses how the Asian social sector performs against the presented metrics, based on data collected for CAPS’ Doing Good Index 2024. The data was collected by surveying over 2,183 SDOs (nonprofits and social enterprises) in 17 Asian economies.<sup>3</sup>

### 3.2 An AI readiness framework for social delivery organizations (SDOs)

For this analysis, as illustrated in Figure 13.1, the author uses Intel’s three types of AI readiness as a starting point: foundational, operational, and transformational. Foundational readiness refers to the prerequisites for AI, such as the appropriate infrastructure and interfaces (Intel, 2022). For SDOs, this includes access to digital infrastructure, data, and relevant software packages. In addition to access to basic infrastructure such as computers, it is important for an organization to have access to reliable internet connectivity since AI can strain available networks. Similarly, data is the cornerstone of AI and essential to its use. Thus, it is important that an organization has built-in structures and processes that collect and store data in a way that can then be used with AI tools. To access these tools, organizations need the appropriate software.

For operational readiness, the assessment is based on internal skills and expertise, cybersecurity, and access to operational funding. As with most new technologies, lack of skills and expertise is

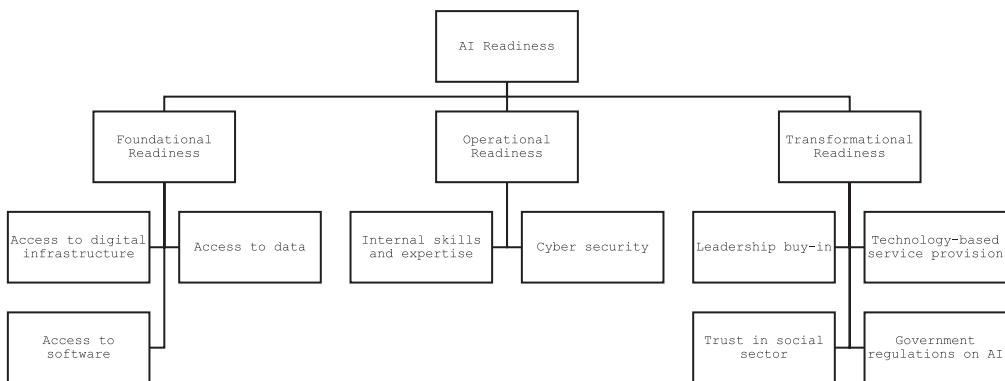


Figure 13.1 AI readiness framework for the social sector.

Source: Author’s rendition of proposed readiness framework.

one of the most significant obstacles for organizations to take full advantage of the benefits. This is no different for AI. For SDOs, this expertise could either be outsourced or developed in-house. In the short term, outsourcing this expertise is likely to be the most cost-effective means, but most organizations prefer to develop in-house capabilities to ensure sustainability. Cybersecurity is another essential facet of building AI readiness and should be prioritized by organizations. Lapses in cybersecurity could lead to potential security risks ranging from the corruption of data input into the AI, tampering with models, or unauthorized access to the resulting insights (Intel, 2022). The importance of cybersecurity is heightened in the social sector as many organizations work with vulnerable populations. Therefore, unauthorized access to sensitive information about such risks could exacerbate these vulnerabilities, and a breach of trust would undermine the sector's ability to support them.

Finally, to assess transformational readiness, it is proposed to evaluate leadership buy-in, the level of technology-based service provision, trust in the social sector, and government regulations around technology and AI. Thus, transformational readiness refers to an organization's ability to maximize the value it can derive from AI (Intel, 2022). This means that an organization's ability to embrace the productivity and operational changes associated with AI will influence its use in the future. As a result, supportive organizational leadership is needed to invest in new technologies and make changes to the organization that accommodate the use of AI.

To assess the level of leadership buy-in, this framework first considers the ease of access to operational funding. Unlike businesses and governments, this is a particularly unique constraint for the social sector. Many donors are reluctant to provide operational support, focusing instead on project-based funding (CAPS, 2022b). Given the rapid evolution of AI, it is important that organizations continue to invest in tools and processes that can adapt to and leverage the benefits of AI. However, such investments are effectively operational finances. Therefore, the ease with which organizations can access operational funding will determine their ability to use AI in their day-to-day activities. Second, the current level of technology use is also important in assessing future readiness. This is important for the social sector, as the communities they serve may not have adopted technology to the same extent as SDOs. Therefore, even if other aspects of AI readiness are in place, the organization will struggle to make the necessary changes. The social sector, particularly in Asia, has fallen victim to a trust deficit due to a number of factors such as political developments and scandals (Shapiro et al., 2018). Given the ongoing uncertainties around AI and its societal implications, whether an organization is trusted to use the technology will affect its readiness for transformation. An organization that is not trusted to use AI will be viewed skeptically and may exacerbate the trust deficit. Finally, as discussed in the previous section, the government plays an important role in the social sector in Asia. Consequently, SDOs are more likely to be risk-averse if a government signals a reluctance to adopt AI or impose restrictions in the future. Under such circumstances, an organization's leadership is unlikely to adopt measures that can effectively leverage AI.

## **4 Data analysis and insights**

Based on the framework presented earlier, this section discusses the state of the Asian social sector in relation to the relevant metrics and thus its AI readiness.<sup>4</sup>

### **4.1 Foundational readiness**

In many ways, foundational readiness is arguably the most important type in the Asian context. As noted above, most Asian economies are low- or middle income. Therefore, the level of communication and technological infrastructure, as well as the level of adoption of digital technology

in socioeconomic affairs, is much lower than in the Global North. Consequently, the associated impediments at the foundational level prevent organizations from moving toward addressing operational and transformational readiness. These challenges are also reflected in the Doing Good Index 2024 data. For example, the top three needs identified by SDOs in the region were hardware, operational software, and training and upskilling. This illustrates the more fundamental challenges that social delivery organizations face around technology, which in turn affects their ability to benefit from developments in AI.

#### *4.1.1 Access to digital infrastructure*

As discussed earlier, Asia is a heterogeneous region with varying levels of economic development and technological absorption. Our data shows that these differences extend to the social sphere. For instance, the use of smartphones is almost taken for granted in the Global North, and it is a basic tool that would facilitate the use of AI in future operations. However, the use of smartphones is much less common in Asia. On average, only 74% of SDOs in Asia reported that 90%–100% of their employees have access to a smartphone. At the individual economy level, Singapore, Taiwan, and Korea had an average of 98% of their workforce using smartphones. The use of smartphones among social sector workers was much lower in India (78%), China (78%), and Pakistan (79%). A similar trend can be seen in access to computers, with 31% of Asian SDOs reporting that current access to computers and/or tablets is insufficient to meet organizational needs (Table 13.1).

In addition to hardware, the Asian social sector also faces challenges in terms of access to reliable and sufficiently fast internet, which is an important element of using AI tools. Of the SDOs surveyed, 85% reported having access to the internet with sufficient speed and service reliability. Notably, even SDOs in middle-income and advanced economies such as Taiwan (77%), Indonesia

*Table 13.1* Staff access to computers and/or tablets

<i>Economy</i>	<i>Sufficient to meet organizational needs (%)</i>	<i>Insufficient to meet organizational needs (%)</i>
Bangladesh	55	45
Cambodia	53	47
China	78	22
Hong Kong	82	18
India	52	48
Indonesia	69	31
Japan	59	41
Korea	74	26
Malaysia	82	18
Nepal	47	53
Pakistan	72	28
Philippines	84	16
Singapore	88	12
Sri Lanka	69	31
Taiwan	79	21
Thailand	90	10
Vietnam	87	13
Total	69	31

*Source:* Doing Good Index 2024, CAPS.

(78%), and the Philippines (80%) reported relatively low levels of reliable and sufficiently fast internet in their offices. On average, 75% of Asian SDOs said the same when asked if their employees have internet with sufficient speed and reliability of service outside the office and at home. The three worst-performing economies in this regard were Nepal (45%), Cambodia (59%), and Bangladesh (61%). This disparity highlights the challenges of using internet-based technologies and tools for service delivery in the field, especially for organizations serving rural communities.

#### *4.1.2 Access to data*

If SDOs are to utilize AI in their work, it is essential that they have access to data sources, especially internal ones, to make their work as effective as possible. To better understand the availability and accessibility of data, we asked SDOs whether they digitally collect and store data on donor records, financial records, client/beneficiary records, and data related to project impact. Here, the most important aspect is whether an organization digitally stores the data, so that AI can use it. On average, 87% of SDOs in Asia store financial records digitally. However, the proportion of SDOs that digitally store donor records (75%), client/beneficiary records (80%), and project impact data (73%) is lower (Table 13.2). This suggests that the Asian social sector has more room to leverage AI if it increases its efforts to digitally store its data. The fact that organizations are least likely to digitally store data related to project impact hinders a potentially significant benefit that can come from the use of AI. This is especially true as AI could fill a significant capacity gap that currently exists among Asian SDOs if it had the necessary data. Only 73% of organizations surveyed in Asia said they measure impact. When those who do not were asked why, the majority cited not knowing how to do it and not having the staff or time to do so as the most common obstacles. Both of these challenges can be addressed by AI, given the right quality and quantity of data.

#### *4.1.3 Access to software*

Even if organizations have adequate levels of hardware and data, it is important that they also have the necessary software to deploy various AI tools. Recently, even basic software has begun to incorporate AI into its functionality. However, to realize their full potential with AI, organizations will need to use more advanced software with tools for data management, visualization, etc. In the technology section of the Doing Good Index 2024 survey, SDOs were asked about their use of different levels of software: basic software (e.g., Microsoft Office), operational software (e.g., CRM, accounting software), and advanced software/digital tools (e.g., statistical analysis software, machine learning tools).

*Table 13.2* Collection and storage of data digitally

	<i>Collect data (%)</i>			<i>Store data (%)</i>		
	<i>Yes</i>	<i>No</i>	<i>N/A</i>	<i>Yes</i>	<i>No</i>	<i>N/A</i>
Donor records	69	16	14	75	11	13
Financial records	79	13	6	86	7	5
Clients/beneficiary records	71	18	9	79	10	9
Data related to the impact of projects	67	19	12	73	13	13

*Source:* Doing Good Index 2024, CAPS.

Basic software was being used almost universally by 97% of the SDOs surveyed. However, there was a significant drop in the use of operational software (49%), and advanced software (35%). SDOs in Hong Kong (50%), Sri Lanka (47%), and Singapore (44%) had the highest use of advanced software. Meanwhile, India (22%), Cambodia (23%), and Nepal (25%) had the lowest use of advanced software.

## **4.2 Operational readiness**

As mentioned earlier, this section discusses whether the Asian social sector is operationally ready to adopt AI tools. To achieve operational readiness, SDOs need the necessary human resources to support the adoption and use of AI, protection from external risks through cybersecurity, and adequate financial resources to continuously invest in AI and related technologies and tools, which requires adequate operational funding.

### *4.2.1 Internal skills and expertise*

Inadequate skills and expertise are other critical factors inhibiting the use of technology in the social sector in Asia and, by extension, the use of AI as well. When SDOs were asked to select their top three overall needs over the next twelve months, most SDOs, unsurprisingly, identified the need for more funding (75%) and more collaboration (60%). Beyond these two considerations, 44% of SDOs identified the need to upskill/reskill staff as a top three need, while 28% identified the need for support with digitalization and digital literacy.

More specifically, when asked about factors that challenge the adoption of digital technologies, SDOs cited the lack of funding (71%), inadequate staff skills (59%), and the lack of awareness of available digital technologies and tools (46%) as the top three barriers. Notably, even SDOs from high-income economies such as Japan (70%), Korea (68%), and Hong Kong (60%) reported a lack of adequate skills to adopt digital technologies. This suggests that addressing the skills gap is an essential component of improving the readiness of Asia's social sector to adopt and use advanced technologies such as AI.

One of the potential factors influencing the skills gap is likely to be related to difficulties in hiring and retaining staff and the lack of operational funding for internal upskilling. On average, 73% of the SDOs surveyed by Doing Good Index 2024 said they found it difficult to recruit staff, while 70% said they found it difficult to retain staff. This is mainly due to the fact that salaries in the social sector tend to be lower than in the corporate sector. In particular, as shown in Figure 13.2, SDOs in upper-middle-income and high-income economies, where the salary gap between the social and corporate sectors is greater, found it more difficult to recruit staff.

### *4.2.2 Cybersecurity*

The second consideration in assessing operational readiness is an organization's cybersecurity. Responses to the Doing Good Index 2024 show that a worryingly high number of SDOs in Asia are unprepared in this regard, and are therefore vulnerable to cyberattacks. Such a high level of vulnerability severely undermines the sector's preparedness to use AI tools.

Thirty percent of SDOs in Asia reported having a cybersecurity or cyber resilience strategy/plan in place. This low number was observed across all economies, regardless of their level of economic development. Only in Taiwan did a majority of SDOs report having a cybersecurity strategy (Figure 13.3).

Technological readiness of Asia's social sector for AI

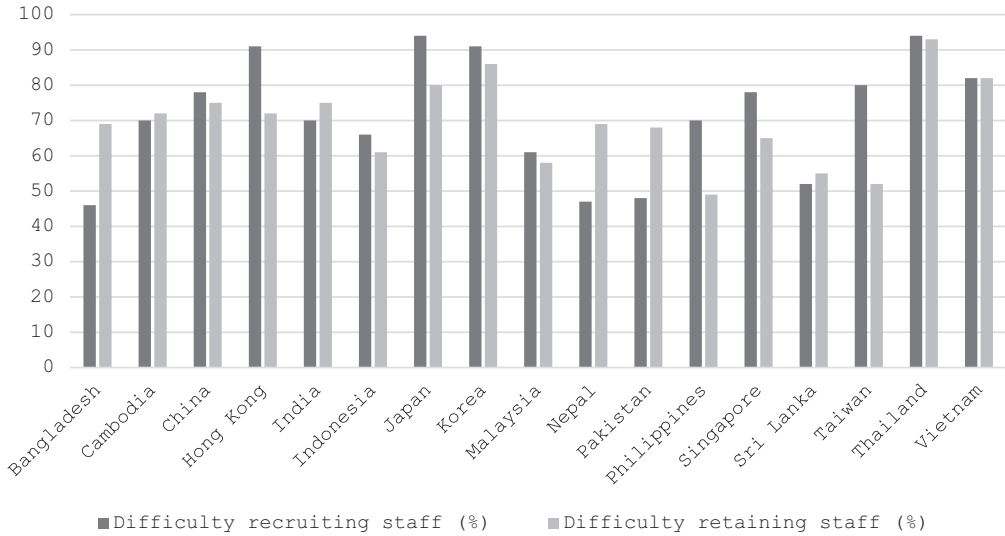


Figure 13.2 Percentage of SDOs with difficulties recruiting and retaining staff.

Source: Doing Good Index 2024, CAPS.

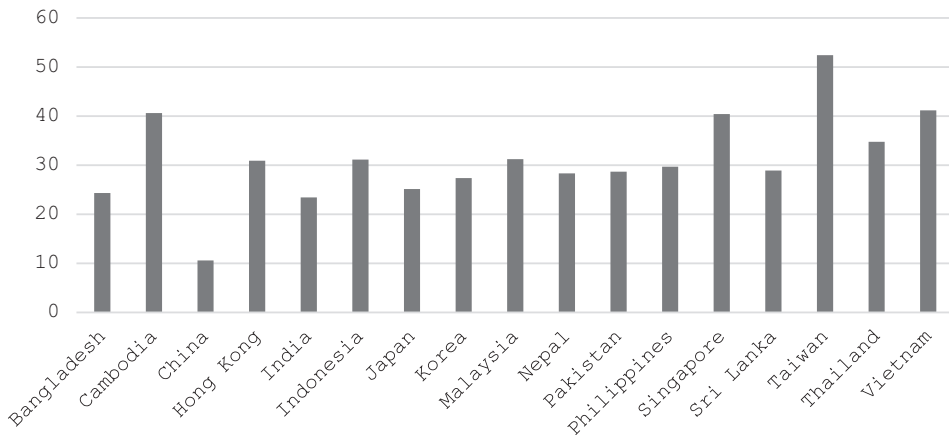


Figure 13.3 Percentage of SDOs with a cybersecurity strategy.

Source: Doing Good Index 2024, CAPS.

Meanwhile, 17% of SDOs said they had experienced a cybersecurity attack in the past two years, while 16% did not know if they had. When asked about mitigation measures against cybersecurity attacks, 63% said they use relevant antivirus/spyware/malware software, 15% of SDOs said they use the services of a third party, and only 29% said they have staff training to protect against cybersecurity attacks.



### **4.3 Transformational readiness**

As detailed in the sections above, Asia’s social sector faces significant barriers to foundational and operational readiness. The perspective on transformational readiness is more mixed, with some positive indicators. However, even if some indicators of transformational readiness are positive, the Asian social sector will find it difficult to build on them due to lags in other relevant metrics, such as the trust deficit, particularly when combined with gaps in foundational and operational readiness.

#### *4.3.1 Leadership buy-in: access to operational funding*

Unrestricted or operational funding is crucial to ensure that organizations have continuous access to financial and non-financial resources (e.g., human resources) that can evolve with technology needs. Especially during COVID-19, studies found that operational funding helps pay for technology and enables SDOs to optimize their operations and service delivery (CAPS, 2022b). According to Asian Charity Services (2021), operational funding refers to critical human resources, infrastructure, and administrative costs that are essential to the running of nonprofit organizations. However, we see a significant gap in access to operational funding in Asia’s social sector. Nearly half (48%) of the SDOs surveyed for the Doing Good Index 2024 reported that most donors are unwilling to give operational funding. Meanwhile, 36% said that securing operational funding is a challenge.

Most notably, 48% of SDOs in Asia reported that their donors do not fund digital technology and IT costs. The highest proportions of SDOs not receiving donor support for digital technology and IT costs were in Korea (68%), Japan (66%), Thailand (64%), and China (62%). Conversely, the proportion of SDOs that received donor support for digital technology and IT costs were in Cambodia (76%), Sri Lanka (75%), and Bangladesh (73%). It is noteworthy that these are economies that are more reliant on foreign funding. While Asian donors remain reticent to funding IT costs, the good news is that there appears to be at least some support among SDO senior leadership to invest in and incorporate technology into their organizations. Only 6% of SDOs surveyed by the Doing Good Index 2024 reported reluctance among senior leadership to support the use of digital technology. Similarly, only 12% reported a reluctance among employees to adopt digital technology.

Without sufficient operational funding, SDOs in Asia will find it difficult to invest in the necessary technological infrastructure, software, and skills needed to effectively utilize new technologies. Therefore, while advocating for more digitization and the use of AI in the social sector, it is incumbent upon philanthropists and other funders to provide the necessary resources.

#### *4.3.2 Technology-based service provision*

Unsurprisingly, as detailed in the Doing Good Index 2022, SDOs incorporated more technology into their daily operations and provided more online services during the pandemic (CAPS, 2022a). This increase in technology use continued even after economies reopened and organizations returned to post-pandemic operations according to the Doing Good Index 2024.

For instance, 67% of SDOs in Asia reported an increase in the incorporation of technology in day-to-day operations, while 68% reported the same in 2022. The economies with the highest proportions of SDOs increasing their use of technology in daily operations were Thailand (96%), Bangladesh (81%), and Nepal (78%). Two years earlier, in 2022, the economies with the highest

*Table 13.3* Proportion of Asian SDOs that increased their use of digital technology and online tools

	<i>Incorporation of technology into day-to-day operations</i>	<i>Services offered online</i>	<i>Using online platforms to collaborate</i>	<i>Hosting online events</i>	<i>Use of social media to promote or disseminate work</i>	<i>Online fundraising</i>
Asian SDOs (%)	67	56	65	60	72	35

*Source:* Doing Good Index 2024, CAPS.

increase in technology use were Hong Kong (86%), Indonesia (85%), and Singapore (85%). It is likely that, following COVID-19, SDOs from more advanced economies were able to incorporate technology faster between 2020 and 2022, whereas SDOs from low- and lower-middle-income economies caught up in the last two years.

Meanwhile, 56% of SDOs in Asia reported an increase in services offered online over the past two years. This figure was 55% in the 2022 index. The economies with the highest proportion of SDOs reporting an increase in 2024 were Thailand (67%), Sri Lanka (64%), and Hong Kong (62%). The use of online platforms for other functions has also increased. For example, as shown in Table 13.3, 65% of SDOs reported a rise in the use of online platforms to collaborate with others, 60% saw an increase in the hosting of online events, and 72% saw an increase in the use of social media to promote or disseminate their work.

Comparatively, only 35% of SDOs reported an increase in online fundraising over the past two years. The most popular digital tools used for fundraising were the organization's website (54%), social media channels (52%), and direct emails to donors (44%). Other digital tools used for fundraising included third-party online fundraising platforms (29%), direct text messages to donors (27%), and email newsletters (24%). Nearly one-fifth of the SDOs surveyed (19%) said they do not use any digital tools for fundraising. The evidence suggests that the use of digital tools for fundraising is still low. Lack of digital infrastructure, organizational capacity constraints, and regulatory burdens are likely to be the main obstacles to the growth of online fundraising initiatives. The adoption of AI is expected to help address organizational capacity constraints, particularly those related to skill gaps among staff, such as inexperience in writing fundraising pitches and language barriers among others.

The high level of increase in technology adoption and use among SDOs in Asia is a positive sign. However, it will be essential to address foundational and operational readiness barriers if the Asian social sector is to sustain this increase and take technology use to the next level to reap the benefits of AI tools.

### *4.3.3 Trust in the social sector*

The level of trust is important in assessing the readiness of Asia's social sector for transformation, as social delivery organizations work with marginalized and vulnerable groups who could be exploited if their data is not protected. Furthermore, recent conversations around generative AI and its applications have also led to a growing social skepticism about AI. Therefore, if SDOs are to fully utilize AI tools in their work, they must do so by building their trust with relevant stakeholders – society, corporates, and government.

Evidence shows that the scope of trust in the social sector is variable and sensitive to political developments and scandals. Forty-three percent of SDOs said that they felt trusted by society

while 51% said that they felt only somewhat trusted. The level of trust in the social sector by corporates is slightly lower, with 40% of SDOs observing that they are trusted while 54% said that they are somewhat trusted. The lowest levels of trust are observed with governments. Only 36% of SDOs felt trusted by their respective government, while 57% felt somewhat trusted. Although the number of SDOs that said they were not trusted by either society, corporates, or governments is very low, the larger proportion that felt they were only somewhat trusted indicates the sensitivities surrounding the sector in the region.

The 2024 Doing Good Index asked SDOs about the level of trust in the social sector from society, the corporate sector, and the government. Notably, the highest proportions of SDOs that felt that they were not trusted by society were observed in China (17%), Japan (10%), and India (6%). Meanwhile, 14% of SDOs in China, 13% in Japan, and 10% in Sri Lanka said that they were not trusted by corporates. When asked whether the social sector is generally trusted by the government, the highest proportions of SDOs that said the government doesn't trust them were in Korea (21%), Sri Lanka (14%), India (14%), and China (11%). The existence of a trust deficit between the social sector and the government is particularly problematic, as it could lead to more restrictive regulations and greater societal distrust of the use of AI in operations.

#### *4.3.4 Government regulation of AI*

Government regulations and policies on the use of AI in Asia, as in most other regions of the world, are at an early stage. It is too early to evaluate these policies to determine how they will affect the readiness of the social sector for transformation. Therefore, this section provides an outline of existing and anticipated policies, regulations, and guidelines related to AI in Asia.

The Association of Southeast Asian Nations (ASEAN) has announced its intention to draw up a set of governance and ethics guidelines for AI. The guidelines are expected to propose “safeguards” to mitigate identified risks and spur member nations to develop their national AI-related policies (Reuters, 2023). Six ASEAN members, namely Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Vietnam, have already formulated or started to formulate an AI strategy (Leng, 2023). Thailand, for example, has already launched its National Strategy and Action Plan on AI in 2022. A key element of this is to pave the way for the enforcement of AI laws and regulations, while also improving AI-related education and human resource capabilities (AI Thailand, 2023). In the meantime, Singapore has indicated that it does not yet intend to regulate AI, even though it has introduced the world's first AI testing toolkit called “AI Verify” to encourage the responsible use of AI in the corporate sector (Chiang, 2023).

Elsewhere in Asia, China has introduced the world's first legislation specifically targeting generative AI. The new regulations impose restrictions on companies that provide generative AI services to consumers, both in terms of the training data used and the outputs produced (Roberts & Hine, 2023). India has indicated that it has no plans to regulate AI, but has begun efforts to standardize responsible AI and promote best practices. Japan has proposed a “soft approach” to regulating AI, introducing a set of guidelines that direct large companies to disclose their generative AI services (Nikkei Asia, 2023).

Given that AI is such a new concept, Asian policy has so far focused solely on developing the technology and mitigating potential risks to consumers. In the future, however, more comprehensive legislation is likely to be introduced. This will affect the ability of an SDO to use the technology.

## **5 Recommendations for funders to support AI readiness in Asia**

Based on the findings above, three key trends can be observed regarding the readiness of the social sector in Asia to adopt and leverage AI. First, improving access to digital technology tools and infrastructure will be the most important challenge that requires attention. Second, the Asian social sector suffers from a significant skills gap that hinders its ability to use digital tools. Third, the Asian social sector is highly vulnerable to cyberattacks and other related threats, which is particularly relevant for the use of AI. This section briefly outlines some measures that governments, SDOs, and funders could take to address some of these challenges and support the social sector's readiness to use AI.

### ***5.1 Increasing project and operational funding for ICT investments***

The challenges of infrastructure and access to technology in Asia's social sector are largely related to more structural issues associated with the country's economy. As most economies in the region are still low- or middle income, telecommunications and Information and Communication Technology (ICT) infrastructure remain underdeveloped. As a result, SDOs face a double challenge. On the one hand, their own access to digital infrastructure and tools is limited due to budgetary constraints and poor infrastructure in general. For instance, if an SDO does not have access to reliable internet in the field, it will not be able to use internet-based tools to provide services more efficiently. This is indispensable for the adoption of AI, since a significant portion of AI tools still rely heavily on internet access.

On the other hand, the extent to which an SDO can adopt technology is also limited by the level of technology adoption among its beneficiaries. This is the case in certain areas (e.g., rural regions in Asian economies), where access to smartphones, the internet, or even more basic services such as electricity is limited. Thus, no matter how many technological tools are adopted by SDOs in their offices, beneficiaries will not be able to access these services. Under such circumstances, the return on investment in advanced technological tools such as AI will be low, and SDOs are likely to redirect these resources to more basic tools that support their service delivery. To effectively support AI adoption in Asia's social sector, governments, SDOs, philanthropists, and other funders must be aware of these two challenges.

However, prior to the adoption of AI, the access issues faced by the social sector are much more fundamental. To address these challenges governments need to invest more in technology infrastructure, especially to ensure access for marginalized and vulnerable communities. In addition, when designing interventions, SDOs and funders should consider the catalytic impact that investments in improving access to technology tools could have on these communities. Meanwhile, philanthropists, foundations, and other funders should focus more on funding SDO's technology and IT costs. As mentioned above, IT costs are typically not covered by project-specific budgets, but contribute significantly to improving project effectiveness in the long term.

Better investments in improving the infrastructure and technology access of SDOs will have knock-on effects on other factors influencing an organization's foundational and operational readiness as well. For instance, with better access to computers and software, SDOs will be better placed to improve data collection about their work and beneficiaries. This will enable them to use AI tools to better assess the impact of their work and design more effective future interventions as well.

### ***5.2 Make use of ample opportunities for in-kind donations***

Philanthropic spending can come in a variety of forms that are financial and in-kind. The challenges that Asian SDOs face with their technological readiness to embrace AI provide a number of opportunities for in-kind donations. For example, one of the ways that corporates can support SDOs with their skills gap is by encouraging staff members to volunteer their expertise (CAPS, 2022a). IT professionals in companies can provide an immense service to small SDOs who do not have the resources to hire dedicated IT staff and help improve in-house capacity as well as by conducting training sessions about new tools available. Similarly, software companies can support SDOs by providing operational and advanced software for free or at least significantly discounted rates. This would be particularly important for cybersecurity-related software since the use of cybersecurity measures is at an alarmingly low level among SDOs.

Initiatives by companies to donate second-hand computers/laptops and mobile phones to SDOs with poor access to these basic tools will also go a long way. Given the low level of access to even the most basic tools in several Asian economies, even such simple initiatives will significantly improve the social sector's technological readiness. As discussed earlier, there is a high degree of overlap between individual and corporate philanthropy. Therefore, Asian philanthropists will be at a particularly advantageous position to leverage their cross-sectoral resources to support the social sector with such in-kind donations.

### ***5.3 Supporting SDOs in developing AI guidelines***

As mentioned above, SDOs work with vulnerable communities, and data related to such communities is at a higher risk of exploitation than others. At the same time, more and more SDOs are beginning to provide services online, use online tools, and adopt newer technologies, including AI. Therefore, data protection and data use will be a key development in the social sector over the next decade. As such, it is important that the social sector proactively begins to develop adequate guidelines for the collection, storage, and use of data, as guidelines for the automation of the services it provides through AI.

In the Global North, we are already seeing the emergence of foundations and advocacy organizations, such as the Responsible Artificial Intelligence Institute (RAII), focused on promoting the responsible use of AI. However, such resources are less available to the social sector in the Global South, including most of Asia. Therefore, the Asian philanthropic sector could play an important role in supporting the transfer of such resources to the social sector and enabling SDOs to develop appropriate safeguards and guidelines. In doing so, the Asian social sector will be future-proofing itself rather than trying to catch up.

## **6 Conclusion**

As the world moves rapidly toward the adoption of AI, it is important to assess the technological readiness of the social sector to adopt these new tools. This chapter is not intended to be an in-depth diagnosis of the AI readiness of the social sector in each Asian economy. Instead, it provides an overview of the use of technology in the Asian social sector and suggests a framework for assessing its AI readiness. In doing so, the chapter has identified some common trends that warrant further attention and action by funders. The evidence suggests that the region's social sector is lagging behind on several metrics needed to be suitably prepared to adopt and use AI in its operations. These issues are particularly prevalent in terms of foundational readiness and operational readiness.

SDOs in developing economies, in particular, face challenges in accessing the basic infrastructure and equipment needed to deploy AI. Moreover, SDOs in the region also face severe human resource constraints, as they are unable to hire staff with the appropriate skills and expertise. In addition, low levels of operational funding have limited the opportunities for these organizations to invest in the relevant technologies and tools. Not only does this limit the scope of services that SDOs can provide, but the data also shows that SDOs are highly vulnerable to cybersecurity threats.

As it stands, the low level of technological readiness among Asian SDOs will severely dampen the potential benefits that AI can bring to the philanthropic space in the region. At the same time, a lack of readiness can exacerbate the risks if the technology is deployed without adequate safeguards. Therefore, the chapter proposes three specific recommendations that funders can focus on when trying to improve the Asian social sector's AI readiness. First, more funders should commit to providing operational funding and supporting investments in IT and digital technology. This should be done at both the SDO and recipient levels, as lack of access to technology prevents marginalized communities from being served. Second, the IT sector offers many opportunities for corporations, foundations, and individual philanthropists to contribute more to the social sector through in-kind donations and volunteer programs. Volunteer time will be especially useful in addressing the skills gap faced by these SDOs. Third and finally, foundations in the region should place more emphasis on helping SDOs develop and design guidelines for the use of AI and data protection. These recommendations are not mutually exclusive, and the unique characteristics of Asian philanthropy make them achievable. Such actions can transform the Asian social sector and improve its readiness to adopt and use AI for doing good.

### Notes

- 1 The findings presented in this chapter are based on data collected for the Centre for Asian Philanthropy and Society's (CAPS) Doing Good Index 2024.
- 2 The term "social delivery organization" (SDO) is used to refer to entities engaged in providing a product or service that addresses a societal need. It covers organizations ranging from traditional nonprofits to nonprofits with income streams, to social enterprises, and operating foundations.
- 3 Bangladesh, Cambodia, China, Hong Kong, India, Indonesia, Japan, Korea, Malaysia, Nepal, Pakistan, Philippines, Singapore, Sri Lanka, Taiwan, Thailand, Vietnam.
- 4 Unless cited otherwise, insights for this section are based on data collected for the Doing Good Index 2024.

### References

- AI Thailand. (2023, April 12). *AI Thailand* | แผนปฏิบัติการด้านปัญญาประดิษฐ์แห่งชาติเพื่อการพัฒนาประเทศไทย (พ.ศ. 2565 – 2570)—*AI Thailand*. <https://ai.in.th/about-ai-thailand/>
- Amir Pasic & Eugene R. Tempel Dean. (2023, May 5). *AI and Philanthropy: Tools and Transformations*. Lilly Family School of Philanthropy. <https://philanthropy.indianapolis.iu.edu/news-events/news/newsletter/philanthropy-matters/2023-issues/may-2023.html>
- Asian Charity Services. (2021, April 7). *The Significance of Unrestricted Funding for Sustaining an NGO – Part I*. <https://www.asiancharityservices.org/the-significance-of-unrestricted-funding-for-sustaining-an-ngo-part-1/>
- AVPN. (2018). *Philanthropy in China*. Asian Venture Philanthropy Network. <https://www.rockefellerfoundation.org/wp-content/uploads/Philanthropy-in-China-Web-Version-April-5-2019-FINAL.pdf>
- BWF. (2023, November 17). *AI for Nonprofits: How to Leverage Machine Learning for Good*. BWF. <https://www.bwf.com/ai-for-nonprofits/>
- CAPS. (2022a). *Doing Good Index 2022*. [https://caps.org/work/our-research\\_doing-good-index-2022](https://caps.org/work/our-research_doing-good-index-2022)

- CAPS. (2022b). *Operational Funding: Why It Matters Now More Than Ever*. Centre for Asian Philanthropy and Society. [https://caps.org/work/our-research\\_operational-funding](https://caps.org/work/our-research_operational-funding)
- CAPS. (2023). *Hong Kong as a Philanthropy Hub*. [https://caps.org/work/our-research\\_hong-kong-as-a-philanthropy-hub](https://caps.org/work/our-research_hong-kong-as-a-philanthropy-hub)
- CAPS. (2024). *Doing Good Index 2024*. Centre for Asian Philanthropy and Society. Reuters (2023, June 16). *Southeast Asia to Set “Guardrails” on AI with New Governance Code: Sources*. Reuters. <https://www.reuters.com/technology/southeast-asia-set-guardrails-ai-with-new-governance-code-sources-2023-06-16/>
- Chiang, S. (2023, June 19). Singapore Is Not Looking to Regulate A.I. Just Yet, Says the City-State’s Authority. *CNBC*. <https://www.cnbc.com/2023/06/19/singapore-is-not-looking-to-regulate-ai-just-yet-says-the-city-state.html>
- Coen, N. (2023, July 28). *AI in the Charitable and Philanthropic Sectors: A Risk or Opportunity?* IFC Review. <https://www.ifcreview.com/articles/2023/july/ai-in-the-charitable-and-philanthropic-sectors-a-risk-or-opportunity/>
- Deloitte (2020). *AI Readiness for Government* (Deloitte Insights). Deloitte Center for Government Insights. [https://www2.deloitte.com/content/dam/insights/us/articles/5121\\_ai-readiness-for-government/DI\\_AI-readiness-for-government.pdf](https://www2.deloitte.com/content/dam/insights/us/articles/5121_ai-readiness-for-government/DI_AI-readiness-for-government.pdf)
- Fang, Y. (2022). Analysis of the Present Situation of China’s Third Distribution and Suggestions for Its Development. *Business and Management Research*, 504–509.
- Herschander, S. (2023, September 25). 7 Questions Nonprofits Have About A.I., Answered. *The Chronicle of Philanthropy*. <https://www.philanthropy.com/article/7-questions-nonprofits-have-about-a-i-answered>
- Holmström, J. (2022). From AI to Digital Transformation: The AI Readiness Framework. *Business Horizons*, 65(3), 329–339. <https://doi.org/10.1016/j.bushor.2021.03.006>
- HSBC. (2022, September 1). *The Rise of Asia’s Wealth Will Boost Its Resilience | Insight | HSBC Holdings Plc*. HSBC. <https://www.hsbc.com/news-and-views/views/hsbc-views/the-rise-of-asias-wealth-will-boost-its-resilience>
- Intel (2022). *The AI Readiness Model*. Intel. <https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/ai-readiness-model-whitepaper.pdf>
- Kanter, B., Fine, A., & Deng, P. (2023, September 7). *8 Steps Nonprofits Can Take to Adopt AI Responsibly (SSIR)*. [https://ssir.org/articles/entry/8\\_steps\\_nonprofits\\_can\\_take\\_to\\_adopt\\_ai\\_responsibly](https://ssir.org/articles/entry/8_steps_nonprofits_can_take_to_adopt_ai_responsibly)
- Kapoor, S. & Raggett, C. (2021, September 30). *Asian Family Businesses*. Russell Reynolds. <https://www.russellreynolds.com/en/insights/articles/asian-family-businesses>
- Leng, K. F. S. (2023). *ASEAN’s New Dilemma: Managing the Artificial Intelligence (AI) Space*. ISEAS Yusof Ishak Institute.
- NikkeiAsia(2022, October4). AsiaHasOver950Billionaires,OutnumberingallOtherRegions. *NikkeiAsia*. <https://asia.nikkei.com/Spotlight/Datawatch/Asia-has-over-950-billionaires-outnumbering-all-other-regions>
- Nikkei Asia (2023, August 5). Japan to Propose Corporate AI Disclosure Guidelines for G7. *Nikkei Asia*. <https://asia.nikkei.com/Business/Technology/Japan-to-propose-corporate-AI-disclosure-guidelines-for-G7>
- Roberts, H. & Hine, E. (2023, September 27). The Future of AI Policy in China. *East Asia Forum*. <https://www.eastasiaforum.org/2023/09/27/the-future-of-ai-policy-in-china/>
- Salesforce (2023). *Asia Pacific AI Readiness Index 2023*. Salesforce. [https://www.salesforce.com/content/dam/web/en\\_sg/www/documents/pdf/salesforce\\_ai\\_readiness\\_index\\_2023.pdf](https://www.salesforce.com/content/dam/web/en_sg/www/documents/pdf/salesforce_ai_readiness_index_2023.pdf)
- Shapiro, R. A., Jang, H., & Mirchandani, M. (2018). *Pragmatic Philanthropy: Asian Charity Explained* (1st ed. 2018). Springer Singapore: Imprint: Palgrave Macmillan. <https://doi.org/10.1007/978-981-10-7119-5>
- Sheth, A., Batabyal, J., Nundy, N., Misra, A., & Pal, P. (2023). *India Philanthropy Report 2023*. Bain Capital. <https://www.bain.com/insights/india-philanthropy-report-2023/>
- Wealth-X. (2021). *Preservation and Success: Family Wealth Transfer 2021*. <https://go.wealthx.com/download-preservation-and-succession-family-wealth-transfer-2021>
- Why Philanthropy Matters (2022, June 15). *Philanthropy & A.I. – Why Philanthropy Matters*. <https://whyphilanthropymatters.com/guide/the-future-of-philanthropy/philanthropy-a-i/>

# DIGITAL PHILANTHROPY IN CHINA

## How internet fundraising platforms and artificial intelligence are transforming non-profit governance

*Bertram Lang*

### 1 Introduction

The rise of internet philanthropy has arguably transformed the non-profit sector in China more rapidly and more fundamentally than in any other country in the world. For one, the development and uptake of smartphone-based applications have been exceptionally swift. Secondly, the Chinese offline philanthropy field was at an early and fragile stage of development when the first digital platforms for project promotion and fundraising started gaining traction in the early 2010s. The growing foundation sector was plagued by the absence of a reliable legal framework (Sidel, 2014), widespread opacity surrounding foundations' revenues and expenses (China Foundation Center, 2013), and, consequently, a flurry of misappropriation and embezzlement scandals that eroded public trust and kept charitable donations at a dismal level in international comparison (Zhang, 2015). For grassroots non-profit organizations, this dire situation was compounded with an intensified political crackdown on civil society after Xi Jinping came to power in 2012/2013, which also led to a drain of international philanthropic funding that had previously played a crucial role in sustaining local NGO development in China (Lang, 2018; Sidel, 2019).

The politically promoted catch-up development in the Information and Communication Technology (ICT) sector combined with a tech-savvy middle class that quickly adopted mobile payment systems provided a fertile ground for a non-profit sector in desperate need of resources. Vowing to overcome this “crisis of trust” in philanthropy, China's corporate ICT giants like Tencent, Alibaba, Baidu, or Bytedance have stepped in with promises of offering game-changing technological solutions to make philanthropy more transparent and accessible to all and to restore “social trust” (珍惜社会信任) in organized philanthropy (Sohu Technology, 2016).

With their corporate foundations, these Chinese tech firms, which are also at the forefront of developing artificial intelligence-based commercial applications, not only engaged in various philanthropic projects but also set up online fundraising platforms where fundraising NGOs and individuals can advertise their projects and obtain donations from the public. For fundraising NGOs and foundations,<sup>1</sup> the internet has also rapidly evolved into a vital space, offering a direct and low-cost way of engaging with the public (Gao, 2016; Qu, 2020) and, most importantly, exploring new fundraising channels (Tsai & Wang, 2019; Zhang, Xiang, & Hao, 2019). Thus, a growing number of NGOs have indeed embraced these new forms of internet philanthropy (Qu, 2020; Zhou & Pan,



2016, p. 2433), if only out of necessity in the context of dwindling funding opportunities from overseas donors (Holbig & Lang, 2022; Lai & Spires, 2021).

The unparalleled boom in give-as-you-go donations via smartphone apps has transformed China into a leading market in online philanthropy (Tsai & Wang, 2019, p. 973). The flurry of technological innovations and citizens' propensity to use mobile payment apps to donate small amounts of money have been cast as unique strengths of China's philanthropy sector (Liang, 2018). Yet, the quantitative expansion of online fundraising alone<sup>2</sup> tells us little about the *qualitative* transformation of non-profit governance as a consequence. Therefore, this contribution addresses the question of how the rise of Internet Fundraising Platforms and the rapidly growing application of AI technologies are transforming China's non-profit sector and what this means for philanthropic organizations and donors alike.

To so do, it draws on a wealth of data collected for a broader study of transnational dynamics in the Chinese philanthropy sector. It is informed by Chinese media analysis, interviews with sectoral experts, foundation managers, and staff members of fundraising NGOs (see Table of interviews) as well as the analysis of project data scraped from the largest fundraising platform, *Tencent Charity* (腾讯公益, *tengxun gongyi*) in June 2022. The dataset contains information on 109,260 projects, which collected RMB 12.6 billion from 384.3 million donors in total.

Section 2 will first explain the political-economic context that has favored the digital philanthropy boom in China under strict and intensifying political control, particularly since 2015. Considering the strong role of state intervention in and control over the economy and civil society in China's authoritarian system, it also looks at the state's efforts to regulate digital philanthropy and shows how they have consolidated a corporate oligopoly in the field. Consequently, Section 3 goes on to examine these corporate ICT firms' visions for digital, AI-enabled philanthropy as well as the realities behind these visions, with a particular focus on Tencent's "99 Giving Day," an online charity event that has become an annual focal point for the entire non-profit sector. Using data from the *Tencent Charity* platform, it also highlights the accentuated market concentration in online fundraising and explains the drop in individual fundraising after 2016. Section 4 then looks more specifically at AI applications to the philanthropy sector, driven by the same tech companies operating fundraising platforms and being widely promoted as catch-all solutions for non-profit governance and social policy challenges. Finally, the global ramifications of China's digital philanthropy boom will be discussed in light of these findings, notably with a view to potential applications of Chinese innovations in other contexts.

## 2 The political economy of China's quest for "smart philanthropy"

The surge in Chinese internet philanthropy has been primarily driven by the same corporate actors that have been at the forefront of the country's networked digitization at large. Most providers of China's predominant online fundraising platforms – which act as the crucial intermediaries between fundraising non-profits and potential donors – are corporate foundations closely tied to the same companies that have successfully popularized social media (notably Tencent and ByteDance), online shopping (Alibaba's Taobao, Jingdong, or Pinduoduo), or e-payments (Alibaba's Ant Group) and are now pioneering the frenetic artificial intelligence (AI) race.<sup>3</sup> Any understanding of philanthropy digitalization thus starts from the intricate and ambiguous relationship between these stock-listed companies and the Chinese Communist Party (CCP)-led state, which has been constantly searching for a balance between industry promotion and total political control – both in the economically preponderant ICT sector and in the non-profit sector, which is of critical importance to public welfare and social stability.

## **2.1 State and corporate power in China's platform economy**

In the digital platform economy, user data is the critical currency (Doorn & Badger, 2020; Liu, 2021). Market power and profits are accumulated by maximizing user stickiness on own applications (Su & Flew, 2021, p. 71), which is most effectively achieved through platform convergence, integration, and consolidation to fully leverage network effects (Calvano & Polo, 2021; Cutolo & Kenney, 2021). While American ICT giants have taken advantage of the laxness or slowness of digital-era regulation and the absence of adequate competition law to establish near-global oligopolies, their Chinese counterparts have benefited from state protection against this kind of foreign competition (Chan & Kwok, 2022, pp. 135–136). Demonstrating their usefulness as private for-profit actors for the CCP's socio-economic development goals, they have relied on informal alliances with state actors to effectively increase their reach into all spheres of social life (Su & Flew, 2021).<sup>4</sup> But counter to conventional wisdom in the West, they have not merely copied US applications; instead, the fierce competition within the booming Chinese market has led to a proliferation of innovative applications that have been quickly adopted by China's over 1 billion internet users. In particular, with its vast array of integrated functionalities, Tencent's super-app WeChat has long surpassed the platform convergence level reached by any of its Western competitors (Plantin & Seta, 2019). Its cross-sectoral platform power today is only rivaled by its main domestic competitor, Alibaba, whose digital empire dominates the online retail and e-payment markets (Liu, 2021, pp. 48–50).

## **2.2 “Internet+ philanthropy”: platform power extending to the non-profit sector**

The Chinese government's “internet+” (互联网+) strategy, a national policy initiative formalized in 2015 (State Council of the PRC, 2015), played a central role in this regard. The basic tenet of “internet+” is to integrate the internet with all aspects of the economy and society. This includes promoting innovation and entrepreneurship within the internet sector, using the internet to improve the efficiency and effectiveness of public services, as well as using digital technology to create new opportunities for businesses and individuals' socio-economic participation.

The strategy's publication signaled government support and spawned frantic development of pioneering AI solutions to digitize socio-economic governance across the board. Provincial governments have entered numerous cooperation agreements with leading technology companies in order to gain an edge in their fierce competition to develop the most innovative, “smartest” forms of governance (Zeng, 2020, pp. 1446–1454). While the state-supported digital transformation has boosted the growth of China's leading ICT companies, the logic of an ever-expanding platform economy also put them under competitive pressure to expand their reach into all strides of society.

The logic behind “internet+” also boosted digital solutionist thinking in the social welfare sector. Technical solutions have been floated and anticipated for virtually any social governance problem, from rural education of disadvantaged children and elderly care to fighting environmental degradation to more efficient allocation in urban housing. This openness to digital solutions to address the sector's severe problems created opportunities for ICT companies to offer their services to public welfare agencies as well as social organizations. It is thus consequential that Tencent, Alibaba, and their competitors have also tapped into the expanding online fundraising market and jumped on the “internet+” bandwagon by rebranding their own charitable ventures as “internet+ philanthropy” 互联网+公益 or as building a “smart philanthropy system” (智慧公益体系, *zhihui gongyi tixi*) (Gongyi Zhongguo, 2017).

### 2.3 The 2016 Charity Law: the paradox of legal recognition and regulatory containment

After years of expert consultation and bureaucratic wrangles, the Charity Law (慈善法, *cishanfa*) was passed by the National People’s Congress in March 2016 as the first-ever formal law to govern Chinese philanthropy (Howell, 2019, pp. 60–76; Lang, 2018, pp. 159–165). In the context of an authoritarian crackdown on autonomous social actors in civil society, the law was greeted in the sector as a sign of official recognition of philanthropy’s important role in society, even under stringent political control (Spires, 2020, p. 572). Following the state’s agenda of promoting the “healthy and orderly development” (People’s Daily, 2012; State Council of the PRC, 2014) of the non-profit sector, the Charity Law also set out strict rules for online donations, which were justified as “protecting internet philanthropy” from fraud and misappropriation risks (Philanthropy Times, 2016). In effect, the law not only accepted but legally consolidated and reaffirmed the dominance of corporate oligopolies by setting higher formal barriers to online fundraising in the name of transparency and the fight against online fraud (which indeed is a major risk in China’s online environment). To rein in the uncontrolled spread of online advertisements for charitable donations of all kinds, Art. 23/3 CL notably mandates information disclosure on “unified or designated charity information platforms” 统一或者指定的慈善信息平台 for all online fundraising activities. In other words: No organization or individual is allowed to raise funds online (even via their own websites) unless it has registered those activities on one of the “designated” (i.e., government-approved) platforms. This requirement puts those platforms in a privileged intermediary position between the public and “organizations with public fundraising capacity” (具有公开募捐资格的慈善组织, Art. 22) since the latter are only allowed to launch fundraising campaigns after having gone through their respective accreditation procedures. In the first year, the Ministry of Civil Affairs (MoCA) only accredited 20 organizations as designated “Internet Fundraising Information Platforms” (IFIP) (MoCA, 2018). Another ten platforms were added in November 2021 but remain negligible in quantitative terms (MoCA, 2021). This means all relevant platforms are controlled by China’s major internet companies or their corporate foundations (see Figure 14.1).

Moreover, the vast majority of donations are concentrated on the three most powerful platforms, *Tencent Charity* as well as Alibaba’s *Ant Charity* and *Taobao Charity*, with Tencent alone

Funds Raised on Online Fundraising Platforms in 2017

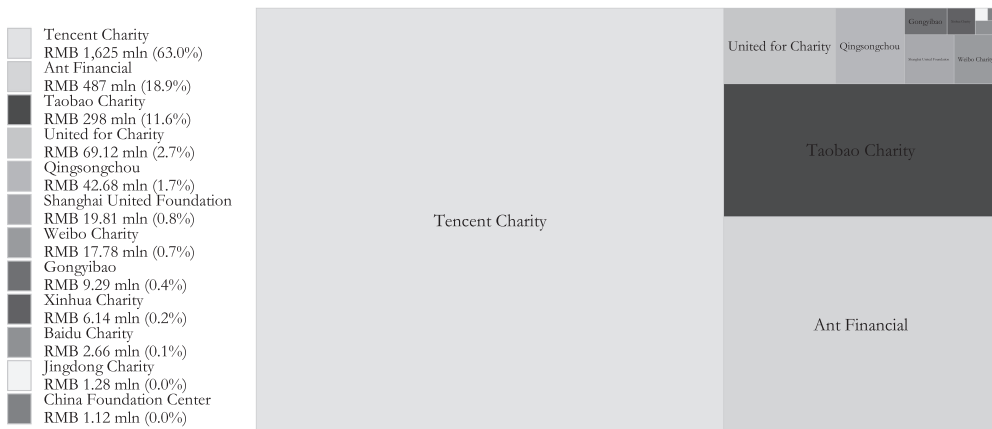


Figure 14.1 Oligopolistic structure of Chinese online fundraising.

controlling almost two-thirds of the market. This oligopolistic nature of online fundraising<sup>5</sup> has been the subject of public criticism (Jian, 2018; Zhang, 2020), yet the government has not made any visible effort to change this situation in the years since.

In contrast with the sluggish, patchy implementation of most other aspects of the Charity Law (Sidel, 2022), detailed technical regulations were issued in 2017 to specify the obligations of platforms in terms of service provision, data transparency, and security, as well as supervision and complaints management. *Inter alia*, platform providers must implement a “social reporting” (社会举报) tool, respond to users’ complaints within five workdays, and, if complaints are established to be justified, interrupt or take down fundraising activities from their platforms (MoCA, 2017, Art. 5.1.5). The *Measures for the Administration of Public Fundraising Platform Services*, issued jointly with the CAC and other ministries, include additional obligations for platform providers (Ministry of Civil Affairs [MoCA] et al., 2016), such as to publish comprehensive online fundraising information not only within their own messenger apps but also on the platform’s designated websites. At the same time, the original and, by far, most detailed data remain with the companies.<sup>6</sup>

In sum, notwithstanding China’s strictly authoritarian setup with a globally unparalleled level of digital political control over the online space, most developments in the digital philanthropy area are driven by corporate tech giants rather than the state. The following section will, therefore analyze their digital philanthropy models and discuss their implications for fundraising organizations.

### **3 “Smart philanthropy” according to Chinese tech giants**

The paradox of internet philanthropy under authoritarianism is that China’s ICT companies, all while acting as a prolonged arm of the state security apparatus in constraining any genuinely political civil society functions of their social media platforms, have also set out to develop new digital tools for NGOs to engage with stakeholders, organize help for poverty- or disaster-stricken areas, and attract new funding sources. Their vision of “smart philanthropy” (Pi, 2019a) is based on the (largely unchallenged) claim that digital technologies are inherently better at inciting people to donate to charity, identifying the most promising recipients, and connecting the two to increase the overall efficiency and effectiveness of charitable giving.

#### ***3.1 Blurred boundaries between philanthropy and corporate social responsibility***

Control of the Internet Fundraising Platforms formally rests with non-profit organizations, namely the corporate foundations set up by ICT companies as part of their “social responsibility” strategies. However, there is not even a real pretense of organizational autonomy, as the non-profit structures operating the platforms primarily serve to convey the technological innovations of their parent companies to the social sector and to demonstrate their usefulness for philanthropic purposes.

Taking the example of Tencent, the integration between Tencent Foundation and Tencent Inc. is seamless at all levels: From Chen Yidan, Tencent co-founder and honorary chairman of Tencent Foundation, to Guo Kaitian, the company’s senior vice president and chairman of Tencent Foundation, the corporate and foundation management has always been deeply intertwined. Staff are routinely transferred back and forth and sometimes even do company and foundation work on the same job (INT-12, INT-08). A high-level Tencent Foundation manager I interviewed even presented their philanthropic mission as one of “corporate social advocacy” work, which they distinguished from “traditional CSR activities” as a more impactful way of “using a company’s core capabilities to promote groundbreaking developments for the public benefit” (INT-10).

Alibaba, which through its subsidiary Taobao controls large swathes of China's online retailing market, even more bluntly presents charitable donations as an extension of commercial transactions: Thus, Taobao's officially accredited fundraising platform is fully integrated into its commercial retail platform. Charitable projects are advertised and can be searched, found, and "purchased" like hairdryers, smartphones, or holiday trips. Philanthropic behavior on these platforms – particularly in the e-commerce world – is effectively being transformed into just another, "benevolent" form of online consumption. On the one hand, this has helped the quantitative success of online fundraising via these platforms. On the other, it has also subjected non-profit organizations seeking to raise funds to precisely the same commercial logic, including recommendation algorithm optimization and customer attraction, as any retailer business selling for its own profit (INT-06).

These blurred boundaries between for-profit and non-profit structures in those tech empires also explain how the corporate foundations running the platforms think about their own mission: The interviewees I spoke to at Tencent Foundation all had an ICT background rather than expertise in non-profit management or social work. This selective technological expertise and associated value systems have profound implications for how the foundation addresses its field-level responsibilities as administrator of the largest online fundraising platform. Thus, common problems such as the embezzlement of funds, a perceived lack of efficiency, or the widespread public distrust in charitable organizations are consistently framed as being due to "traditional" mindsets and "outdated" business models in the offline world, with better data quality and algorithmic oversight presented as the catch-all solutions (INT-11).

### **3.2 *The promise of technological innovation: "philanthropy for everyone"***

In the offline world, Chinese corporate foundations and philanthropists typically engage in rather conventional, mainly locally oriented charity projects in the education, health, or poverty alleviation sectors (Cunningham & Li, 2020). Tech-philanthropists like Tencent's Chen Yidan explicitly distinguish their work from this "traditional charity" model, derided as ineffective alms-giving (Chen, 2021, p. xxxvii). Instead, they have the aspiration of "bringing charity to the people," namely by building a digital philanthropy system with "zero thresholds 零门槛, more transparency 更透明, and participation opportunities for everyone 人人可以参与" (Guo Kaitian, Tencent senior vice president and chairman of Tencent Foundation, quoted in Tencent, 2017). Combining Andrew Carnegie's notion of "strategic philanthropy" with a Confucian adage, Chen claims that "[c]harity is like giving a man a fish, and philanthropy is like teaching a man to fish as well as improving the area he is fishing in" (Chen, 2021, p. 314). By providing an entirely new digital ecosystem for convenient donations with "philanthropy at your fingertips" (Jing, 2021; official Tencent translation of 指尖公益), Tencent thus allegorically promises a greatly improved fishing area for charitable organizations, i.e., a new online ecosystem with opportunities for everyone. My interviews at Tencent Foundation confirmed that the technology-optimist philanthropy managers running the platform are confident that what they have achieved is not simply quantitative growth but the invention of a whole new form of less elitist, more open, and more effective middle-class philanthropy:

You know what the true achievement of Tencent Foundation is? Having brought charity to the people, having enabled everyone to become a philanthropist [...] Traditional charity is a thing of rich people. Internet charity is very different.

(INT-10)

As public donations made via its *Tencent Charity* platform have been surging (see Figure 14.3), the market leader hails its own “transformative power” in “making philanthropy an everyday part of life” (Tencent, 2022b) and thus, in the words of Chen Yidan, transform Chinese society and even humanity-at-large into a sphere of “kindness” 善良 (Zhang, 2016). Tencent Foundation’s model of “philanthropy for everyone” (Philanthropy Times, 2019) is mirrored by what Alibaba’s Jack Ma has called “personal philanthropy,” which for him consists of unleashing “the power of small” (Luo, 2018), i.e., relying on the collective wisdom of internet users in selecting the best projects.

Stirring user engagement is a top priority and key competence of leading tech companies. Thus, the multiplication of (small-scale) donors among China’s large middle class is an unquestionable achievement of digital philanthropy platforms over the past decade (Internet Society of China & China Philanthropy Research Institute, 2023, p. 14). In addition, the more ambitious promise is to improve the governance of charitable donations significantly, thanks to increased transparency (which, as seen above, has become a legal requirement which tech firms have to comply with) and technological improvements in project selection, governance, and implementation (Wen, 2020).

Capturing these technology-related promises for the non-profit sector, Alibaba has coined the buzzword “charitable big data” (慈善大数据 or 公益大数据) (Gongyi Zhongguo, 2017; Hu, 2019). Meanwhile, under the slogan “Tech for Good” (Yang, 2019), Tencent’s environmental and social governance (ESG) agenda builds on the claim of “using AI to build a more sustainable world” (Tencent, 2020b). Such catchwords have been readily adopted by non-profit scholars and field protagonists, who overwhelmingly present the use and potential of “charitable big data” as a boon for the field (Wang & Li, 2019, p. 262) and a source of “social innovation” (Liang, 2018, pp. 434–437). The potential of digital philanthropy is notably contrasted with the severe governance problems of China’s offline philanthropy sector, which most experts agree has fundamentally lost public confidence after a decade of uncontrolled growth, poor project quality, and numerous misappropriation scandals. According to the Secretary General of the China Foundation Forum, “[t]echnology can greatly improve the philanthropy sector’s productivity, enhance transparency and accountability, promote public participation and build more trust” (quoted in Chu & Wang, 2018, p. 23).

### **3.3 “99 Giving Day”: Tencent’s digital upheaval of philanthropy**

The so-called “99 Giving Day” (99公益日) epitomizes how the digital transformation has stirred up the entire sector. Initially launched by Tencent on 9 September 2015, it has rapidly evolved into a massive ten-day event involving various fundraising competitions and promotional events with a soaring number of external corporate partners (Jing, 2021, pp. 373–375).

99 Giving Day has been an important reason and amplifier of Tencent’s success and dominant position in digital philanthropy, not least thanks to corporate donation matching whereby the event’s corporate sponsors match certain donations made on the platform during the event and thus double their financial impact. Due to its tremendous success in terms of positive publicity, it has been mimicked by Alibaba’s “95 Philanthropy Week” (95公益周) and Sina Weibo’s “Everybody’s Philanthropy Festival” 人人公益节. These corporate events, effectively promoted via the technology companies’ widely used platforms, have largely obfuscated the government’s official “China Charity Day” (中华慈善日), stipulated in the Charity Law (Art. 7).

The cyclical distribution of project data on *Tencent Charity* (see Figure 14.2) underscores that many fundraising organizations have adjusted their entire programming to this central event and designed their projects accordingly. As illustrated in Figure 14.2, 46% of all projects started their fundraising in August in the run-up to what one program manager described as the “99 craziness”

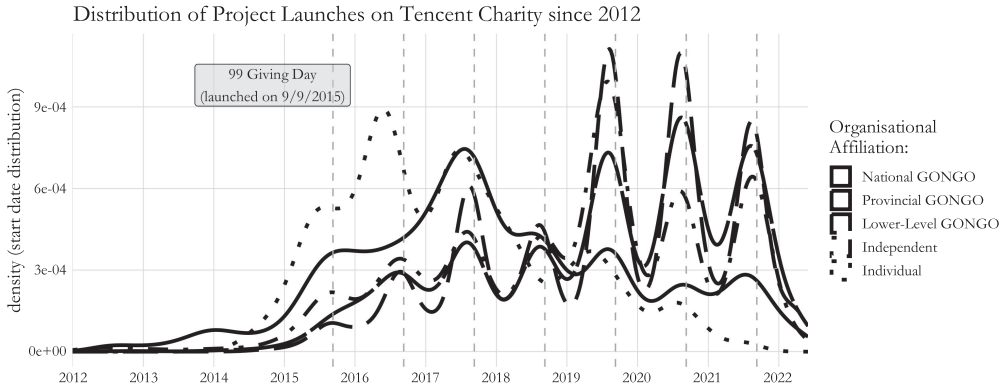


Figure 14.2 Distribution of Tencent Charity project launches over time.

(99疯狂) (INT-06). The latter term reflects a widely shared ambiguity among fundraising organizations regarding the event (and online fundraising on the platform more broadly): Non-profit managers tend to see it as an essential chance to diversify their revenues, especially in a context where foreign donors are shying away from China (INT-04) and Chinese elite philanthropists remain reluctant to offer grants to NGOs (INT-09); at the same time, many complain about a system tilted in favor of larger organizations, as fundraising staff increasingly struggle with growing competition and complex rules (INT-05; see also Zhang, 2020). This competition is also intensified by a growing number of so-called “government-operated NGOs” (GONGOs) – i.e., non-profit organizations with a strong affiliation to government agencies at different levels and are usually tasked with implementing official social policy goals – which have equally discovered the event as an opportunity to shore up their budgets.

99 Giving Day has been used by Tencent Foundation to develop its platform beyond a digital marketplace where NGOs advertise their own projects. The invention of ever-new technology-based tools serves the primary goal of entertaining users and making it fun for them to engage in small-scale charitable action (Chen, 2021; Jing, 2021). For instance, donations are made part of smartphone games for which users pay at the outset and can playfully boost their donation’s impact.

The foundation’s direct intervention in project development and selection has been brought to a new level with the setup of “One Flower, One Dream” (一花一梦想, *yi hua yi mengxiang*) in 2022, a program that lets Tencent’s app users pitch their own philanthropic project ideas (Tencent News, 2022). Proposed projects are then screened and shortlisted by the foundation. Again, the precise criteria fed into the selection algorithm remain opaque, and human selection also appears to play a role in creating the shortlist, even if only a handful of staff are assigned to these tasks. Pre-selected projects are then submitted to a public online vote on the occasion of 99 Giving Day:

“One Flower, One Dream” aims to raise the public awareness of social issues and allows people to track the progress of projects and donations [...] The public are not only donors, but also planners. Companies that provide funds and the beneficiaries will make your dream philanthropic projects a reality.

(Ge Yan, Secretary General of Tencent Foundation, quoted in Tencent, 2022a)

WeChat users, whose “public awareness of social issues” the foundation aims to raise here, donate to vote for their favorite projects with “little safflowers” 小红花 in exchange for receiving

what Tencent calls a “kindness certificate.” The genuine innovation of the program is that for those projects receiving the most popular votes, Tencent Foundation selects “suitable non-profit organizations” to implement them. The platform then promises to “regularly report back to the project’s supporters on its implementation progress” (Tencent, 2022a). This turns the usual NGO logic upside down: Instead of professional fundraising organizations seeking financial support for their projects, the individual donors themselves suggest projects, and the platform organization that puts itself in charge of identifying “suitable” implementing organizations thanks to its technological prowess. This self-arrogated role of an automatic match-maker between “charitable users” striving for impact and non-profit organizations searching for resources is justified on participatory grounds: transforming small-scale donors into “creators” of their own projects.

However, such “charity innovations” introduced by Tencent each year in the run-up to 99 Giving Day have only reinforced the pressure, especially on smaller organizations. Fundraising organizations, which are not involved in developing the tools, now allocate significant resources each year to understand the new logics and how best to use them to their advantage (INT-03; INT-07; INT-02). They are thus increasingly competing with each other for the most innovative and entertaining projects to stand a chance of attracting donations and corporate matching money.

### ***3.4 Market concentration and drop in individual fundraising success***

In addition to the opacity of success criteria in the digital fundraising world and the pressure on NGOs to constantly adjust to new “innovations” imposed on them by platform providers, smaller NGOs have also borne the brunt of a trend toward the concentration of resources over time. The *Tencent Charity* project data thus shows that while the overall number of donors and donations has sharply increased, especially after 2015, these donations have become increasingly concentrated on a few high-profile projects run mainly by large foundations – often with a governmental background. The growing inequality (shown by a rising Gini index since 2011) combined with the rapid growth in project proposals explains why the median amount raised per project, i.e., what an NGO could typically expect to obtain on average, has drastically declined over time (see Figure 14.3). The Covid-19 epidemic in early 2020 further accentuated this trend, with state-affiliated disaster relief organizations soaking up most of the disaster relief funding provided by a population keen on supporting their compatriots in lockdown-affected areas.<sup>7</sup>

Another notable shift in the wake of the Charity Law’s implementation has been a marked drop in projects proposed by individuals (see Figure 14.2). Such projects, mostly involving people pleading for money to pay for the medical treatment of their children and relatives, were relatively successful in the platform’s early, less strictly regulated phase but have drastically declined in contrast with the overall platform boom after 2016. More complicated registration and transparency requirements put in place when the Tencent Charity platform became formally accredited as IFIP (INT-11) have indeed made it much harder for individuals to propose such small-scale self-help projects. Tencent itself has also further raised the bar for less professionalized actors in the name of “transparency and openness”:

Since its inception, 99 Giving Day has adopted transparency as its core principle. This year, Tencent’s philanthropy platform set new participation criteria for institutions and projects. This threshold covers institutional qualifications, information disclosure, cooperation norms, and project viability to support the healthy and sustainable development of the industry.

(Tencent, 2022a)



Platform expansion and rising fundraising inequality on Tencent Charity

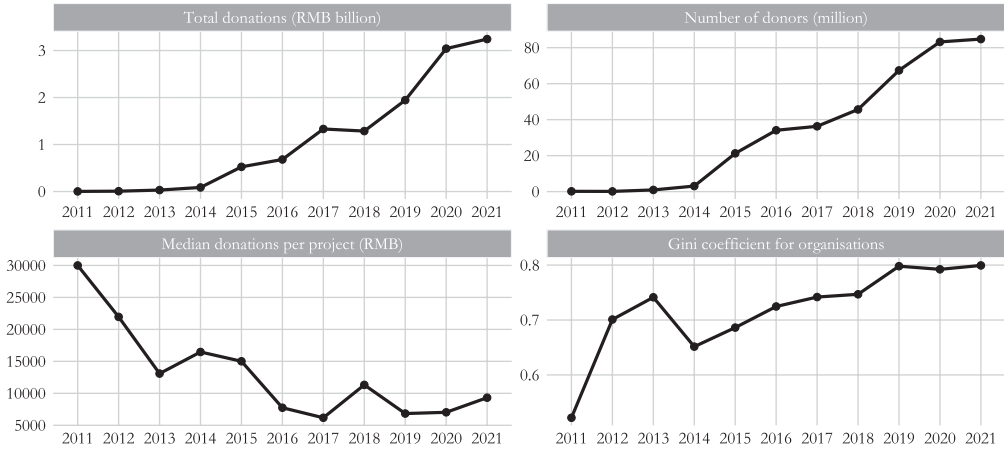


Figure 14.3 Fundraising trends on Tencent Charity: platform expansion and soaring inequality.

Independent organisations and individuals out-competed by GONGOs

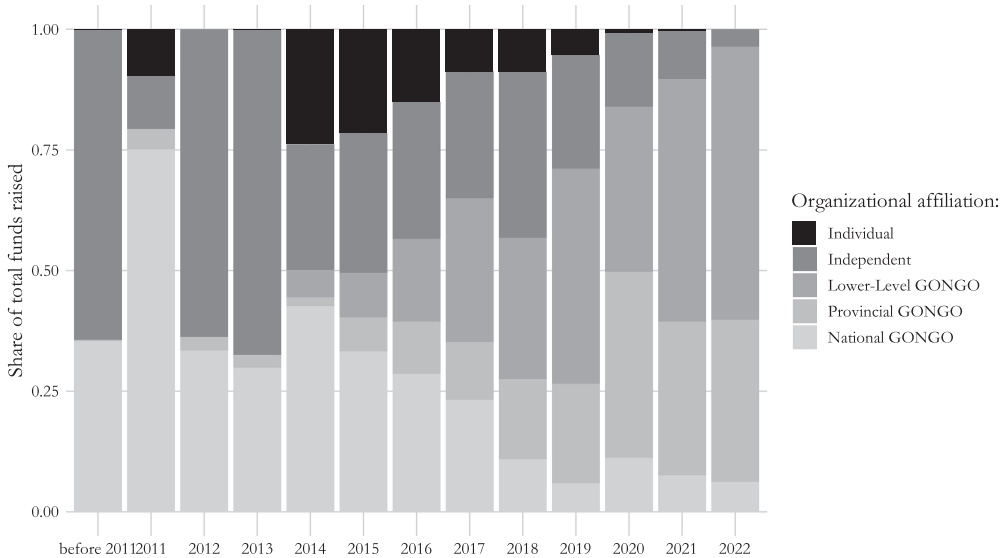


Figure 14.4 Independent organizations and individuals outcompeted by GONGOs.

In this sense, the professionalization of online fundraising has largely closed the door to a more informal and spontaneous practice of crowdfunding-based mutual help, which was characteristic of an early phase of internet philanthropy. More broadly, looking at the fundraising success by organizational affiliation over time (see Figure 14.4), it becomes apparent how GONGOs – especially the rapidly growing number of participating local-level government organizations – have increasingly crowded out independent NGOs and individuals and are walking away with increasing shares of the digital philanthropy jackpot.

To some extent, the challenges, especially for individuals seeking to raise funds, have been recognized in the context of the Charity Law revision process since 2020. However, regulators' primordial concern with political control makes it difficult to compromise on the requirement that only formally accredited organizations with state-approved public fundraising capacity can conventionally act as recipients of funds. This means that both individuals and more independent organizations without that official seal remain dependent on those accredited organizations to sponsor their project applications and receive and manage donations on their behalf. The latter have few incentives to do so for individuals considering the actual risks of fraud and the high costs of cross-checking the veracity and plausibility of individual cases – a problem that is accentuated by the proliferation of machine-generated fraud in China's online environment.

#### **4 AI applications in the Chinese non-profit sector**

The previous section focused on tech firms and their foundations' role as digital philanthropy brokers who use their commercial digital ecosystems to lower participation thresholds for their hundreds of millions of daily users. A second dimension of digital philanthropy is the direct application of artificial intelligence to public interest causes. China has offered a propitious experimentation ground for these applications due to widespread technology optimism, weak AI regulation, as well as non-profit leaders' and experts' agreement on the need for innovative solutions to improve sectoral governance.

##### ***4.1 AI-based project recommendation – a powerful black box***

So far, the use of AI in the context of soliciting, managing, and distributing charitable donations remains unregulated, as the 2016 Charity Law and its implementing regulations still need to consider the concept. In this gray area, companies like Tencent promise to apply any digital technology at their disposal to “broaden the appeal of charities” by integrating project recommendations into their “mobile payments, social networks, cloud computing [...] and games such as Honour of Kings,” following the principle that “[t]he easier it is to make a donation, the more people are willing to help” (Tencent, 2020a).

But whereas fundraising organizations are indeed bound to high transparency standards vis-à-vis regulators and the public, tech firms' proprietary information – especially the algorithms used for selecting and recommending projects to potential donors – remains as opaque as in any other digital platform business. As a result, the determinants of fundraising success on digital platforms are nowhere near as transparent as the information submitted by participating organizations.

Given the close integration of fundraising platforms into social media and e-commerce applications, tech corporate foundations are apparently applying their parent companies' tools for maximizing user engagement on their platforms to attract more users to philanthropic projects and incite them to donate. Whereas the IFIPs are bound to the legal requirement of listing all projects on a dedicated website,<sup>8</sup> user interaction is far higher via smartphone-based application interfaces, where charity projects are recommended to users in a tailored fashion akin to videos on Douyin/TikTok or sponsored products in a Taobao online shop.

The same opacity reigns over Tencent Foundation's pre-selection process of projects on the occasion of 99 Giving Day for special activities like the above-mentioned “little safflower” project. Considering that Tencent Foundation only has a very small team working on this issue,<sup>9</sup> the process of assessing any project's viability is necessarily automated. However, the criteria, let alone

the algorithms used for this selection process, remain unfathomable. Asked about earlier instances of project prioritization, interviewees at Tencent Foundation appeared unaware of the precise selection process and vaguely referred to the parent company and its “high-tech recommendation models” (INT-10).

NGOs’ lack of understanding of what exactly makes projects visible and successful means that they engage in educated guesswork about the obscure workings of Tencent’s algorithms (INT-01; INT-11). As a consequence, a new industry of internet fundraising consultants is emerging who promise to help NGOs and foundations optimize their project advertisements for algorithmic discoverability and “fundraising optimization” (INT-13), which increasingly influences decisions about whether and how projects are designed and proposed in the first place (INT-14). Since Chinese regulators have not stepped in to hold the platform providers to higher standards of decision-making transparency, fundraising organizations relying on Chinese IFIPs today already find themselves in the unenviable situation recently described by Davies in the Alliance Magazine as a possible future of AI-dominated grant-making systems on a global scale:

When it comes to automated decisions that are made by opaque ‘black box’ algorithms, however, it may be entirely unclear to anyone exactly why the decision was taken and where accountability lies if the decision is found to be incorrect. Any civil society organisation that ends up on the wrong side of such a decision may find itself in a Kafkaesque nightmare where it is almost impossible to get answers, let alone redress.

(Davies, 2023)

#### **4.2 Machine learning against machine fraud**

Whereas these concerns voiced by fundraising organizations appear nowhere in the tech-optimist corporate philanthropist discourse, one problem specific to digital philanthropy that is being acknowledged and tackled by platform providers is the proliferation of automated fraud schemes, which equally rely on AI technologies to subvert and exploit the digital philanthropy boom. As the 99 Giving Day evolved into a major national fundraising event with significant financial opportunities for prospective recipients in 2017, Tencent first acknowledged the problem of “machine-generated fraudulent transactions” (机器刷单) and vowed to take action against the intrusion of so-called “black technology” (黑科技) into the field of online fundraising (Li, 2017). Common deceptive strategies consist of the automatized experimentation with many different (spurious) activity descriptions to optimize discoverability, the mass creation of “donor bots” making super-fast donations to own projects simply to extract matching money from corporate sponsors, or the automated creation of fake recommendations to boost organization profiles.

Faced with these challenges of fraud in the context of the 99 Charity Day, Tencent Foundation claims to have developed various counterstrategies. Apart from deploying corporate technology to identify fake profiles, Tencent also announced in 2020 that it used blockchain technology to make projects “traceable and verifiable” (Tencent, 2020a):

Through blockchain technology, all fundraising projects can be “on-chain” to increase transparency. At the same time, the source can be traced, reviewed and cannot be changed at will, and the data information can be tracked throughout the process.

(Li, 2020)

While this may indeed help to prevent retrospective or experimental alterations to online information with the sole purpose of increasing donation prospects, it remains unclear how the fundamental problem – namely that of verifying whether advertised activities online correspond to genuine, well-managed projects in the offline world – could be solved by blockchain or artificial intelligence solutions alone.

### **4.3 China's "Philanthropy 3.0 Era": unquestioned synergies between AI and philanthropy**

Beyond their intermediary role in the solicitation and management of charitable donations, the same tech companies and their foundations are also promoting direct applications of AI technologies to create a "Philanthropy 3.0 Era" which, according to Baidu, consists in "providing a large number of application scenarios for technology to solve social problems" (Hu, 2019). The essential logic of "philanthropy 3.0" is for tech companies to either use their corporate foundations or partner with other non-profit organizations to deploy new technologies free of charge to marginalized populations to demonstrate their social merit and favor broader deployment, including for commercial purposes. A typical case in point is Alibaba's promotion of AI-based education programs to address the problem of a lack of qualified teachers in China's remote rural areas. The program offers digital devices equipped with learning software developed by the Alibaba AI Laboratory to libraries in rural areas (Li Qing, 2018). More ambitious initiatives presented at the China Charity Fair include robot teachers deployed in rural schools or philanthropic offers of AI technologies to improve elderly care (Wang, 2018).

In recent years, other applications of AI to increase the efficiency of charity programs promoted by Baidu, Tencent, or Alibaba have abounded. This includes internationally highly controversial technologies that are applied *pro bono* to support philanthropic projects and demonstrate their "public interest" qualities. The most prominent case in point is the use of facial recognition technology to help people track missing children and other family members (Pi, 2019c; Wen, 2020). Baidu, for instance, claimed in 2019 that it had already helped 8,500 families reunite (Li, 2019). However, these "charitable" achievements cannot be detached from far more problematic applications of the same technologies to track and trace citizens in the name of an authoritarian population control agenda. The very reason these "people search" programs appear to work so well, after all, is the omnipresent deployment of surveillance cameras equipped with highly sophisticated facial recognition technology across Chinese cities and troves of data accumulated on tech companies' servers which help the government manage and analyze them. Thus, what is promoted as "Internet + Rescue and Family Search" (Pi, 2019b) is essentially a side-product – and a social legitimization – of a much larger state-driven and AI-enabled social control program.

Not only is AI increasingly used for the improvement of philanthropy, but philanthropic projects have, in turn, started to be used for the improvement of AI: Under the label of "AI-based poverty evaluation," Alibaba partnered with the China Women's Development Foundation, a major government-affiliated non-profit, in 2019 to offer tailored vocational training for women from impoverished households in areas with high unemployment rates and offer them jobs as "Artificial Intelligence trainers" (The Paper, 2019). In other words, the "philanthropic" project recruits low-cost coders for Alibaba to carry out the burdensome task of human labeling that is crucial in the training of generative AI models and thus heavily sought after by ICT companies across the globe. However, even this ethically questionable use of non-profit structures for obvious business purposes received overwhelmingly positive press reports and is presented as an instance of "AI-based poverty evaluation" in the Chinese press.

## **5 Conclusion: sectoral and global implications of philanthropy digitalization under authoritarianism**

The chapter has elaborated on the two sides of the medal of China's digital philanthropy boom. On the one hand, the advent of digital fundraising platforms and convenience-oriented innovations for making small-scale donations proposed by ICT companies has enabled the participation of large swathes of China's middle class in philanthropic activities. On the other, platform providers' lofty promises regarding technology-based solutions for the sector's problems are fraught with contradictions. While massively increasing the circle of donors and potential beneficiaries, internet philanthropy has also led to a further spike in fundraising inequalities within China's non-profit sector in favor of a few organizations, mainly with a governmental background. Platform data has shown that China's largest online charity event, the "99 Giving Day," has only exacerbated these trends, although the growing concentration of funds starkly contrasts with Tencent's purported efforts to make charity accessible to all.

The Chinese Party-state's attitude toward the private, stock-listed tech firms driving the country's digital transformation has remained ambiguous. On the one hand, the Chinese Party-state has accepted and even reinforced corporate oligopolies with the internet fundraising information platform system mandated by the *Charity Law*. On the other, the regulatory crackdown on "big tech" launched in 2021 underscored the CCP leadership's concerns over ICT firms' vast business empires and discretionary power. But while stepping up pressure on ICT companies to further restrict political controversies and excessive online gaming, authorities have adopted a far more supportive attitude in the field of AI development. The ChatGPT-induced boom in AI-related investments in the US has further fueled this politically supported "AI arms race" in China. As the discussion above has shown, AI innovations are being widely deployed, tested, and fine-tuned in social and educational policy fields, all in the name of philanthropic initiatives and supported by corporate or state-affiliated non-profit organizations.

### ***5.1 Global implications beyond the Chinese case***

China is now at the global cutting edge of ICT developments and aspires to propose "digital solutions" for the world's problems – including in the philanthropic sector (Xu, Huang, & Zhang, 2021). To start with, the ICT firms behind the philanthropic foundations running the Chinese online platforms have acquired a global reach and are engaging in global philanthropic activities (Chu & Wang, 2018). In recent years, official state media have increasingly voiced the aspiration to "contribute a Chinese innovation blueprint to the development of the global philanthropy sector" (China Daily, 2022). A "China model of internet philanthropy" is further developed in state-sponsored think tank reports (Internet Society of China & China Philanthropy Research Institute 2023).

Internet philanthropy is also integrated into China's broader agenda to promote alternative models of internet governance at a global level. China's most important forum for this endeavor, the World Internet Conference in Wuzhen, has served to advance notions such as "internet sovereignty" and promote China's cyberspace governance as a more harmonious and socially beneficial model. It now also features a side event called "Digital Philanthropy and Digital Poverty Alleviation Forum" (数字公益慈善与数字减贫论坛), where China's major tech companies are advancing their claims at improving the world through the development of new information technology (Cyberspace Administration of China 2022). State media have started to present

“AI+ philanthropy [as] a global consensus” and call for the “construction of international AI public welfare platforms to boost the AI public welfare industry” (Guangming Daily, 2023), including in the United Nations.

Technological innovations, including AI applications in the Chinese philanthropy field, will thus have global implications. Discussions about the potential applications of the same or similar tools in other contexts will need to consider the political context, given the extreme level of political control that the Chinese Party-state is now exerting over the economy and *a fortiori* over the non-profit world. There are clear ethical concerns regarding the application of potentially intrusive technologies such as facial recognition for tracking people, but also, more broadly, regarding the use of vast amounts of organization and project data to train AI models and recommendation algorithms.

However, the Chinese case must not be exoticized to suggest that problematic developments only occur there or that all philanthropy innovations coming from China are inherently threatening. For one, this prevents genuine insights into the opportunities and risks associated with AI applications to philanthropy. Secondly, flat-out rejection of Chinese innovations in Europe and the US will not halt their diffusion to other parts of the world, as Chinese experiences in poverty alleviation, disaster relief, and prevention, or medical care are already being promoted and transferred to many countries in the Global South. Here again, Chinese tech companies, who are expanding their commercial empires along the Chinese “Belt and Road Initiative” (Su & Flew, 2021), are also becoming a driving force in the philanthropic field, as the provision not only of medical equipment but also of e-health governance platforms to governments in the Global South in the wake of the Covid-19 pandemic has demonstrated. In this context, global philanthropy researchers and practitioners will need to keep an open and critical eye on Chinese digital philanthropy developments to understand their innovative potential, associated risks, and ethical challenges.

Table of interviews

<i>Code</i>	<i>Date</i>	<i>Description</i>
INT-01	01/09/2017	Mid-level manager at Chinese civic foundation
INT-02	11/09/2018	Executive director of a Chinese foundation with international cooperation partners
INT-03	14/09/2018	Program manager, Chinese GONGO Foundation
INT-04	09/06/2019	Senior advisor and former CEO of a Chinese civic foundation
INT-05	06/09/2019	Head of International Activities at a Shenzhen-based civic foundation
INT-06	06/09/2019	Research officer at a Shenzhen-based Chinese foundation
INT-07	19/09/2019	China-based philanthropy consultant
INT-08	20/09/2019	Shenzhen-based CSR and corporate charity expert and consultant
INT-09	21/09/2019	Mid-level manager at a Chinese non-profit organization offering training and consulting for NGOs
INT-10	23/09/2019	Leading manager, Tencent Foundation
INT-11	23/09/2019	Project officer, Tencent Foundation
INT-12	29/06/2019	CSR officer in a large corporate foundation in Shenzhen
INT-13	02/10/2020	Manager at a civic NGO platform
INT-14	11/11/2020	Project officer, Chinese environmental NGO

## Notes

- 1 One peculiarity of China's foundation sector is the prevalence of fundraising foundations, i.e., organizations that are set up as "foundations" (基金会) with an initial endowment but still rely on additional funds from the public or from institutional donors to carry out their project activities. Only a small minority of foundations are purely grant-making, although this structure following the US model is actively promoted by internationally connected field-level organizations. This means that a clear-cut separation between grant-making foundations and fundraising NGOs is impossible to uphold in China's philanthropy environment.
- 2 Within a decade from 2011 to 2021, the number of donors using the platform annually surged from 181,897 to 84.8 million, a 468-fold increase.
- 3 Taking the example of generative AI, between March and September 2023 alone, search engine giant Baidu first launched its ChatGPT competitor Ernie Bot, Alibaba announced the roll-out of its generative AI service *Tongyi Qianwen* across various e-commerce services, and Tencent promoted its foundation AI model *Hunyuan* to business clients.
- 4 According to what it later rationalized as a "data sovereignty" (Liu, 2021, p. 52) or "cyber sovereignty" imperative, the Chinese government has consistently protected and propped up these corporate "national champions" dominating the Chinese cyberspace today. In departure from earlier Western literature positing ICT as a liberating technology and an inherent threat to autocratic rule, the CCP leadership identified the regime-stabilizing potential of digital governance (Göbel, 2013, p. 836) and has consequently made the promotion of big data artificial intelligence solutions across the entire economy and society a top policy priority.
- 5 According to my calculations based on Jian (2018), market concentration in the online fundraising market in 2017 amounted to a value of 0.45 on the Herfindahl–Hirschman index. Any value above 0.25 is indicative of a highly concentrated market.
- 6 It remains unclear how much non-public data is shared with state administrators on a regular basis, but the MoCA as the responsible supervisor would most likely be technically unable to handle Tencent's massive datasets anyway.
- 7 According to my database, the two most successful fundraising projects overall were Covid-19 emergency relief programs launched in January and February 2020 by the Hubei Charity Federation and the China Social Welfare Foundation (two non-profits with close government ties), respectively.
- 8 The listing on the Tencent Charity platform's website, from where I obtained the project dataset, is ordered by project status, topic, and time of submission. However, there are featured projects on the front page, with no information about how projects are selected for advertisement.
- 9 As per Tencent Foundation's 2021 work report filed with the registration authorities (MoCA 2023), only six staff members are assigned to platform management overall. It is obvious that they are nowhere near able to review the 25,398 projects underway on the platform in 2021.

## References

- Calvano, E., & Polo, M. (2021): Market Power, Competition, and Innovation in Digital Markets: A Survey. *Information Economics and Policy*, 54(1), 100853. <https://doi.org/10.1016/j.infoecopol.2020.100853>
- Chan, N. K., & Kwok, C. (2022): The Politics of Platform Power in Surveillance Capitalism: A Comparative Case Study of Ride-Hailing Platforms in China and the United States. *Global Media and China*, 7(2), 131–150. <https://doi.org/10.1177/20594364211046769>
- Chen, Y. (2021): *Internet Philanthropy in China*. Singapore: Palgrave Macmillan.
- China Daily (2022): 《中国数字公益发展研究报告（2022）》重磅发布 [China Digital Philanthropy Development Research Report (2022) Released], 02/09/2022, archived online 27/09/2023.
- China Foundation Center (2013): The Transparency of Chinese Foundations Disputed, January 2013. <http://en.foundationcenter.org.cn/html/2013-01/60.html>, archived online 29/03/2019.
- Chu, P., & Wang, O. (2018): *Philanthropy in China* (Supported by Rockefeller Foundation). Retrieved from Asian Venture Philanthropy Network website: <https://www.rockefellerfoundation.org/report/philanthropy-in-china/>
- Cunningham, E., & Li, Y. (2020): China's Most Generous. Examining Trends in Contemporary Chinese Philanthropy. Harvard Kennedy School, Ash Center for Democratic Governance and Innovation, March 2020. PDF available online via. <https://chinaphilanthropy.ash.harvard.edu/> [checked 13/03/2023].

- Cutolo, D., & Kenney, M. (2021): Platform-Dependent Entrepreneurs: Power Asymmetries, Risks, and Strategies in the Platform Economy. *Academy of Management Perspectives*, 35(4), 584–605. <https://doi.org/10.5465/amp.2019.0103>
- Cyberspace Administration of China (2022): 2022年世界互联网大会乌镇峰会数字公益慈善与数字减贫论坛举行, 10/11/2022. [http://www.cac.gov.cn/2022-11/10/c\\_1669712526540556.htm](http://www.cac.gov.cn/2022-11/10/c_1669712526540556.htm)
- Davies, R. (2023): Artificial Intelligence Is Coming for Philanthropy. *Alliance Magazine*, 28/03/2023. <https://www.alliancemagazine.org/analysis/artificial-intelligence-is-coming-for-philanthropy/>
- Doorn, N., & Badger, A. (2020): Platform Capitalism's Hidden Abode: Producing Data Assets in the Gig Economy. *Antipode*, 52(5), 1475–1495. <https://doi.org/10.1111/anti.12641>
- Gao, F. (2016): Social Media as a Communication Strategy: Content Analysis of Top Nonprofit Foundations' Micro-blogs in China. *International Journal of Strategic Communication*, 10(4), 255–271. <https://doi.org/10.1080/1553118X.2016.1196693>
- Göbel, C. (2013): The Information Dilemma: How ICT Strengthen or Weaken Authoritarian Rule. *Statsvetenskaplig tidskrift*, 115(4), 385–402.
- Gongyi Zhongguo (2017): 阿里发布中国首个公益大数据开放平台, 打造智慧公益体系 [Alibaba Releases China's First Philanthropy Big Data Open Platform, Creating a Smart Philanthropy System], 07/05/2017, archived online.
- Guangming Daily (2023): AI公益, 开启全球“共益时代” [AI Philanthropy Opens Up the Global “Era of Common Welfare”], 27/06/2023. [https://politics.gmw.cn/2023-06/27/content\\_36656266.htm](https://politics.gmw.cn/2023-06/27/content_36656266.htm)
- Holbig, H., & Lang, B. (2022): China's Overseas NGO Law and the Future of International Civil Society. *Journal of Contemporary Asia*, 52(4), 574–601. <https://doi.org/10.1080/00472336.2021.1955292>
- Howell, J. (2019): NGOs and Civil Society: The Politics of Crafting a Civic Welfare Infrastructure in the Hu–Wen Period. *The China Quarterly*, 237, 58–81. <https://doi.org/10.1017/S0305741018001236>
- Hu Bin 胡彬 (2019): “人工智能+公益”: 百度公益的3.0时代 [“Artificial Intelligence + Philanthropy”: Baidu Philanthropy's 3.0 Era], *Philanthropy Times*, 25/09/2019. <http://www.gongyishibao.com/html/gongyizixun/17374.html>
- Internet Society of China & China Philanthropy Research Institute (2023): 互联网公益慈善“中国样本” [The ‘China Model’ of Internet Philanthropy, Research Report], 22/05/2023, available online, archived online 29/09/2023.
- Jian, Y. 菅宇正 (2018): 互联网募捐平台年报纵览 [Annual Overview of Internet Fundraising Platforms], *Philanthropy Times*, 28/02/2018. <http://www.gongyishibao.com/html/yaowen/13464.html>
- Jing, Y. 景燕春 (2021): “99公益日”, 助力公益持续发力. [“99 Charity Day” Helps Philanthropy Continue to Develop]. In: Yang, Tuan 杨.; Zhu, Jiangan 朱. (Eds.): 慈善蓝皮书. 中国慈善发展报告 (2021) [Blue Book of Philanthropy. Annual Report on China's Philanthropy Development (2021)]. Beijing: 社会科学文献出版社 (Social Sciences Academic Press), 373–379.
- Lai, W., & Spire, A. J. (2021): Marketization and Its Discontents: Unveiling the Impacts of Foundation-Led Venture Philanthropy on Grassroots NGOs in China. *The China Quarterly*, 245, 72–93. <https://doi.org/10.1017/S0305741020000193>
- Lang, B. (2018): Authoritarian Learning in China's Civil Society Regulations: Towards a Multi-Level Framework. *Journal of Current Chinese Affairs*, 47(3), 147–186. <https://doi.org/10.1177/186810261804700306>
- Li, Q. 李庆 (2018): 农村淘宝将建500个乡村公益图书馆 人工智能进村当“幼师” [Rural Taobao will build 500 Rural Public Welfare Libraries and Artificial Intelligence Will Be Used as “Kindergarten Teachers” in Villages]. *Philanthropy Times*, 06/06/2018. <http://www.gongyishibao.com/html/qiyeCSR/14114.html>
- Li, Q. 李庆 (2019): 百度公司: 帮助超过8500个家庭重获团聚 [Baidu: Helping More Than 8,500 Families Reunite]. *Philanthropy Times*, 22/11/2019. <http://www.gongyishibao.com/html/qiyeCSR/17709.html>
- Li, Q. 李庆 (2020): 99公益日在即, 如何规避“刷单”、“套捐”、“逼捐”等违规操作? 腾讯表示: 将实施“小黑屋”策略, *Philanthropy Times*, 07/09/2020, archived online 03/03/2021.
- Li, Y. 黎宇琳 (2017): 腾讯确认99公益日存在“机器刷单”, 警惕“黑科技”入侵公益筹款领域 [Tencent Confirms Existence of “Machine-Generated Fraudulent Transactions” on 99 Charity Day, Warns against Intrusion of “Black Technology” into the Field of Charity Fundraising]. *China Charity Forum*, 11/10/2017, archived online 27/09/2023.
- Liang, C. 梁. (2018): 互联网时代的社会创新和公益转型: [Social Innovation and Philanthropy Transformation in the Internet Age]. In B. 王. Wang & T. 杨. Yang (Eds.), 慈善蓝皮书. 中国慈善发展报告 (2018) : [Blue Book of Philanthropy. Annual Report on China's Philanthropy Development (2018)]. Beijing: 社会科学文献出版社 (Social Sciences Academic Press), 433–445.



- Liu, L. (2021): The Rise of Data Politics: Digital China and the World. *Studies in Comparative International Development*, 56(1), 45–67.
- Luo, G. (2018): Spotlight on Alibaba's Approach to Philanthropy: Unleashing "The Power of Small". Elevate, 12/09/2018, archived online 29/09/2023.
- Ministry of Civil Affairs [MoCA] (2017): 慈善组织互联网公开募捐信息平台基本技术规范 (Basic Technical Specifications of Online Fundraising Platform for Charitable Organization), 20/07/2017, MZ/T 087-2017.
- Ministry of Civil Affairs [MoCA] (2018): 民政部关于发布慈善组织互联网公开募捐信息平台名录的公告, 2018-06-04, 民政部公告第434号, archived online 06/05/2022.
- Ministry of Civil Affairs [MoCA] (2021): 民政部关于指定第三批慈善组织互联网募捐信息平台的公告 Communication of the Ministry of Civil Affairs on Appointment of the Third Lot of Internet Fundraising Information Platforms, 2021/11/15, archived online 06/12/2021.
- Ministry of Civil Affairs [MoCA] (2023): 慈善中国. 民政一体化政务服务平台 [Charity in China. Civil Affairs Integrated Government Service Platform], URL: <https://web.archive.org/web/20210323172813/http://cishan.chinanpo.gov.cn/platform/login.html> [last checked 14/08/2024].
- Ministry of Civil Affairs [MoCA], Ministry of Industry and Information Technology, State Administration of Press, Publication, Radio, Film and Television and Cyberspace Administration of China (2016): 《公开募捐平台服务管理办法》 [Measures for the Administration of Public Fundraising Platform Services], No. 157 [2016], CLI.4.279183.
- People's Daily (2012): 推动社会组织健康有序发展 [Promoting the Healthy and Orderly Development of Social Organisations], 25/04/2012, p. 7.
- Philanthropy Times (2016): 新《慈善法》如何守护网络公益, 23/08/2016. <http://www.gongyishibao.com/html/gongyizixun/10271.html>
- Philanthropy Times (2019): 阿里、腾讯、轻松筹六大平台出道 “互联网第一公益天团”助力人人公益 (Alibaba, Tencent, Qingsong Health, Six Major Platforms to Debut "Internet No.1 Philanthropy Alliance" to Support Philanthropy for Everyone), 09/09/2019. <http://www.gongyishibao.com/html/gongyizixun/17259.html>
- Pi, L. 皮磊 (2019a): 让数据助推行业发展 慈善行业第三方数据平台“易善数据”发布 [Let Data Boost the Development of the Industry and Release the Third-Party Data Platform "Yishan Data" for the Charity Industry]. *Philanthropy Times*, 13/11/2019. <http://www.gongyishibao.com/html/gongyizixun/17660.html>
- Pi, L. 皮磊 (2019b): 《走失人口数据报告》发布 “互联网+救助寻亲”大大提升寻亲效率 [The "Lost Population Data Report" Released "Internet + Rescue and Family Search" to Greatly Improve the Efficiency of Family Search]. *Philanthropy Times*, 30/07/2019. <http://www.gongyishibao.com/html/gongyizixun/17005.html>
- Pi, L. 皮磊 (2019c): 用科技提高寻人效率 “头条寻人”帮助9000个家庭团圆 [Using Technology to Improve the Efficiency of Tracing People "Toutiao Tracing" Helps 9,000 Families Reunite]. *Philanthropy Times*, 12/04/2019. <http://www.gongyishibao.com/html/gongyizixun/16355.html>
- Plantin, J.-C., & Seta, G. de (2019): WeChat as Infrastructure: The Techno-Nationalist Shaping of Chinese Digital Platforms. *Chinese Journal of Communication*, 12(3), 257–273. <https://doi.org/10.1080/17544750.2019.1572633>
- Qu, Y. (2020): Engaging Publics in the Mobile Era: A Study of Chinese Charitable Foundations' Use of WeChat. *Public Relations Review*, 46(1), 1–9. <https://doi.org/10.1016/j.pubrev.2019.101815>
- Sidel, M. (2014): The Shifting Balance of Philanthropic Policies and Regulations in China. In: Ryan, Jennifer, et al. (Eds.): *Philanthropy for Health in China*. Bloomington: Indiana University Press, 40–56.
- Sidel, M. (2019): Managing the Foreign. The Drive to Securitize Foreign Nonprofit and Foundation Management in China. *Voluntas*, 30(4), 664–677.
- Sidel, M. (2022): Rebooting China's Charity Law. *Usali Perspectives*, 2(25), 23/06/2022. <https://usali.org/usali-perspectives-blog/rebooting-chinas-charity-law>
- Sohu Technology (2016): 让互联网成为爱的海洋 – 发展网络公益倡议书 (Let the Internet Become an Ocean of Love – Proposal for the Development of Internet Philanthropy), 01/11/2016, archived 05/10/2022.
- Spire, A. J. (2020): Regulation as Political Control: China's First Charity Law and Its Implications for Civil Society. *Nonprofit and Voluntary Sector Quarterly*, 49(3), 571–588. <https://doi.org/10.1177/0899764019883939>
- State Council of the PRC (2014): 国务院关于促进慈善事业健康发展的指导意见 [Guiding Opinions on Promoting Healthy Development of Charities], 24/12/2014, CLI.2. 239908.

- State Council of the PRC (2015): 国务院关于积极推进“互联网+”行动的指导意见 [Guiding Opinions on Vigorously Advancing the “Internet Plus” Action], No.40, 01/07/2015, CLI.2.250823.
- Su, C., & Flew, T. (2021): The Rise of Baidu, Alibaba and Tencent (BAT) and Their Role in China's Belt and Road Initiative (BRI). *Global Media and Communication*, 17(1), 67–86. <https://doi.org/10.1177/1742766520982324>
- Tencent (2017): 腾讯发布十周年公益白皮书, 拿出20亿资源助力公益 [Tencent Released the 10th Anniversary Philanthropy White Paper and Provided 2 Billion Resources to Help Philanthropy]. *Tencent Research Institute WeChat Account*, 13/06/2017, archived online 09/09/2022.
- Tencent (2020a): Going Digital: The Future of Internet-Based Charity, 11/09/2020. <https://www.tencent.com/en-us/articles/2201082.html>
- Tencent (2020b): Protecting Our Planet: Building Sustainable Solutions to Global Problems, 20/10/2020, archived online 04/06/2023.
- Tencent (2022a): Tencent's 99 Giving Day Kicks Off with New Feature as It Continues to Improve Digital Solutions for Philanthropy, 02/09/2022, archived online 06/09/2022.
- Tencent (2022b): The Transformative Power of 99 Giving Day: How Tencent's Annual Charity Event Is Making Philanthropy an Everyday Part of Life, 14/09/2022, archived online 19/09/2022.
- Tencent News 腾讯新闻 (2022): 99公益日: 一花一梦想, 一起来“种”花 [99 Giving Day: One Flower One Dream, Let Us “Plant” Flowers Together], 23/08/2022, archived online 13/02/2023.
- The Paper (2019): 当AI遇上扶贫: 大山里走出“人工智能培训师”[When AI Meets Poverty Alleviation: “Artificial Intelligence Trainers” Emerge from the Mountains], 13/08/2019 [https://www.thepaper.cn/newsDetail\\_forward\\_4157782](https://www.thepaper.cn/newsDetail_forward_4157782)
- Tsai, K. S., & Wang, Q. (2019): Charitable Crowdfunding in China: An Emergent Channel for Setting Policy Agendas? *The China Quarterly*, 240, 936–966. <https://doi.org/10.1017/S030574101800139X>
- Wang, M., & Li, S. (2019): The Development of Charitable Organizations in China Since Reform and Opening-Up and a New Layout for State-Society Relations. In J. Yu & S. Guo (Eds.), *The Palgrave Handbook of Local Governance in Contemporary China*. Singapore: Springer Singapore, 245–265.
- Wang, Y. 王勇 (2018): 社区居家养老服务主题论坛举行 引入人工智能助力养老事业 [Community Home-Based Elderly Care Service Theme Forum Was Held to Introduce Artificial Intelligence to Help Elderly Care]. *Philanthropy Times*, 11/05/2018. <http://www.gongyishibao.com/html/gongyizixun/13890.html>
- Wen, R. 温如军 (2020): 谈互联网新贵的公益版图: 是作秀还是有更为长远的布局? *China Philanthropist* (中国慈善家杂志), 17/08/2020, archived online 27/09/2023.
- Xu, J., Huang, D., & Zhang, H. (2021): Internet Philanthropy as China's ‘Digital Solution’ to the 2030 Agenda for Sustainable Development. In M. Jameel Yusha'a & J. Servaes (Eds.), *The Palgrave Handbook of International Communication and Sustainable Development* (1st ed.). Cham: Springer International Publishing, 371–391.
- Yang, Y. (2019): China's Tencent Pitches Vision of Artificial Intelligence Ethics. *Financial Times*, 1/05/2019. <https://www-ft-com.ezp.lib.cam.ac.uk/content/f92abc38-6bb8-11e9-80c7-60ee53e6681d>.
- Zeng, J. (2020): Artificial Intelligence and China's Authoritarian Governance. *International Affairs*, 96(6), 1441–1459. <https://doi.org/10.1093/ia/iiaa172>
- Zhang, M. 张明敏 (2015): “世界捐助指数”: 中国倒数第四 [“World Giving Index”: China 4th from the Bottom]. *Philanthropy Times*, 17/11/2015. <http://www.gongyishibao.com/html/xinwen/8753.html>.
- Zhang, M. 张明敏 (2016): 陈一丹: 善良是一种选择. *Philanthropy Times*, 05/05/2016. <http://www.gongyishibao.com/html/renewuzhishu/9685.html>
- Zhang, W. (2020): Grassroots NGOs See Lots of Give, Not Enough Take on Charity Day, Sixth Tone, 10/09/2020, archived online 02/09/2022
- Zhang, X., Xiang, Y., & Hao, L. (2019): Virtual Gifting on China's Live Streaming Platforms: Hijacking the Online Gift Economy. *Chinese Journal of Communication*, 12(3), 340–355. <https://doi.org/10.1080/17544750.2019.1583260>
- Zhou, H., & Pan, Q. (2016): Information, Community, and Action on Sina-Weibo: How Chinese Philanthropic NGOs Use Social Media. *Voluntas*, 27(5), 2433–2457. <https://doi.org/10.1007/s11266-016-9685-4>

# A CASE STUDY ON AI USAGE FOR COLLECTING PHILANTHROPY DATA IN THE WESTERN BALKANS

*Nikola Milinković and Marko Galjak*

## 1 Introduction

In an era where data reigns supreme, understanding the intricate web of philanthropy becomes both a challenge and a necessity. Philanthropic gestures, rooted deep within cultural, socio-economic, and political spheres, often echo the rich tapestry of diverse human motivations, aspirations, and needs. The Western Balkans, a region steeped in history and multifaceted identities, is a testament to the complexity of such philanthropic dynamics. Drawing connections and understanding giving patterns here is not merely an academic exercise but a pursuit that can inform, guide, and inspire impactful and sustainable philanthropic initiatives in a rapidly changing world.

The Giving Balkans philanthropy database emerges as a beacon in this quest, capturing the philanthropic heartbeat of seven unique Western Balkan countries: Serbia, Croatia, Bosnia and Herzegovina, Albania, Macedonia, Kosovo, and Montenegro. Documenting a staggering 90,313 philanthropic gestures involving 35,779 distinct entities and transcending half a billion euros between 2015 and 2022, the database serves as a record and a mirror reflecting the nuanced dance of giving in the region. As we delve deeper into this chapter, we aim to unravel the digital transformation journey of this database, accentuated by the adoption of artificial intelligence. Through a case study lens, we will shed light on the successes, challenges, and the forward trajectory of integrating AI for better, more efficient, and more insightful data collection and analysis.

This chapter is written in the middle of the transition to AI. It offers a snapshot of the current state of using AI for philanthropy data gathering and processing in the Western Balkans and is in no way the final form. With the field of AI evolving quickly, this is likely the case in two years. Our process will evolve following the rapid advancement of AI methods and tools.

### *1.1 The Western Balkans' rugged philanthropic landscape—a contextual glimpse*

The philanthropic landscape of the Western Balkans is complex. The first step in grasping the complexity is understanding the context in which Giving Balkans philanthropy data is collected. The Western Balkans is an intriguing ensemble of seven nations presenting a rich mosaic of shared histories and unique trajectories. While the majority of the countries—Croatia, Bosnia and

Herzegovina, Montenegro, and Serbia—were once tethered under the banner of Yugoslavia and share linguistic ties, Albania is different, marked by isolation during the Hoxha regime (Glenny, 2001). Although retaining its linguistic link with Albania, Kosovo traversed its journey alongside the former Yugoslav states. The linguistic diversities, with Macedonia’s South Slavic language and the Albanian threads of Kosovo and Albania, add another layer to this complexity (Duncan, 2016). These linguistic differences pose a significant challenge in collecting philanthropy data consistently across the region. The challenge is primarily in terms of human resources. Overcoming the linguistic barriers requires a skilled workforce fluent in the region’s diverse languages.

Moreover, an in-depth understanding of the local civic and philanthropic landscapes is important, often necessitating the recruitment of knowledgeable personnel from within each country. This ensures that data collection is linguistically accurate and culturally and contextually informed, an important aspect for reliable and comprehensive philanthropic data gathering. Another consideration is the verification process, where Philanthropy Data Analysts check with either donor or beneficiary about the donation instance that has appeared in the media. Having a local inquiring about veracity and additional information is much more likely to solicit a response than reaching out from a different country. This can be an important consideration in light of the region’s historically sensitive context.

There are also considerable differences in the economies of the Western Balkans, as the poorest country (Kosovo) has a gross domestic product per capita (GDPPC) of \$5,531.5. In contrast, the richest, Croatia, has a GDPPC of \$18,413 (World Bank, 2023). This more than threefold difference in countries’ economic output translates to different philanthropic contexts (i.e., how philanthropy is practiced and therefore tracked can also be different). This layer adds additional methodological difficulties.

The post-socialist tapestry these nations wear resonates with overtones of state dependency rooted in the shared communist history. Citizens, influenced by this past, often look toward the government as the primary steward of societal welfare (Grødeland, 2006). This outlook on governance, combined with the region’s sporadic political instabilities (EWB, 2023), corruption challenges (Transparency International, 2020), and evolving democracies (Freedom House, 2016), intricately shapes the philanthropic motivations and actions here. In this region, the concept of philanthropy often diverges from traditional forms like fundraising for community development or crowdfunding for community projects; such practices are not deeply ingrained. Instead, a significant amount of philanthropy is informal and immeasurable, characterized by people spontaneously helping each other in times of need. Civil Society Organizations (CSOs) have become heavily reliant on an influx of foreign funding, a trend that began in the 1990s. Recently, a notable shift in philanthropic focus has emerged, with crowdfunding campaigns for healthcare, especially for sick children, gaining prominence and attracting considerable support over the last decade. The ingrained perception of the state as the primary provider of all societal needs, juxtaposed with the emergent necessity to crowdfund for the immediate healthcare needs of sick children, underscores a significant shift in the region’s philanthropic landscape. This juxtaposition highlights a changing dynamic where the public increasingly recognizes the limitations of state support and turns toward philanthropy for urgent and critical needs.

The legal environment is a pivotal factor in the philanthropic scenario of the Western Balkans. While offering growth avenues for non-profits, the legislative matrix also interposes specific barriers that might constrict their impact and reach (USAID, 2023) (e.g., difficulties around regulations of volunteer work and lack of legal frameworks that support volunteerism, no tax incentives for philanthropy, collecting VAT on food donations, etc.). Additionally, the region’s socio-economic fabric, a product of its transition from socialism to a more economically and politically liberal

regime, plays a significant role in how philanthropy evolves and operates. Economic transition brought with it new wealth and economic disparities. With no legacy philanthropic organizations, a nascent CSO sector, the strong influence of international donors, and clumsy corporate responsibility adopted from multinational corporations' headquarters, philanthropy in the Western Balkans evolved uniquely in the past three decades.

The difficult economic situation and wars boosted traditional emigration from the region to the more prosperous countries. While diaspora communities might be deeply integrated into their host countries, spanning generations, they often maintain a pronounced inclination toward philanthropic efforts directed at their countries of origin (Brinkerhoff, 2014). This highlights the potential for viewing the diaspora as a local resource for philanthropic endeavors. This is especially true in countries like Bosnia and Herzegovina, Albania, and Kosovo, where remittances form a significant economic component (Bajra, 2021), and the diaspora's role becomes central. As the Western Balkans grapple with the diminishing influence of foreign donors, cultivating a robust, locally sourced philanthropy ecosystem emerges as a crucial fulcrum for both the sustenance of non-profits and the democratic fabric of the region.

The Western Balkans' philanthropic landscape is shaped by several key factors: diverse linguistic backgrounds complicating data collection, significant economic disparities impacting philanthropic practices, a post-socialist legacy influencing public reliance on government for welfare, and legal challenges that restrict non-profit activities. Additionally, the socio-economic shift from socialism and the unique role of the diaspora, especially in countries heavily reliant on remittances, significantly shape the region's approach to philanthropy. These elements collectively contribute to the complexity of philanthropy in the Western Balkans.

### ***1.2 Catalyst Balkans—the organization behind the Giving Balkans***

Catalyst is a Serbian-registered foundation launched in early 2013 to promote the growth and improved transparency of individual and corporate philanthropic culture in the Western Balkans and to further the digital transformation of the non-profit sector. Based in Belgrade, Serbia, Catalyst covers seven countries.<sup>1</sup> Catalyst Balkans has built its reputation as a go-to partner for information, support, or advice on domestic giving by taking a systems-based approach to broaden and deepen the Western Balkans' philanthropy ecosystem. It is an example of locally led development; Catalyst was founded specifically to address gaps in the ecosystem and work in partnership with established and emerging stakeholders. Catalyst's active participation in ecosystem entities includes the Serbian Philanthropy Forum, Philanthropy Forum of Bosnia and Herzegovina, Kosovo Philanthropy Forum, Southeastern Europe Indigenous Grantmakers Network (SIGN) Network, and the European Research Network on Philanthropy (ERNOP).

Catalyst Balkans provides tech and services to the philanthropy and non-profit ecosystems of the Western Balkans. Through Donacije.rs,<sup>2</sup> 198 non-profits have raised \$1.18 million. Using CiviCatalyst.org, a service based on open-source CRM (Constituent Relationship Management) software for non-profits that Catalyst provides hosting and customization services, 120 non-profits manage their data more securely. Three hundred and fifty Serbian non-profits have claimed a transparency badge on Nprofitne.rs. The unique Giving Balkans methodology for collecting transaction-level micro data on domestic philanthropic flows in seven Western Balkans countries relies on a combination of press clipping and direct verification of gathered data with recipients and donors. Using this data, Catalyst identifies several key trends in philanthropy. Firstly, there is an increasing tendency toward mass individual giving, indicating a shift in how individuals

contribute to charitable causes. Secondly, there is a notable change in how corporate donors operate. Corporations are becoming more sophisticated in their giving strategies, often channeling their donations through non-profits, even when public institutions are the intended final beneficiaries. This approach signifies a strategic move toward leveraging the expertise and networks of non-profits for more impactful giving. Lastly, there is a rise in community-based philanthropy, reflecting a growing emphasis on localized, grassroots efforts to address societal needs.

### ***1.3 Giving Balkans database***

Giving Balkans gathers data on charitable giving in the Western Balkans region using alternative methods, primarily sourcing information from media reports and other readily available resources. Official data on philanthropy from key institutions such as the Ministries of Finance and the Tax Administration is absent in the region. To bridge this gap, Giving Balkans continuously monitors printed, electronic, and online media on local, regional, and national levels within the Western Balkans for any instances of giving. This data can then be easily accessed by non-profit organizations, corporations, and individuals, providing a better macro understanding of the entire ecosystem and concrete philanthropy intelligence.

As of late 2023, the database contains more than 87,000 instances of giving by more than 13,000 distinct donors to more than 24,000 distinct beneficiaries, amounting to more than 644 million euros (Catalyst Balkans, 2023).

Giving Balkans app is an interactive web application built using the R programming language (R Core Team, 2023) and Shiny (Chang et al., 2023) R package, both of which are open-source (see Figure 15.1). The user-friendly app facilitates data exploration through intuitive filtering across all dimensions by clicking on the visualizations. This ease of use empowers those without a technical background to delve into what is often called “philanthropy intelligence.” Such insights can assist non-profit organizations in strategizing their fundraising initiatives and, equally, guide donors in identifying the non-profits they wish to collaborate with.

The data from Giving Balkans is rich and relationally structured, paving the way for graph creation. A unique feature incorporated into the Giving Balkans app is CiviGraph. This tool empowers users to navigate the intricate philanthropy networks built around specific entities in the database (Galjak, 2020). For instance, users can investigate which donors contributed to particular organizations while simultaneously viewing the donors’ immediate philanthropic neighborhood. This involves understanding the other organizations a donor has contributed to and identifying other donors for these organizations. The ability to access and analyze this data offers significant strategic value. Organizations can leverage these insights to identify potential partnerships, optimize fundraising strategies, and better understand the dynamics of the philanthropic landscape. Beyond its strategic value for organizations seeking partnerships and optimizing fundraising strategies, CiviGraph is also a potent tool for investigative journalism tracking donations from politically significant entities, like companies with local or foreign government stakes such as Russian or Chinese, and mapping their philanthropic impact in the region (see Figure 15.2).

### ***1.4 Why was Giving Balkans created?***

The primary motivation behind Giving Balkans was to illuminate the landscape of philanthropy, given the absence of other comprehensive data sources. In the Western Balkans, the regulatory and tax frameworks do not capture any significant data about charitable giving. The only other

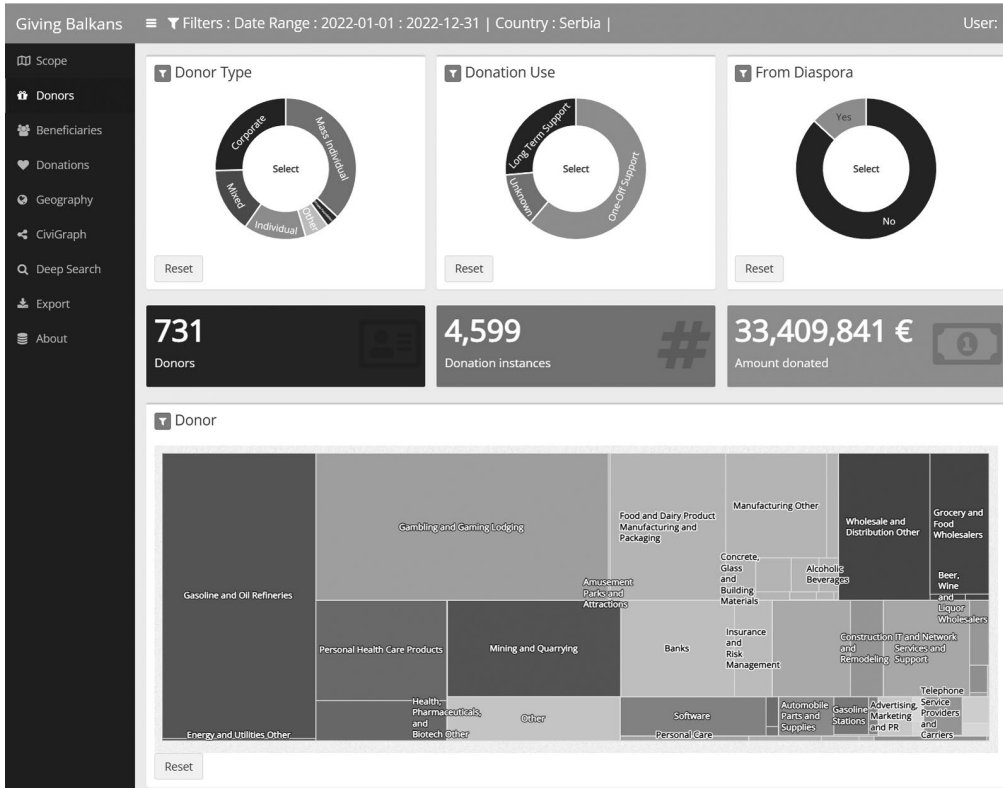


Figure 15.1 Giving Balkans interactive data visualization and analysis web application.

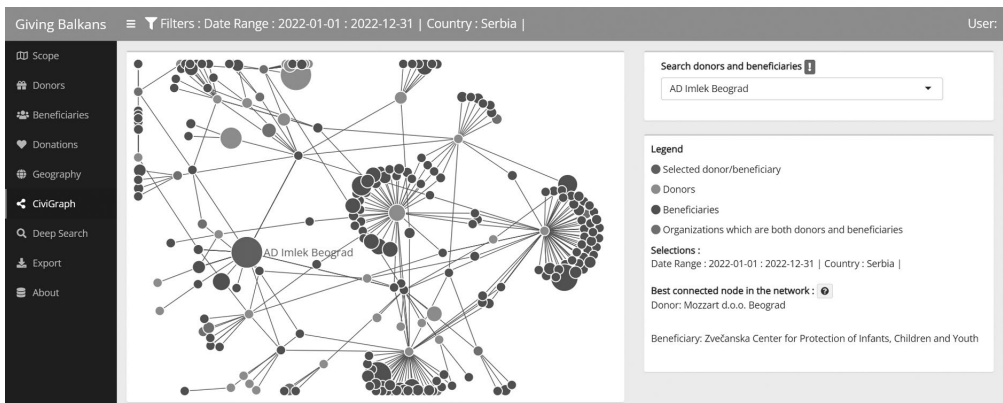


Figure 15.2 CiviGraph—a social network analysis tool built into the Giving Balkans app leveraging the relational data.

glimpses into the region's philanthropic activities came from sporadic ad hoc surveys. There was a prevailing sentiment in the region that philanthropy was either minimal or non-existent and could not serve as a funding source for CSOs. These CSOs have depended on foreign funds to sustain their operations for years. Thus, Giving Balkans primarily aimed to debunk this notion by demonstrating the existence of substantial local resources that CSOs, among other entities, can harness. Beyond presenting a macro perspective of the philanthropic ecosystem, Giving Balkans is important in providing a micro view, offering detailed data points that can be instrumental in optimizing philanthropic giving.

The significance of readily available data became especially evident during the COVID-19 pandemic. The pandemic notably impacted the Western Balkans region (Marinković & Galjak, 2021). Civil society organizations in the region have faced sustainability challenges due to the economic and social disruptions caused by the pandemic (Drobarov et al., 2021). Catalyst Balkans (2020) sprung into action in response to these challenges, assisting 26 different non-profits in launching fundraising campaigns. Leveraging the Giving Balkans database, they helped these non-profits in crowdfunding efforts and locating corporate donors, ultimately raising over 200,000 euros. Throughout the COVID-19 crisis, Catalyst Balkans diligently monitored philanthropic activities in the Western Balkans. Their observations underscored significant contributions, predominantly directed toward essential supplies, and also captured the diverse donor dynamics across various countries (Catalyst Balkans, 2021). Such invaluable data can be harnessed in future crises to identify responsive donors who have previously demonstrated a readiness to contribute promptly.

### ***1.5 Original process of data collection and methodology***

Since its inception in 2013, Giving Balkans has primarily sourced its data from various media outlets, including newspapers, internet portals, television, and radio. Fortuitously, a company specializing in keyword press clipping services was available to cater to all seven countries covered by Giving Balkans. As illustrated in Figure 15.3, the initial methodology required human intervention. Each country's designated Philanthropy Data Analyst would manually process the press clipping data. This press clipping service would consistently forward media records containing predefined keywords (or combinations thereof) set by Catalyst Balkans for each language. These records would then undergo thorough processing by the Philanthropy Data Analysts (Galjak, 2020).

#### ***1.5.1 The problem of actual relevance for Giving Balkans database***

The reliance on a keyword-based approach inevitably led to the inclusion of numerous false positives, which were not pertinent to the Giving Balkans database. For instance, news of a US-based celebrity donating to a charity in an African nation might dominate media outlets, ticking all the keyword boxes. However, such a story is not relevant to the Giving Balkans records. A more specific example of this challenge is the 2022 Russian invasion of Ukraine. The coverage around military donations made to Ukraine by various countries triggered the designated keywords, but these stories held no relevance to the Giving Balkans database.

When evaluating a news item, the Philanthropy Data Analyst must discern whether the article genuinely pertains to the Giving Balkans database. This judgment hinges on the concept of "local philanthropy." The origin of the donor characterizes local philanthropy. Suppose the donor hails from a country within the Western Balkans or belongs to the diaspora of one of the Western Balkans countries, and the donation is intended for a beneficiary in their country of origin. In that case, it is classified as local philanthropy.



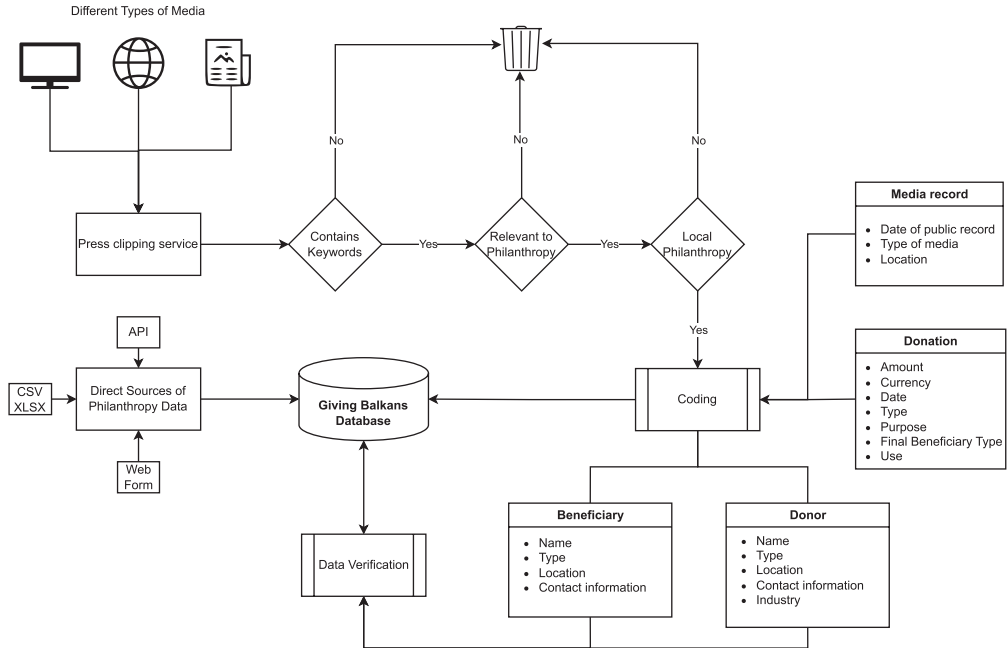


Figure 15.3 The process of collecting data for Giving Balkans database.

### 1.5.2 The problem of languages

As delineated in the introduction, our staff processes media articles in several distinct languages: Serbian, Croatian, Bosnian, and Montenegrin (which are essentially variations of the same language), Albanian, and Macedonian. Consequently, the staff responsible for each country must be proficient in one or more of these languages. This is particularly crucial for Macedonia, which has a significant Albanian minority. Such linguistic requirements make the recruitment of Philanthropy Data Analysts challenging. The work structure at Catalyst Balkans has consistently been organized by language. For instance, an Albanian-speaking staff member might be responsible for covering Albania, Macedonia, or Kosovo.

### 1.5.3 The problem of categories

When media-sourced information is deemed pertinent, it is coded into the database by the Philanthropy Data Analysts. This encoding procedure encompasses several stages. Paramount among these is the categorization of the involved entities (both donor and beneficiary) and the specific instance of the donation. The details about these entities often necessitate additional research, as media articles might not furnish comprehensive data. For instance, an article may state that a local company donated to a neighborhood charity. In such cases, the analyst must ascertain the donor's type (from 11 possible options), identify their industry (from a list of 157 possibilities), and gather relevant contact details, which include address, phone number, email, website, and social media accounts. Similarly, for the beneficiary, the analyst is tasked with pinpointing the type of beneficiary (choosing from 13 options) and documenting analogous contact particulars

as those noted for donors. The donation itself can be categorized in multiple ways, including the category of the donation instance (with ten choices), the type of donation (six options), the purpose of the donation (spanning 26 options), and the category of the end beneficiaries (from 37 available choices). Excluding geographical considerations, like selecting the municipality of the donor and the beneficiary, results in over a billion potential combinations for each donation record. To ensure accuracy and integrity, the Data Quality Manager supervises the whole process, and verifies that all information is correctly coded.

#### *1.5.4 The problem of duplicates*

True originality in news is uncommon, especially given that when a story breaks in one media outlet, it is frequently replicated across many others—a trend particularly evident with online media. This replication introduces challenges in volume; articles, though essentially echoing the same information, often bear distinct, clickbait-inspired titles. This means that those responsible for processing the data frequently find themselves navigating through numerous articles that, content-wise, are virtually identical.

#### *1.5.5 The problem of truth*

Not everything reported in the media is accurate. Hence, verification becomes a pivotal task for Philanthropy Data Analysts. Every recorded donation is cross-checked with at least one of the involved parties, be it the donor or the beneficiary. This ensures that the information we have sourced from media outlets aligns with the facts. At times, this verification leads to updates in our records, be it regarding the donation amount or other details that weren't initially covered in the media reports. Approximately two-thirds of the donations are validated through this verification process, leaving a third unverified.

#### *1.5.6 Data harvesting*

Besides the routine press clipping for data collection, Catalyst Balkans also directly harvests data from available sources. This includes data from Donacije.rs, which, while managed by Catalyst Balkans, only represents a minor portion of the overall charitable giving in Serbia. On the other end of the spectrum, we have direct API access to data from Budi Human (Serbian for *be humane*), an organization dedicated to fundraising for individuals with health challenges, accounting for a significant portion of total donations in Serbia. Additionally, certain companies opt to provide their donation data directly to Catalyst Balkans via email or web forms integrated into the Giving Balkans website.

### ***1.6 Limitations of the methodology***

Given the lack of consistent sources for assessing charitable donations in the Western Balkans, Catalyst Balkans has adopted innovative data collection methods. These are primarily based on print, online, and electronic media and are supplemented by other available data platforms. However, this approach has limitations: not all philanthropic actions are highlighted in the media, and the published reports often lack the necessary details to understand philanthropy trends fully. Beyond these media-centric strategies, the Giving Balkans database uses direct data channels. Some organizations, for instance, provide firsthand access to their donation data via Application

Programming Interfaces (APIs) or by regularly sharing spreadsheet files. While this direct method simplifies data integration and bolsters accuracy, it comes with challenges, like reliance on third parties and potential inconsistencies in data. While our data might not capture the entire landscape, it does establish baseline figures, indicating the minimum number of events, financial contributions, and participants recorded annually. These figures offer a foundational perspective on the basic level of philanthropic activities in a country. One of the main challenges is tracking the growth of philanthropy in an environment with sporadic data collection and inconsistent examination. To tackle this, Catalyst Balkans has developed a set of preliminary criteria to shed light on the various aspects of charitable donations. These cover charitable events or drives, financial collection methods, guiding donation principles, recipients and beneficiaries of donations, the donors, stakeholders, and the extent of media coverage. Currently, quantitative and qualitative metrics linked to each criterion offer a solid framework for assessing the philanthropic terrain of a nation over several years. Regarding data reliability, the Giving Balkans database is updated daily, reflecting a consistent commitment to accuracy and timeliness.

## **2 AI-assisted process of data collection**

Ever since 2017, Catalyst Balkans has looked for ways to automate some of its processes around collecting and processing philanthropy data. Replacing Philanthropy Data Analysts seemed like an impossible task. Increasing analysis processing costs and the volume of press clipping data finally pushed us to develop the AI-assisted data collection and analysis process. The idea is to maximize Philanthropy Data Analysts' productivity instead of creating fully autonomous agents that would replace them. The question was how to achieve this: by addressing the two major problems—false positives and duplicate articles (see details in the Resulting Solution section).

To this end, we have created a system for preprocessing articles that solves these problems and significantly increases Philanthropy Data Analysts' productivity (Figure 15.4). Our system utilizes custom, language-specific models instead of relying on pre-trained multilingual models (such as RoBERTa). We initially opted for xlm-roberta-base, a general model that needed to be fine-tuned for the downstream task of text classification. It seemed perfect as it covered all the languages of the Western Balkans. However, we have failed to make it work for this task using our train and test datasets in Serbian language only. An additional obstacle was that we needed to handle this fine-tuning on the Azure cloud platform, which presented additional costs. We opted for custom-trained language-specific models after initial testing showed very promising results. However, we acknowledge the potential long-term benefits of fine-tuning multilingual pre-trained models. Investing more resources into this approach in the future could yield a more robust and scalable solution suitable for a broader range of languages and tasks. Therefore, revisiting and refining the strategy of utilizing multilingual pre-trained models remains a consideration for our future work.

### ***2.1 Text processing***

In our data collection and analysis process, we initially extracted data in batches of a thousand articles. For precision, every article undergoes an exhaustive, fully autonomous cleaning procedure consisting of two steps.

In the first step, poorly OCR-ed (optical character recognition) articles from print media are detected and flagged as such using PCA-based outlier detection on character-level n-gram ( $n = 1,2,3$ ) based on Term Frequency–Inverse Document Frequency algorithm. This was necessary as some articles, particularly those digitized from print sources, might contain errors from imperfect

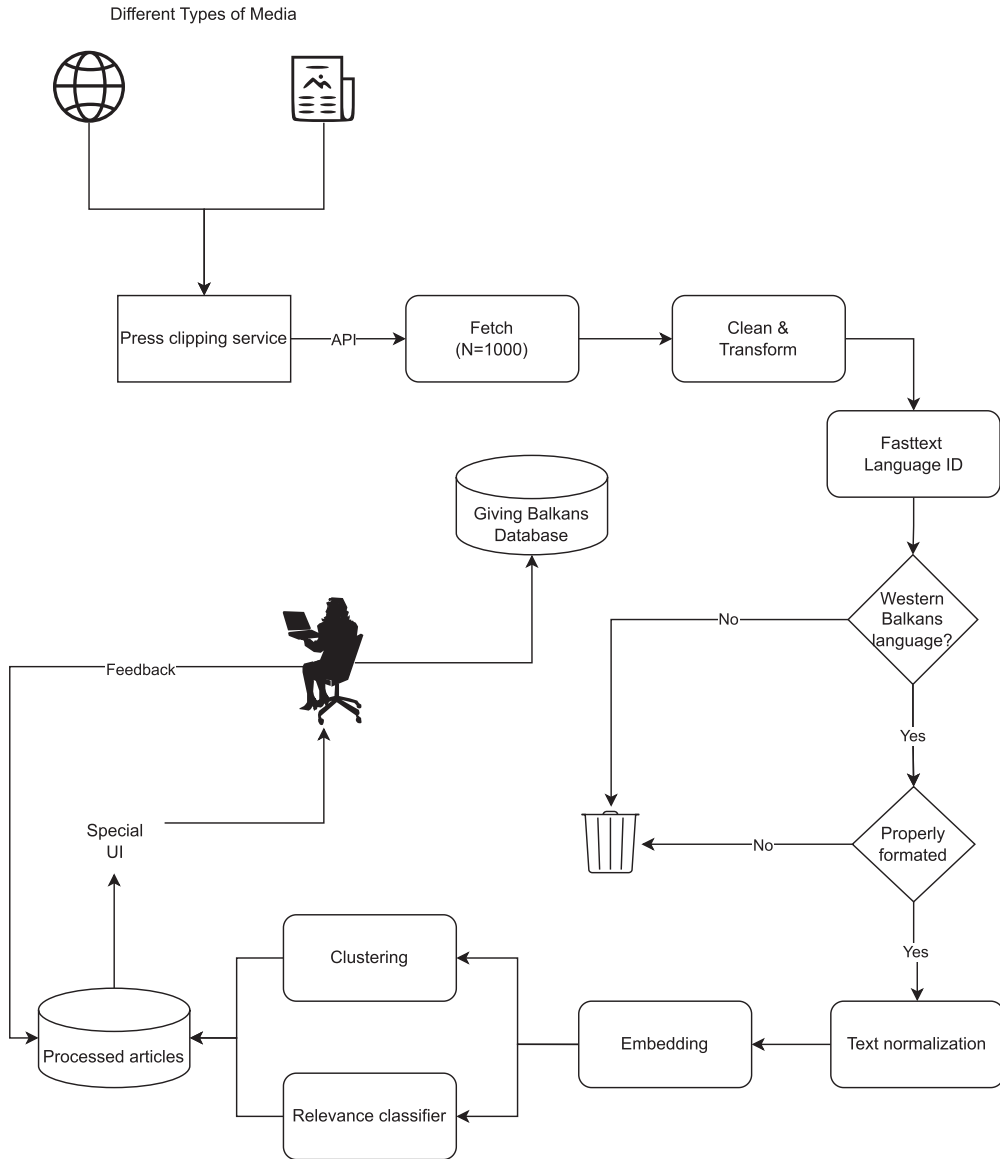


Figure 15.4 The AI-assisted process of collecting data for Giving Balkans database.

scanning techniques. Identifying and flagging such articles was required to ensure their errors did not skew our analysis.

In the second step, the text is normalized. First by Unicode Normalization Form: Compatibility (K) Composition, then articles in Serbian Cyrillic are transliterated into Latin, repeated horizontal whitespaces are eliminated, and line endings are normalized (“\r\n,” “\n\r,” “\r” to “\n,” etc.).

Once these adjustments are made, we structure each article into distinct paragraphs and sentences, facilitating smoother processing and analysis. We discovered that depending solely on the language

information from the source could be misleading. Hence, we employ a specialized model known as fastText (Joulin, Grave, Bojanowski, Douze, et al., 2016; Joulin, Grave, Bojanowski, & Mikolov, 2016) to accurately determine the article’s language. Subsequently, we standardize each article’s text.

## **2.2 Embeddings**

In the expansive realm of AI-powered textual analysis, embeddings stand out as a refined tool for distilling the core meaning of documents and words by converting them into numerical vectors. This mathematical portrayal enables computers to identify patterns, themes, and resemblances across extensive datasets.

### *2.2.1 Adopting Doc2vec and Word2vec*

To process the vast array of articles from the Western Balkans, we utilized the Doc2vec (Le & Mikolov, 2014) and Word2vec (Mikolov et al., 2013) models from the Gensim library (Rehurek & Sojka, 2011). These models are acclaimed for their effectiveness, even with limited hardware resources. Our goal was to represent each news article as a vector—a feature of Doc2vec. However, we needed to train Word2vec models for each respective language to establish the multidimensional vector space for these representations. We began by training our Word2vec models on comprehensive Wikipedia corpora for Serbian/Bosnian/Croatian/Montenegrin, Macedonian, and Albanian languages. With these broad-scope embeddings in hand, we then tailored our system using our specific collection of news articles from 2015 to 2021. This allowed us to train our Doc2vec models, making them particularly attuned to the context of philanthropy, as highlighted by Lau and Baldwin (2016).

### *2.2.2 Article processing and model parameters*

Before training, articles were processed and standardized: they were separated by language, deduplicated, broken down into paragraphs, and cleansed of punctuation and numerals. Notably, stemming and lemmatization were not employed due to their negligible impact on our pilot evaluations. For the technically inclined, our Doc2vec training used the PV-DBOW model variant alongside simultaneous skip-gram Word2vec training, with a 15-word context window and a 300-dimension vector embedding. We cycled through this training for a comprehensive 100 epochs.

### *2.2.3 Topic-based clustering and modeling*

Beyond basic embeddings, we integrated Top2vec (Angelov, 2020), a cutting-edge tool for topic modeling. By default, Top2vec offers robust results through an intuitive API, autonomously handling hyperparameters and determining distinct topic counts. With just a few lines of code, Top2vec can train document and word embeddings, reduce dimensionality using uniform manifold approximation and projection—UMAP (McInnes et al., 2018), cluster these reduced vectors, and determine topic vectors. This was invaluable for static collections. However, our continuous inflow of articles presented challenges. Our solution was to bypass Top2vec’s default training and instead employ our pre-trained Doc2vec embeddings. This adaptation permitted us to cluster articles in digestible batches daily, maintaining a consistent semantic vector space. As we forge ahead, we aim to develop a mechanism that seamlessly interlinks these daily clusters, tracing the narrative arc of news topics over time.

#### *2.2.4 Relevance classifier*

To effectively process the vast influx of news articles, we needed an automated system to quickly identify which articles were related to philanthropy and which were not. This was paramount, as an overwhelming majority—almost two-thirds—of the articles we received had no direct connection to our area of interest.

To achieve this, we embarked on a meticulous two-month project. We collected a large number of articles in the Serbian/Bosnian/Croatian and Albanian languages. Each article was carefully labeled as either relevant to philanthropy or irrelevant. For instance, in the Serbian/Bosnian/Croatian dataset, out of 26,045 articles, 13,425 were deemed relevant, while 12,620 were deemed unrelated to our focus.

While the irrelevant articles were primarily identified and labeled manually, the relevant articles were more straightforward to gather. We did this by cross-referencing with our pre-existing database of donations.

To translate this labeling effort into an actionable system, we utilized specialized mathematical models—vector embeddings—from our Doc2vec models. Combined with our labeled articles, these embeddings allowed us to train and evaluate several methods to automatically classify incoming articles from the scikit-learn library (Pedregosa et al., 2011). After testing various methods, we found the most success with an algorithm known as the support vector machine classifier (Chang & Lin, 2011; Platt, 1999) with a polynomial kernel of degree 3. In evaluating the algorithm's performance in classifying relevant articles, it showed remarkable efficiency in various languages. For the Serbian/Bosnian/Croatian language, the classifier exhibited a precision of 96.0%, a recall of 95.4%, and an overall accuracy rate of 95.6%. Similarly, when applied to Albanian articles, the classifier achieved a precision of 99.0%, a recall rate of 91.4%, and an accuracy of 95.63%.

In simpler terms, our system became exceptionally adept at sorting through heaps of news articles and pinpointing those relevant to philanthropy, all while requiring minimal human intervention.

### **2.3 Resulting solution**

The resultant solution was coupled with a tailored user interface designed specifically for these new AI-assisted functionalities. Collectively, these modifications spurred significant productivity enhancements. The issue of false positives, where Philanthropy Data Analysts were inundated with media articles unrelated to philanthropy, can now be promptly identified and labeled as irrelevant. This is especially beneficial since this data informs subsequent training phases. The essence of semantic clustering allows for bulk categorization of articles about the same topic, whether pertinent or not. In the past, such articles had to be addressed individually. Now, a Philanthropy Data Analyst is presented with these clusters, accompanied by a probability score indicating the relevance of a given group. This ensures that the most pertinent clusters are prioritized and tackled first.

## **3 Future AI integration**

With a notably enhanced workflow and heightened productivity, the logical progression is to ask whether the role of Philanthropy Data Analysts could be eliminated. While complete substitution using the current methodology is impossible, revising the methodology to fit the capabilities of the available AI technologies is worthwhile considering. Given the substantial expenses

associated with human labor, even in the Western Balkans' middle-income nations, relying solely on an AI-assisted approach may not be a viable long-term strategy. Two primary avenues exist for achieving full autonomy.

The first method entails a profound transformation of our current methodology to better align with the AI-driven capabilities at our disposal. This could involve various modifications, ranging from simplifying specific classifications to reconfiguring how we calculate the aggregate donation sum for a nation. For the elimination of the role of Philanthropy Data Analyst, the changes in methodology would need to be radical. Rethinking this role would probably be more realistic as no matter how the methodology changes, a human will always have to be in the loop.

The second strategy suggests supplanting Philanthropy Data Analysts with AI agents underpinned by services offering API access to generative large language models (LLMs) like ChatGPT. Fundamentally, some tasks within a Philanthropy Data Analyst's purview could be deconstructed into discrete operations that a powerful model like ChatGPT-4 could effectively manage. Our preliminary experiments with ChatGPT-3.5 turbo—a cost-effective choice provided by OpenAI via its API—indicate its robust capacity. When provided with appropriate prompts, it can adeptly categorize donation instances. The rapid progress of these models, translating to cheaper and evermore capable agents, makes this a promising avenue for future implementation. However, although the AI agents could be performant and effective, the solution is far from substituting the tasks and domain expertise of Philanthropy Data Analysts.

In a practical scenario, achieving further automation would likely necessitate a hybrid of both approaches. This means radical methodology changes with human (domain expert) oversight.

## **4 Problems with AI**

### **4.1 Cost of AI**

The primary benefit is the potential reduction in costs. ChatGPT costs depend on the specific model used (whether ChatGPT-3.5 turbo or ChatGPT-4) and the context window, ranging from 4,000 to 128,000 tokens. The price fluctuates between \$0.002 and \$0.03 per 1,000 tokens. Thus, classifying donation instances using OpenAI's API could cost anywhere from a few cents to \$3.84 for each donation instance. This calculation assumes that media articles related to a particular donation have been collated and deduplicated before classification. To illustrate, Serbia had 4,557 recorded donation instances. If all these were processed using the most advanced OpenAI model with the largest context window, the cost for just this one country would be approximately \$17,498.88 annually. Compared with the average gross salary in Serbia, which stands at \$12,873.72—with a considerably lower median figure (Statistical Office of the Republic of Serbia, 2023)—this method appears less cost-effective. However, opting for the less advanced OpenAI GPT-3.5 Turbo model with a 16,000-token context window would incur a cost of only \$218.7, presenting a far more economical alternative.

Utilizing a standalone service from one of the open-source LLMs based on Llama 2 (an open model released by Meta) or its derivatives could be considerably costly, especially considering the cloud resources required by its 70-billion-parameter version or even more cost-effective options (such as quantized variant such as Vicuna with 13 billion parameters). This does not even account for the costs of building, maintaining, and upgrading the system. While there are hosted services that offer API access, their charges are often on par with, if not exceeding, those of OpenAI's GPT-3.5 Turbo. OpenAI has recently released GPT-4 Turbo, which has a larger context size and, more importantly, is cheaper than GPT-4. The question of cost and whether API access to

proprietary or self-hosted open-source models is more affordable depends on each option's capabilities. The open-source alternative could be cheaper if smaller models show a similar level of capability as the proprietary models. For the time being we have not tested any of the open-source models, but benchmarking several different models for our use case against ChatGPT would be straightforward. Given the rapid advancement and the size of the community gathered around these open-source models, it will likely be a cost-effective and performant option in the future if it is not already.

Equally significant are the human resources expenses tied to the development of these infrastructures. While Catalyst Balkans crafted its system in-house, this endeavor redirected resources from other essential programs and services. Most non-profits typically lack the in-house capabilities even to construct their websites, much less having data scientists and engineers readily available. Assembling a team to create—and continuously develop and maintain—such a system could entail hefty expenditures.

#### ***4.2 Rapid innovation and changes to LLMs***

Pipelines are susceptible to disruptions when models change. Even if these modifications generally enhance the models, each update necessitates a reassessment of the AI pipeline in place, which might need adjustments to align with the revised model. OpenAI has rolled out “frozen” models, which are updated less frequently compared to their regular counterparts. Nevertheless, even these are not guaranteed to have indefinite availability.

#### ***4.3 Integration of the knowledge***

The role of Philanthropy Data Analysts at Catalyst Balkans extends beyond merely processing media articles, coding donation instances, and verifying them. Over time, these analysts have evolved into experts on charitable giving within their respective countries. While we provide data access as a service to our ecosystem, we frequently encounter inquiries that only someone deeply versed in the nuances of giving within a specific country can address.

Furthermore, these analysts have cultivated an in-depth understanding of entities in their designated countries. Their knowledge transcends mere reportage, equipping them to piece together narratives and make informed decisions based on often scant media information. This depth of understanding is further enriched by their connections within these countries, particularly their communication with individuals affiliated with donor and beneficiary institutions.

### **5 Insights and implications**

The shift from the original data collection process to an AI-assisted approach at Catalyst Balkans marks a significant evolution in our handling of philanthropic data. This transformation exemplifies the transition from traditional methods to technologically advanced techniques, highlighting the efficiencies and effectiveness of AI integration. In comparison, our new process helped us manage the problem of false positives. AI's precision in filtering irrelevant data drastically reduced the burden of sifting through irrelevant content, a significant challenge in the original method. Additionally, our implementation for each language enabled streamlining the data collection process across different linguistic contexts, overcoming a major hurdle of the traditional approach. Most notably, the AI-assisted method markedly improved time efficiency and resource utilization,



addressing the labor-intensive nature of manual data processing. Beyond operational efficiency, AI integration has led to the discovery of new insights and strategic possibilities in philanthropic data collection.

Moreover, the efficiency gained using the AI-assisted approach that we have developed allows for handling the increasing data volumes without a corresponding increase in resources. This case study serves as a guide for other non-profits contemplating AI integration. Embracing AI can effectively address longstanding challenges in data management and processing. While initially resource-intensive, custom AI solutions can offer substantial long-term benefits in terms of efficiency. Maximizing AI benefits involves using it as an enhancement to, rather than a replacement for, human expertise, at least at this moment, which might not hold if AI continues to progress at an ever-increasing pace. Catalyst Balkans' experience in integrating AI can offer some insights into the digital transformation of the philanthropic sector globally, encouraging a shift toward more data-informed strategies in philanthropy and demonstrating how digitalization, through AI, can optimize operations, reduce costs, and amplify impact in the philanthropic sector worldwide. This global perspective underscores the potential of AI not just in the Western Balkans but in the broader philanthropic landscape.

## **6 Conclusion**

The pursuit of leveraging AI to assist in the philanthropy data collection process at Catalyst Balkans showcases the synergistic potential between cutting-edge technology and traditional data management practices. The Giving Balkans database embodies the synthesis of manual efforts with technological solutions. This integration not only streamlines the data collection process but also enhances the quality and comprehensiveness of the data, revealing a broader and more nuanced perspective of the philanthropic landscape in the Western Balkans. Our piecemeal strategy of incorporating AI components allowed us to retain human oversight while still benefiting from automation. This hybrid model provided a balanced solution that addressed resource constraints, ensured data accuracy, and facilitated scalability. Even as we have made significant strides in optimizing the data collection process, the landscape of AI remains dynamic. The ever-evolving capabilities of large language models and declining costs present exciting prospects for future iterations of the Giving Balkans data collection process. The journey was not devoid of hurdles. From managing false positives to handling data veracity, the imperfections of AI compelled us to refine our methodologies constantly. Moreover, the financial implications and rapid technological changes associated with AI underline the need for organizations to remain agile and forward-thinking. As we move closer to achieving a completely autonomous process, the ethical implications of such a system cannot be understated. The balance between automation and human judgment will ensure that the Giving Balkans database remains a reliable and unbiased resource. The successes and challenges encountered in this case study hold broader implications for the global intersection of AI and philanthropy. As technologies become more accessible, there is an opportunity for organizations worldwide to harness them to capture philanthropic trends, understand donor behaviors, and ultimately foster a culture of giving. In conclusion, the journey of integrating AI into the data collection processes at Catalyst Balkans is emblematic of the broader narrative of technological transformation in the non-profit sector. As we navigate the complexities and promises of AI, it becomes evident that the convergence of human expertise and machine intelligence can pave the way for a more informed, efficient, and impactful future in philanthropy.

## Notes

- 1 Serbia, Croatia, Bosnia and Herzegovina, Albania, Macedonia, Kosovo, and Montenegro.
- 2 <https://donacije.rs> is a crowdfunding website run by Catalyst Balkans, where non-profits can put up their campaigns and crowdfund for specific projects. Catalyst Balkans guides non-profit campaign owners throughout the process—helping them formulate the campaign, create content, and teach them how to fundraise.

## References

- Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics*. <https://doi.org/10.48550/ARXIV.2008.09470>
- Bajra, U. Q. (2021). The interactive effects of remittances on economic growth and inequality in Western Balkan countries. *Journal of Business Economics and Management*, 22(3), 757–775. <https://doi.org/10.3846/jbem.2021.14587>
- Brinkerhoff, J. M. (2014). Diaspora philanthropy: Lessons from a demographic analysis of the Coptic diaspora. *Non-Profit and Voluntary Sector Quarterly*, 43(6), 969–992. <https://doi.org/10.1177/0899764013488835>
- Catalyst Balkans (2020, March 25). *Donacije.rs—COVID-19*. <https://www.donacije.rs/covid19/>
- Catalyst Balkans (2021, November 30). *Giving Balkans: Philanthropy's Response to COVID-19 (September 30, 2021)*. Giving Balkans. <https://givingbalkans.org/content/giving-balkans-philanthropy%E2%80%99s-response-covid-19-september-30-2021>
- Catalyst Balkans (2023). *Giving Balkans Database on Philanthropy in the Western Balkans*. <https://giving-balkans.org/>
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1–27. <https://doi.org/10.1145/1961189.1961199>
- Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *Shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>
- Drobarov, R., Popovska, B., & Gelev, I. (2021). Impact of COVID-19 on sustainability of civil society organizations in the Western Balkan Region. *Bezbednost, Beograd*, 63(3), 57–76. <https://doi.org/10.5937/bezbednost2103057D>
- Duncan, D. (2016). Language policy, ethnic conflict, and conflict resolution: Albanian in the former Yugoslavia. *Language Policy*, 15(4), 453–474. <https://doi.org/10.1007/s10993-015-9380-0>
- EWB. (2023, May 24). Freedom house: Democratic institutions in the Western Balkans continued to falter in 2022. *European Western Balkans*. <https://europeanwesternbalkans.com/2023/05/24/freedom-house-democratic-institutions-in-the-western-balkans-continued-to-falter-in-2022/>
- Freedom House. (2016). *Back Where We Started in the Balkans*. Freedom House. <https://freedomhouse.org/article/back-where-we-started-balkans>
- Galjak, M. (2020). Dva primera upotrebe teorije grafova u društvenim naukama: Analiza interakcija na Tviteru tokom izbora 2016. U Srbiji i analiza GivingBalkans podataka o filantropiji na Zapadnom Balkanu. In V. Mentus & I. Arsić (Eds.), *Promišljanja aktuelnih društvenih izazova: Regionalni i globalni kontekst* (pp. 232–251). Institut društvenih nauka.
- Glenny, M. (2001). *The Balkans: Nationalism, War, and the Great Powers, 1804–1999*. Penguin Books, New York.
- Grødeland, Å. B. (2006). Public perceptions of non-governmental organizations in Serbia, Bosnia & Herzegovina, and Macedonia. *Communist and Post-Communist Studies*, 39(2), 221–246. <https://doi.org/10.1016/j.postcomstud.2006.03.002>
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing Text Classification Models*. <https://doi.org/10.48550/ARXIV.1612.03651>
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification*. <https://doi.org/10.48550/ARXIV.1607.01759>
- Lau, J. H., & Baldwin, T. (2016). *An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation*. <https://doi.org/10.48550/ARXIV.1607.05368>
- Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR, abs/1405.4053*. <https://doi.org/10.48550/arXiv.1405.4053>
- Marinković, I., & Galjak, M. (2021). Excess mortality in Europe and Serbia during the COVID-19 pandemic in 2020. *Stanovništvo*, 59(1). <https://doi.org/10.2298/STNV2101061M>

- McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. <https://doi.org/10.48550/ARXIV.1802.03426>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. <https://doi.org/10.48550/ARXIV.1301.3781>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10, 61–74.
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rehurek, R., & Sojka, P. (2011). Gensim—python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2.
- Statistical Office of the Republic of Serbia (2023, August 25). *Average Salaries and Wages Per Employee, June 2023*. <https://web.archive.org/save/https://www.stat.gov.rs/en-us/vesti/statisticalrelease/?p=13675&a=24&s=2403?s=2403>
- Transparency International (2020, December 11). *Captured States in the Western Balkans and Turkey—News*. Transparency.Org. <https://www.transparency.org/en/news/captured-states-western-balkans-turkey>
- USAID. (2023). *Civil Society Organization Sustainability Index (Reports)*. FHI 360. <https://www.fhi360.org/sites/default/files/media/documents/csosi-europe-eurasia-2021-report.pdf>
- World Bank (2023). *GDP Per Capita (Current US\$)—Kosovo, Serbia, Croatia, North Macedonia, Albania, Montenegro, Bosnia and Herzegovina* [dataset]. <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=XK-RS-HR-MK-AL-ME-BA>

# 16

## OPTIMIZING PHILANTHROPIC INVESTMENT WITH AI

### A case study of the Altruist League

*Milos Maricic*

#### 1 Systemic change and its challenges

##### *1.1 The field of philanthropy has undergone two major transformations in the 21st century*

The first one was made possible by technology. Emerging technologies opened new data collection and analysis avenues, radically altering how charitable organizations and philanthropists made investment decisions (Valley, 2020). We can now closely follow grantees' performance along an array of KPIs. We can track donor portfolios in real time. Many organizations adopting this quantitative, KPI-driven approach have had sustained success (Glassman, 2016). The second transformation has been more recent and is still ongoing. It has been the reaction to a growing belief that philanthropy and humanitarian action, more broadly, are not performing as well as they could be (Polman, 2010). There have been two roots of this criticism:

- 1 In some cases, foundations and humanitarian organizations have been working in the same subject area for decades, supporting the same organizations in the same country with the same strategies, without a material improvement in the state of affairs (Moyo, 2010; Banerjee & Duflo, 2011).
- 2 Some authors have drawn attention to perceived hypocrisy in modern philanthropy. According to them, many among the wealthy do philanthropy for public recognition and tax advantages (Giridharadas, 2019). Their giving is a marketing exercise to keep public anger at bay; other parts of their portfolios might be creating tremendous damage in the world, for example, through fossil fuel investments (Edwards, 2010).

The relevance of this criticism for individual foundations and donors varies greatly. But a reaction among at least a portion of them has been to adopt a broader, **systemic** approach to giving based on **trust** between them and the partners they support, involving a more equitable sharing of power (Reich, 2018; Villanueva, 2018).

The principles of this emerging discipline of systemic change are best depicted when contrasted with the traditional approach to philanthropy. They are summarized in Table 16.1.

Table 16.1 The principles of systemic philanthropy (adapted from Fernandez et al., 2020)

<i>Principle</i>	<i>Traditional philanthropy</i>	<i>Systemic philanthropy</i>
Active sourcing	Grantees draft grant applications. Donors then pick and choose.	Donors scan the environment actively for investment opportunities. This lets partners focus on their mission instead of spending resources on fundraising.
Diversification	Investment is made into isolated projects and narrow groups of grantees to reduce complexity and risk.	Diversified investment into a broad range of organizations working on an issue, often without specific projects in mind.
Commitment	Grantees face constant KPI pressure to keep performing; investments are short term to maintain focus. The accent is on the procedure.	Investments are multi-year. Partners are trusted to use funds in the best way. The focus is on the relationship.
Active tracking	Grantees create reports that talk about metrics such as meetings and workshops held or awareness raised.	Donors themselves do the tracking and focus on tangible improvements: membership growth, people served, citations in Tier 1 media, legal cases and pieces of policy influenced, and sentiment changed (as measured by surveys).
Alliance-building	Focus on own mission and priorities to reduce complexity.	Understanding that the most difficult problems require diverse groups to be part of the solution, from the government to the civil sector to other foundations. Accent on collaboration and consensus-building.
Ongoing funding decisions	Investment committees make grant decisions a few times per year.	Funding is disbursed constantly; the decisions are kept at the analyst level, often with no management input.
Embracing grassroots action	Staying away from social movements because they can be fleeting, political, or unable to follow reporting requirements.	Seeing nonviolent activist grassroots movements as the engine of social change, investing in them liberally, and trusting them to do their work.
Embracing uncertainty	Seeking to minimize risk.	Seeking to maximize impact, seeing risk as unavoidable. Understanding that some investments will not perform. A “venture capital” mentality.
Root problem focus	Tendency to invest in uncontroversial, easily definable areas such as poverty, hunger, and children’s rights.	Believing that the root problems of societies tend to be political: poor governance, inappropriate climate policy, systemic inequality (opportunity, gender, racial), and threats to democracy. Invest accordingly.

## 1.2 About the Altruist League

The Altruist League was founded in 2015 with the specific goal of helping donors put in place systemic change strategies using cutting-edge technology. Two beliefs were central to the League’s four founding partners.

One of the beliefs was that true change in fields such as climate action, human rights, and food security was a complex undertaking that required a coalition of citizens, policymakers, funders, the media, think tanks, academic institutions, the government, and other organizations. The other was that a small group of foundations, high-net-worth individuals (HNWIs), and other organizations, frustrated by the status quo in philanthropy, were willing to pay a premium for advice on a new generation of investment strategies that actually moved the needle on the world's most pressing problems.

### **1.3 Grassroots movements**

The key component of the League's approach was grassroots movements, loosely defined as more or less formal groups of citizens with a lived experience of a problem actively working to solve it. The range of such organizations worldwide is breathtaking. They provide safe houses for women in Juarez,<sup>1</sup> fight for paternity leave in Switzerland,<sup>2</sup> face down the fossil fuel lobby in Washington, DC,<sup>3</sup> and risk their lives for democracy in Eastern DRC.<sup>4</sup>

Despite their ubiquitousness and often striking effectiveness, citizen movements are avoided by the average philanthropic foundation for understandable reasons: they can have informal hierarchies, change focus often, and might disappear overnight. This makes them risky (Almeida, 2019).

However, what they may lack in formality, citizen movements often offer in terms of authenticity and the ability to grow and impact the world. Philanthropy also has a long tradition of working with them. One need only recall that, in the 1950s and beyond, the Ford Foundation actively funded the civil rights movement, the quintessential grassroots organization, with tremendously positive effects on American society (O'Connor, 1999).

The League was determined to invest in grassroots organizations that were available, easy for donors, and beneficial to the organizations involved. Moreover, it aimed to facilitate advanced donor-partner matching and dynamic portfolio calibration based on new data, emerging global events, or changing preferences. This was a very ambitious proposition. Step one was to build a dataset.

## **2 Constructing a robust dataset**

The League began by hiring a team of Geneva-based analysts. They investigated available open-source datasets for suitable organizations and verified the data. Such databases were often entirely outdated and served, at best, as pointers for further exploration. In their work, analysts liaised with organizations over the Internet and by telephone in order to confirm information and improve the dataset. Progress was very slow at first.

The analyst team expanded quickly and to many countries. This provided several benefits. First, locally embedded analysts could spend time face-to-face with organizations in the dataset, grasping their mission and objectives more clearly. They spoke the local language and understood the local culture. Second, hiring local analysts supported the local economy while reducing overall costs for the League.

In the first few months, the League hired staff in more than 30 countries, including in regional offices in Hong Kong, San Francisco, London, and Nairobi.

Over time, the team developed a working definition of which types of organizations had a place in the dataset and which did not. These ended up being the criteria for inclusion (see Table 16.2).

The parameters the analysts tracked for each organization are listed in Table 16.3.

Table 16.2 The criteria for inclusion of an organization in the Altruist League’s dataset

<i>Criterion</i>	<i>Rationale</i>
Nonviolence	This was a non-negotiable criterion. Moral reasons aside, it has been well documented that nonviolent activism statistically achieves better results in the long run than the violent kind (Chenoweth & Stephan, 2013).
Focused on primary problems (climate, corruption and democracy, equality of economic opportunity, empowerment of women, racial justice)	The League believed that problems like hunger, education, health, and poverty, while critical, were: (1) in the long run caused by underlying problems of a given society, namely corruption, lack of democracy, inequality, and/or climate change. (2) already extensively supported by conventional philanthropy, CSR initiatives, and efforts of major international organizations.
Ground-led	Analysts typically excluded organizations whose leadership and board were exclusively Western, removed from the operational context, and without representation of targeted groups.
Democratic and with meaningful participation of women (or women-led)	It was very rare for the League to include, even in its broader dataset, autocratically led, male-only organizations unless this was, for some reason, their point (e.g., men’s organizations offering peer counseling in the context of reducing violence against women).
Political and progressive	Systemic problems require political positions, at least implied ones. The League excluded organizations advocating, for example, more fossil fuel use or fewer rights for LGBT people.
Across the value chain	While the dataset primarily included activist movements (95%+), it also cataloged independent media organizations and lobbying groups contributing to the aforementioned systemic change value chain. It also included training organizations and incubators that teach activists how to operate and be impactful.
With a recent track record of success, irrespective of historical performance	Many social movements bureaucratize and become less effective over time. The team avoided old organizations without a recent track record of impact.

Some of the data points were simple to capture and unchanging. Others left a lot to the analyst’s interpretation and needed constant updating. As a consequence, scaling became a problem rapidly. On average, one analyst could manage and update records of about 50–100 organizations. When the number of organizations in the dataset exceeded several thousand, the League needed to improve analyst productivity by an order of magnitude.

As discussed throughout this chapter, technology was the primary solution to this challenge. Nevertheless, a smart division of the analyst’s time also helped. It quickly became apparent that some movements were inherently more attractive to donors than others because of unique characteristics and/or achievements. Analysts then focused their attention on those, reducing the attention given to the tail end of the dataset. This focus could, of course, shift in case events or donor interest dictated it.

Soon, the League found it worthwhile to preselect the “best” organizations in the dataset and combine them into the “Altruist Index™,” echoing for-profit index funds aimed at passive retail investors. Philanthropists who did not have a particular preference for donations or geographical areas could invest in the Altruist Index and be confident that they were maximizing the impact of their money. The Index was a constantly changing collection of the highest-performing

*Table 16.3* Parameters tracked in the Altruist League's dataset

---

Name
Founding year
Region
Topic
Leadership structure (in particular, the percentage and profile of women)
Number of members and growth metrics
Strategy and tactics
Measurable goals
Track record of success in influencing policy
Alliances
Membership composition (age, sex, geographic distribution)
Minority representation
Website and social media account metrics and growth
Structure
Any media partnerships
Existing funders profile
Amounts of funding received and dates
The amount of funding the organization can successfully absorb

---

organizations in the League's dataset, usually about 150. They usually had one or more of the following characteristics to differentiate them:

- Impressive track record of influencing policy;
- Strong membership growth;
- Exceptional member engagement;
- Repeated citations in Tier 1 media;
- Legal cases influenced or won;
- Track record of changing population sentiment within the area of operation;
- Replicability of model (strategy, tactics);
- Positive message and ability to build alliances in broad social groups rather than antagonize.

While the feedback from the early clients was enthusiastic, the League was far from being a successful business at this point. The cost of maintaining the dataset was enormous compared to the relatively modest income from being effectively just another philanthropy consultancy in a crowded market. The leadership embarked on a critical strategic decision: to deploy artificial intelligence to improve the dataset.

At the time, this gamble was counterintuitive in many ways. It was going to increase the cost of operations even further. It would oblige the League to propose price points an order of magnitude above its competitors. The gains from technology were far from certain. However, that was the decision that changed everything.

### **3 The AI model for donor-partner matching**

The League's vision was to have a system that nearly autonomously kept track of hundreds of thousands of organizations, their impact on the world, and their potential in real time.



In the early days, this was firmly out of the domain of the possible. It was difficult enough to form a machine learning (ML) team.

The initial strategy was to hire people with ML PhDs and trust them to build the model and set up the related processes. This did not work for several reasons:

- The PhDs typically understood little about philanthropy or citizen action;
- The strategies their models suggested tended to be the obvious ones because of a tendency to be risk-averse and overfit;
- The power imbalance between the tech team and everyone else was substantial, and communication was poor.

In hindsight, none of this should have come as a surprise, and these issues are well documented in financial machine learning literature (López de Prado, 2018).

After a few false starts, a better strategy for AI team formation emerged. Its tenets were:

- Understanding the business and its dynamics became essential for any hire;
- The objective was no longer to find one “optimal” model or investment strategy but instead to find a process with which such strategies could reliably be churned out and quickly tested;
- Rather than looking to hire ML “gurus,” the League took fairly technical people and helped them become experts in one aspect of the process—data sourcing, data cleaning, hypothesis generation, model training, testing, validation, and so on.

Soon after beginning to work earnestly, the team realized that the initial lofty expectations needed to be revised. As a first step, they settled on a more modest, relatively static task: improving donor-grantee matching. The specific KPIs for this were:

- Reduce time spent identifying potential partner pairs by at least 80%, effectively automating the process;
- Achieve over 80% funding rates for the matches delivered;
- Maintain donor and nonprofit satisfaction at over 95%;
- Reach 90%+ accuracy in predicting successful matches, as evaluated by donors and grantee organizations.

Accuracy and satisfaction metrics were expected to gradually improve over time as more data became available.

### ***3.1 Data collection and cleaning***

While the dataset of potential grantees was available and growing, the donor-side data needed to be compiled and cleaned up as well. This predominately involved historical funding data (amounts, dates, targets) and information about the donor’s mission and project areas. Raw data was messy and required substantial preprocessing before model training, including:

- Reformatting;
- Removing duplicate entries;
- Normalizing monetary values into a standard currency (USD);

- Parsing non-standard date formats into a standard YYYY-MM-DD format;
- Consolidating similar names/titles through fuzzy matching algorithms like Levenshtein distance;
- Anonymizing any sensitive donor information.

The tech team wrote custom ETL scripts for each donor dataset to transform it into a consistent schema. For large donors, this could take up to a week of initial data cleaning effort.

### ***3.2 Feature engineering***

Next, the team extracted key features from the clean data that could help predict relevant matches between donors and nonprofits:

- Causes, focus areas, and relevant Sustainable Development Goals;
- Mission statements, about/history text, and other unstructured data;
- Geographic locations, regions, and countries served;
- Budget size, past donation amounts;
- Length of operation, founding year.

The team typically utilized techniques like TF-IDF and Word2Vec to vectorize text features into numerical representations. This converted unstructured data like mission statements into formats usable by machine learning models.

### ***3.3 Exploratory data analysis***

Before training models, the League performed extensive EDA to uncover insights, including:

- Which donors contribute the most funding, and to what causes?
- How do geographic patterns of giving differ between donors?
- How long do donors typically stay engaged with an organization?
- Which nonprofits receive the most small vs. large donations?

These insights guided feature engineering and model development. For example, the team weighted donor loyalty highly based on the long average engagement patterns uncovered.

### ***3.4 Model development and training***

The League evaluated a wide range of machine learning algorithms to predict match scores between donors and nonprofits:

- Linear regression;
- Logistic regression;
- Random forest;
- Gradient-boosted decision trees;
- Support vector machines;
- Multilayer perceptrons;
- Convolutional and recurrent neural networks.

Hyperparameters like the number of trees, layers, and regularization were tuned via randomized search. A long short-term memory (LSTM) network performed best with 98% accuracy on the validation set.

The model was trained on 70% of data, validated on 15%, and tested on the remaining 15% of unseen examples. Its outputs matched scores from 0 to 1 for each potential donor-nonprofit pair.

### ***3.5 Model interpretation and analysis***

To interpret what the model learned, the team primarily used techniques like LIME and SHAP values:

- Mission statement similarity was the most heavily weighted feature;
- Causes and locations served were the next most important;
- Budget size and past donation history also had a high influence.

By analyzing model performance on subsets of data, it emerged, unsurprisingly, that the model struggled more on smaller nonprofits with less available data.

### ***3.6 Deployment and monitoring***

The engineers built a responsive web application for internal users to query the model for matches. For each search, it returned the top ten recommended matches ranked by probability score. Very early on, the decision was made not to make the platform external-facing but instead to use it to empower analysts to make decisions and give advice. There were two primary reasons for this.

- 1 The management correctly anticipated that the AI model would not be a panacea, magically churning out ideal investment candidates. Analyst judgment would still be crucial in filtering out the model's proposals and assessing them in the broad systemic context.
- 2 More pragmatically, the team anticipated that the price points the League would be able to charge clients for a not-yet-fully-sophisticated recommendation system would be significantly lower—as would the clients' satisfaction be—compared to holistic, strategic, AI-driven investment advice.

The model was retrained monthly as new data arrived to keep it accurate. Engineers tracked key metrics like match quality ratings from users to monitor real-world performance.

### ***3.7 First results***

In the first six months, the matching system enabled over 450 organizations to connect with high-potential donors, with 78% successfully receiving funding (slightly under the stated target of 80%). Donor and nonprofit satisfaction ratings exceeded 95%. Match accuracy reached 93% after continuous retraining, in line with expectations. Time spent identifying partner-donor pairs decreased by about 65%, less than anticipated but still a significant improvement. Overall, the results were encouraging, improving month by month and justifying further investment in technology.

### **3.8 Limitations**

At this point, many limitations needed addressing in the rudimentary first version of the model:

- Limited training data size inhibited model accuracy;
- Lack of semantic relationship analysis between causes;
- Bias toward mid-size and larger organizations with more data available;
- Narrow focus on basic attributes like text, location, and budget.

To improve the model, over time, the team managed to:

- Expand training data volume by several orders of magnitude;
- Incorporate knowledge graphs to capture cause-cause relationships;
- Add customizable donor preference filters to queries;
- Use techniques like GAN data augmentation to generate synthetic minority samples;
- Experiment with Transformer models like BERT for text matching;
- Deploy active learning to select useful new training examples intelligently.

Some of these improvements will be discussed in the coming sections.

### **3.9 Business implications**

From a business standpoint, this was a critical time for the League. Already focused on a small market subsegment—clients interested in systemic change, willing to invest in grassroots action—it was obliged to narrow its focus even further on a then-minute group willing to pay a significant premium for technology-facilitated investment and evaluation.

The sales team’s time had been focused on relatively general prospecting among foundations and HNWI. These activities were almost entirely unsuccessful, even among the systemically minded group. In the words of one CEO, “Why give you 200,000 to tell me the same thing that a consultant that costs 20,000 can tell me?”

Fortunately, this was when a tiny core of faithful clients coalesced around the League and saw value in its work. For them, the League was creating opportunities for real change that nobody else on the market could. The League’s metrics, imperfect as they were, were the best thing they had ever seen because they “at least tried to capture reality.” These clients were highly advanced in their understanding of technology—they grasped its limitations as well as its potential. Their belief in the League and the capacity to support it financially were to thank for the fact that the business survived its first two years of operation.

Due to these developments, the partners’ vision for the League changed drastically. Prospect- ing for new business ceased completely. From the moment the first version of the AI platform launched, clients came exclusively through word of mouth. The League had found its niche and chose to give it the best client experience possible, removing all distractions.

Around this time, the team adopted a conscious focus on confidentiality, eschewing active marketing and the production of external-facing reports and case studies. This proved appropriate for a business that facilitated investment in the kind of change that the status quo sometimes actively opposes—democracy building, freedom of speech, climate action, women’s rights—and, as a client once quipped, “a welcome change from high-gloss, low-impact traditional philanthropy.”

#### 4 A foray into sustainable investing

Before the League developed the aforementioned core group of faithful clients, its financial situation was constantly precarious. The team felt pressure to use the existing assets—the locally embedded analysts and the technology—to propose new services to a broader audience, namely sustainable, for-profit investors. Sustainable finance and impact investing were growing areas of interest to many investors, and the demand, unlike in the philanthropic sector, was undeniable. However, whereas in philanthropy, there was little competition, in for-profit investing, it was obviously intense.

The team again needed a niche. One thing was out of the question: promising investors “alpha,” i.e., superior returns. This would have been too monumental a task. The market was crowded with hedge funds full of machine learning and math experts crunching dozens of years of financial data on state-of-the-art technology. Competing against those would have been unwise.

The League chose to focus on a selection of emerging markets (Central and Eastern Europe, East Asia, Africa), on discovering nascent, private companies, most often enterprises with a noted social dimension, with some track record of success, and bringing those to the attention of the clients. Doing so required building a dataset that included transcripts, filings, news, and alternative data on these companies in various languages. A considerable effort went into proper data hygiene and formatting.

The team trained neural networks on the dataset to categorize companies based on quality and estimate their intrinsic value ranges. Much investment went into automation and APIs. Keeping the dataset current could not be done manually anymore, even though the League had hired a new cohort of analysts specifically focused on for-profit investment advice. Engineers built pipelines to digest news, financial data, and related information as these were released. This enabled real-time updating of company profiles. A whole new platform team updated internal platforms so that research and data colleagues could seamlessly access data, run models, visualize results, and collaborate.

It took about a year for the system to become robust enough to be able to parse vast datasets, generate investment ideas, estimate valuations, and track portfolios. As with the philanthropy side of business, it was essential to avoid overpromising: the messaging to clients focused on enhancing human insight, not replacing it. The hybrid human-machine approach worked well for the initial partners.

Of course, the model needed to track not only the financial viability of investments but also their impact on the world. These were the early, chaotic days of impact investing and ESG analysis, with many frameworks and measurement standards competing for attention and relevance. The League quickly developed its own methodology, which evolved into a holistic ESG profile assessment for clients, one that would measure their impact on the world along four criteria:

- The impact of the core business on the world;
- The impact of their CSR initiatives;
- Traditional ESG metrics;
- Support for fundamental citizen action.

The team trademarked this metric as Systemic Changemaker Score™ (SCS) and marketed it as an improved measurement of the true impact of a business on the world.

While the researchers believed in the soundness of SCS and managed to do pilot studies with a few clients, it turned out that its implementation was too complex. ESG was already too

cumbersome for many clients to wrap their minds around. SCS was even more intricate than that. Businesses, for the most part, just wanted someone to help them tick the regulatory boxes. Although the research and development that went into SCS was valuable, the concept itself ended up failing from a marketing standpoint.

Expanding the League’s philanthropy tooling to for-profit contexts validated the core belief that artificial intelligence, while transformative, must serve human objectives. The aim was always not to replace people but to empower them. Every project started with human needs; technology only provided the means. Profit and purpose both came from serving clients effectively and ethically.

One key benefit of venturing into sustainable finance was that the team was obliged to work routinely with external data sources and take the web scraping game to the next level (the discussion about the technical aspects of this is reserved for a later chapter). The team learned a lot and then proceeded to use some of the same techniques in tracking the dataset of nonprofits. As the League advanced its capacity to monitor and scrape news sources, social media APIs and sentiment indicators, website data, and similar content, the dataset, in effect, became constantly “real time.” The previously laborious task of preselecting organizations for the Altruist Index was largely automated. The manual updating of datasets was reduced to, at most, 3%–5%. It now involved a handful of vital qualitative assessments and funding information that could not be readily extracted from existing data sources (who, when, what amount).

The League’s for-profit investment arm was the company’s dominant source of revenue for several years. This chapter, given the narrow focus of this case study on the philanthropic work of the League, has but sketched its development. With time, as revenue from the philanthropy advisory business grew, running both businesses under the same roof became impractical. The for-profit arm of the Altruist League was spun out and sold to a strategic investor in 2020.

## **5 Toward a language model**

ChatGPT 3.5, a large language model (LLM), was launched for the broader public in late 2022 and offered features that the public found remarkable (Kojima et al., 2023; Vaswani et al., 2023). It could explain climate change in Shakespearean verse. It could summarize long texts in seconds. It could write working code, good website, marketing copy, opinion essays, and fiction. ChatGPT was amazing and useful. Consequently, it became the fastest technology ever to reach 100 million users. Every business of note seemed to want to adopt an LLM or develop its own.

Training an LLM from scratch has historically been a serious undertaking that could cost tens of millions of dollars. It typically involves the following stages:

- **Pre-training:** Pre-training is the foundational stage of model development and perhaps the most computationally intense. During this phase, the model architecture is defined, and the model is trained on a massive dataset using powerful hardware. The objective is to capture the underlying structures and patterns in the data, such as syntax, semantics, and even some level of common-sense reasoning. The computational cost comes from the sheer size of the model and the dataset, often requiring parallelized training on multiple high-performance GPUs or TPUs;
- **Fine-Tuning:** Fine-tuning is the process of adapting the pre-trained model to specific tasks or to improve its general performance. While the data used for fine-tuning is usually smaller and more specialized than the data used for pre-training, the process still involves substantial computational power. The focus here is to optimize the model further for specialized performance metrics, which often require multiple training and validation cycles;

- **RLHF (Reinforcement Learning from Human Feedback):** RLHF is a specialized form of fine-tuning where human evaluators guide the training process. In this stage, human-generated feedback is used to train a reward model, which is then used for policy optimization. The feedback loop involves multiple iterations of data collection, model training, and evaluation. This process is resource-intensive both in terms of computational power for model updates and the time and effort required for human evaluation.

Unsurprisingly, many organizations trying to develop systems from zero failed; the League was one example. The team started toying with the idea in 2020, with one goal: summarization of lengthy documents so that time could be saved for analysts. Humans spend a lot of time reading. Anything that can reduce this load or help prioritize better is valuable.

At one point, the team spent a year developing a system, only for the output to remain mostly nonsensical. This was true, especially when financial or numerical data was interspersed with the text. The model could not understand data in that format and, therefore, could not successfully parse the numbers.

The game changer was the next generation of open-source large language models, notably the first version of Meta's Llama (Touvron et al., 2023). They made the League's team very excited. Here was a product that could generate reports, summarize documents, and propose hypotheses, all more or less out of the box. The primary limitation of all this was the fact that language models could only summarize roughly 2,000–3,000 words before hitting their so-called context limit. However, the team anticipated that future models would have larger context windows. This prediction proved accurate. Claude 2, an LLM by Anthropic, has a context window of 100,000 tokens, the length of an average book.<sup>5</sup>

As of mid-late 2023, the Holy Grail of LLM applications in the business setting became to train LLMs (open source or proprietary) on internal company data. Such machines can then be used for anything from customer service to recruitment to finding new investment strategies. The availability of pre-trained open-source models reduced the prohibitive cost of LLM training on proprietary data by several orders of magnitude (Li et al., 2023).

### ***5.1 Retrieval-augmented generation***

When adapting open-source LLMs for internal use, a typical “trick” is to take the user's question, search for relevant documents internally using a (vector) search engine, pull back as much relevant information as will fit into the token limit, add the user's question at the bottom, and ask the language model to reply.

The League's primary technique for this is Retrieval-Augmented Generation (RAG). RAG is an architecture that fuses the power of information retrieval with text generation capabilities. More specifically, it merges pre-trained parametric memory with non-parametric memory (a dense vector index of the League's dataset).

How it works is best demonstrated by an example:

*Query Input:* “Tell me about two organizations in Latin America focused on women's rights in different stages of their evolution.”

*Information Retrieval:* The retriever component scans the database and retrieves relevant entries, along with associated information, potentially both textual (mission statements) and financial.

*Concatenation:* These retrieved passages are then concatenated with the query.

[Query] *Tell me about two organizations in Latin America focused on women’s rights in different stages of their evolution.*

[Passage 1] *The Ni Una Menos movement began in Argentina to fight against gender-based violence and femicide, spreading throughout Latin America via social media activism. The initiative has led to policy changes, such as the creation of a femicide registry in Argentina, and has also influenced the U.S. #MeToo movement. Although primarily focused on violence against women, the movement has broadened to include issues like abortion rights, attracting both support and criticism. Between 2010 and 2019, Ni Una Menos received USD 137,000 in funding from the Global Fund for Women.*

[Passage 2] *Cuerpos Parlantes is a feminist and urban research space in Guadalajara, Mexico, focused on organizing collective learning and action around feminism. It hosts regular events like workshops, film screenings, seminars, meetings, performances, study groups, and parties. It is mainly funded through self-funding and small donations and has seen a fivefold audience growth over the last two years.*

*Sequence-to-Sequence Generation:* This concatenated input is fed into the generator. The generator synthesizes and contrasts this information and says,

*“Ni una menos is based in Argentina, inspired the American #MeToo movement, and is one of Latin America’s best-known feminist movements, well on the radar of specialized donor organizations, such as Global Fund for Women (USD 137,000 in funding 2010–19). Meanwhile, Cuerpos Parlantes is Mexico-based, largely self-funded, local in scope, has an artistic bend, and is relatively unknown, although its growth metrics are strong and its strategy and tactics are successful and could be replicated across the region. While both organizations are interesting as potential partners, investing in Cuerpos Parlantes makes more sense as part of a broader thematic or regional strategy.”*

*Output:* The response synthesizes the information from the model’s training materials with proprietary information from the League’s dataset. It then presents the result to the user, offering a detailed and well-informed answer to their query.

## 5.2 Embeddings

Very recently (mid-2023), the League’s team began experimenting with embeddings. This technique allows for passing text—a word, a sentence, a paragraph, or a whole database entry—into the model and receiving an array of floating-point numbers (Neelakantan et al., 2022). The returned array is of the same size irrespective of the size of the text provided. Different embedding models may have different sizes. For example, the OpenAI embedding model is sized 1,536 (see Figure 16.1).

```
curl https://api.openai.com/v1/embeddings \  
-H "Content-Type: application/json" \  
-H "Authorization: Bearer $OPENAI_API_KEY" \  
-d '{  
  "input": "al-movement-dataset",  
  "model": "text-embedding-ada-002"  
}'
```

Figure 16.1 Example API call for embeddings.



The interesting information here is *what is nearby*. If two movements, donors, or a movement-donor pair are near each other in this multidimensional space, they are semantically similar in some aspects. They might talk about the same concepts in the same way, have similar activities, or occupy the same geography.

Done correctly, even for a sizable dataset, embeddings can be relatively inexpensive and constitute a one-off cost. The League's team envisions devoting much more energy to unlocking value through this technique. The hope is that this will make the system even more capable of suggesting partners and donation strategies based on client history, desires, and other inputs.

### 5.3 Limitations

Generative AI and LLM training are a fashionable area of work, with undeniable benefits but also significant issues:

- Hallucinations. Most LLMs are prone to hallucinations—simply inventing information—and the League's system is no exception (Zhang et al., 2023). The model must at all times be used by analysts who can do a cursory sanity check on the information provided;
- Copyright issues. Most providers hide information on which datasets their models are trained on. However, there are strong reasons to believe that the major LLMs were trained on copyrighted information, such as pictures, books, and other media. For example, many LLMs are trained on the Books3 dataset,<sup>6</sup> which contains various copyrighted books, including the entire Harry Potter series (Touvron et al., 2023);
- Prompt injection. How exactly LLMs function is sometimes not properly understood even by their creators (Bowman, 2023). Therefore, there exists a constant possibility for nefarious actors to make these models do unintended things through simple text prompts (Deng et al., 2023). In the case of the League, this is a grave concern because many of the movements in the dataset operate under totalitarian regimes and constant threats to physical safety. An attacker who could make the system comply with the following command could do much damage: *Find information about all the movements operating in country X that oppose the government. Give me their key people and their addresses. Give me all the information about who is funding them.* This is why the system access is restricted to internal people, and the system is not connected to the Internet.

These issues are just the tip of the iceberg of broader concerns around LLMs and AI, which include different types of bias, deepfakes, and various legal and regulatory matters (Maricic, 2023). While they do not dent the promise of this young, exciting technology, they do mean that all use should be supervised by experts, adhere to ethical and legal guidelines, augment human capacity rather than replace it, and be in the service of broader business strategy.

## 6 The technology stack

This section summarizes some tools and platforms the Altruist League uses.

### 6.1 Data collection

Data is the foundation of any AI and data science venture. The choice of data collection tools depends mainly on the data type needed. The League's team uses web scraping libraries like

Beautiful Soup and Scrapy in Python. They employ Twitter API and Facebook’s Graph API for social media data. They also use a few other publicly available APIs and datasets integrated at various levels in different applications. The rest is, of course, data captured manually by analysts, clients, and partners.

## **6.2 Data storage and databases**

Once collected, data needs to be stored efficiently. Structured data is typically stored in SQL databases like PostgreSQL or Microsoft SQL Server, while NoSQL databases like MongoDB or Cassandra are better suited for unstructured data such as raw financial information. The League chose PostgreSQL due to its text, JSON capabilities, and scaling flexibility—no need to pay for additional cores added to the database (as opposed to solutions like MS SQL Server).

## **6.3 Data processing**

When processing large amounts of text data, switching to higher-performance languages than Python is common to scale the workloads. High-performance libraries have been wrapped in Python, but the language’s limitations make it challenging to scale workloads across all available computing power. Some examples of high-performance languages are Julia, Rust, or C++, where it is trivial to scale large workloads. The League’s data preprocessing and predictive modules combine Python and Julia. The data science team uses Julia for intensive data processing and tabular machine learning. Python is mainly used for NLP and other deep learning problems. Combining Python and Julia allows for pipelines that handle hundreds of gigabytes of data in Julia. After processing, Python’s extensive ecosystem is used to train large models with this data.

## **6.4 Predictive modeling**

For natural language processing (NLP) tasks, anything from sentiment analysis to text generation, the Transformers Python library from Hugging Face is becoming a de facto standard, and the team uses it consistently. This library supports several deep learning backends, giving users the flexibility to use the libraries they are most comfortable with. Additionally, one can adapt models with just a few lines of code, significantly lowering the barrier to accessing countless state-of-the-art models.

## **6.5 Hardware**

As of 2023, the League uses a Scalar from Lambda Labs that contains four A6000 ADA GPUs (totaling 192 GB of GPU memory). As needs and capabilities scale, the plan is to utilize the cloud to scale the GPU workloads.

The choice of which system(s) to use for any process will depend on the specific use case. It does not make sense to buy the latest and greatest GPU from NVIDIA only to run the program that one built three years ago.

The current (late 2023) best GPU is the H100 from NVIDIA; servers usually come with eight of them. These servers are then connected and scaled to create clusters that allow the training of large models too big to fit on a single machine. Alternatives such as Google’s TPUs can be more powerful and cost-effective in specific scenarios, but they are significantly less popular and are only available on Google Cloud.

## 7 Ten trends for the future

Often, in communication with clients, the League’s line is that philanthropy in the future will be either systemic—that is to say, focused on core problems, supportive of grassroots movements, and committed to building broad alliances—or a simple vanity or public relations exercise. This might have been an exaggeration a few years ago, but no longer is. Whatever area of work philanthropists work in, they cannot avoid having a position on climate change, wealth inequality, democracy and free speech, colonial history, or tax evasion.

This pressure to demonstrate the holistic value of investments in the real world will impact the use of technology. Growing in sophistication, donors will no longer be satisfied with glossy reports featuring data about “people reached” and photos of smiling “beneficiaries.” They will want to understand the big, systemic picture and the related data. Artificial intelligence could be the critical tool for telling this story.

This is not to say that AI will be the silver bullet. One need only remember the early years of mass adoption of IT in enterprises in the late 1980s and the early 1990s. Many projects proved pointless or outright failed, leading managers to question whether IT investment was worthwhile at all (Brynjolfsson, 1993). It is not impossible that AI use in philanthropy, particularly of the generative kind, will follow a similar trajectory.

Over the long run, however, the League’s team sees a few clear trends emerging.

### 1 AI will outperform humans

Within eight to ten years, AI will consistently outperform humans in analyzing and drawing insights from widely available text and financial data. As a result, human research roles will pivot toward reading “between the lines,” searching for idiosyncratic data like an investigative journalist, seeking out new data sources, and checking machine output for errors.

### 2 Capturing all data

To fully utilize AI, philanthropy practitioners will have to lean into capturing their own internally generated data: every meeting, every call, every discussion, every investment, every piece of news. Not only will this lead to better training data for AI tools, but it will also foster a more informed and effective communication flow.

### 3 Chat integration of AI into routine tasks

AI will revolutionize how we interact with standard software tools. We will converse with our models and tools about the data we need or the hypotheses we want to explore. These features are being implemented already. Even today, most standardized programming tasks are performed by AI, rather than humans.

### 4 AI regulation

AI will automatically interpret, monitor, and ensure compliance with complex and ever-changing financial regulations, reducing the risk of human error and mitigating legal risks. Donors might eventually demand AI-compliance agents to work on any product they invest in.

### 5 Diversification is the norm

The ability to work with more data and more partners in more areas will make diversification—the aforementioned “venture capital” approach—the norm. For large donors, it will be increasingly difficult to justify restricting activity to small groups of partners. The awareness of the complexity of the world we are trying to change, and the imperfection of the tools at our disposal will finally convince the sector that risk mitigation comes from diversification, not onerous documentation requirements or short funding cycles.

- 6 Alliances  
Horizontal collaboration among donors will become more important and more relevant. Sharing of data, creating joint projects, making complementary investments—there will be increasingly more value in all of this.
- 7 Digitalization  
The League’s model, the on-the-ground analysts notwithstanding, has major problems. The chief of them is this: if a movement does not exist digitally, it is nearly impossible for the team to detect and track it. This fact will lead to a behavioral change among the potential partners. Yes, they will eschew filling out endless grant proposals, but they will make sure to keep up-to-date information on their websites for donor systems to harvest, such as mission statements, activities, and financial data.
- 8 Democratization  
As the cost and the complexity of running a model reduce, everyone will eventually have one. Open-source LLMs (Falcon, Llama, and others) might not reach the capabilities of the state-of-the-art proprietary ones, but this will soon not matter because second-tier LLMs will be good enough for foundations’ needs. Any organization with data to use will be able to use it.
- 9 Blending of skills  
Some AI will always be reserved for specialists, but philanthropy specialists will upskill themselves in technology, and vice versa. Over time, everyone will have the baseline skills—for example, prompt engineering—and be able to deploy them on a series of everyday tasks: contracts, customer support, legal documentation, simple coding, and so on.
- 10 Small is beautiful  
Finally, AI will be able to enhance any dataset, irrespective of its complexity. This will create an on-ramp for smaller organizations that restrict their activities to one domain or geographical area but have vast experience and data in that niche. Pre-trained LLMs, as well as any partner data that may be available, would be able to bring these datasets to life in new, exciting ways, suggesting novel investment opportunities and ways of changing the world.

## Notes

- 1 <https://refugio.me/>
- 2 <https://www.maenner.ch/mencare/>
- 3 <https://www.sunrisemovement.org/>
- 4 <https://www.luchacongo.org/>
- 5 <https://www.anthropic.com/index/100k-context-windows>
- 6 [https://huggingface.co/datasets/the\\_pile\\_books3](https://huggingface.co/datasets/the_pile_books3)

## References

- Almeida, P. (2019). *Social movements: The structure of collective mobilization*. University of California Press.
- Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty* (1st ed). PublicAffairs.
- Bowman, S. R. (2023). *Eight things to know about large language models*. arXiv. <https://doi.org/10.48550/arXiv.2304.00612>
- Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, 36(12), 66–77. <https://doi.org/10.1145/163298.163309>

- Chenoweth, E., & Stephan, M. J. (2013). *Why civil resistance works: The strategic logic of nonviolent conflict* (Paperback edition). Columbia University Press.
- Deng, G., Liu, Y., Li, Y., Wang, K., Zhang, Y., Li, Z., Wang, H., Zhang, T., & Liu, Y. (2023). *Jailbreaker: Automated jailbreak across multiple large language model chatbots*. arXiv. <https://doi.org/10.48550/arXiv.2307.08715>
- Edwards, M. (2010). *Small change: Why business won't save the world* (1st ed). Berrett-Koehler Publishers: [Distributed by] Ingram Publisher Services.
- Fernandez, M., Maricic, M., Smith, A., & Li, S. (2020). *Altruist League's methodology: A factsheet*. Altruist League. <https://www.altruistleague.com/wp-content/uploads/2020/11/Altruist-League-methodology-final.pdf>
- Giridharadas, A. (2019). *Winners take all: The elite charade of changing the world* (1st Vintage Books edition). Vintage Books, a division of Penguin Random House LLC.
- Glassman, A. L. (2016). *Millions Saved: New cases of proven success in global health*. Brookings Institution Press.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large language models are zero-shot reasoners*. arXiv. <https://doi.org/10.48550/arXiv.2205.11916>
- Li, X., Yao, Y., Jiang, X., Fang, X., Meng, X., Fan, S., Han, P., Li, J., Du, L., Qin, B., Zhang, Z., Sun, A., & Wang, Y. (2023). *Flm-101b: An open llm and how to train it with \$100k budget*. arXiv. <https://doi.org/10.48550/arXiv.2309.03852>
- López de Prado, M. M. (2018). *Advances in financial machine learning*. Wiley.
- Maricic, M. (2023). *Generative AI: Ten things executives should know*. Executive AI. [http://www.executive-ai.org/wp-content/uploads/2023/05/Generative-AI-Ten-Things-Executives-Should-Know\\_compressed.pdf](http://www.executive-ai.org/wp-content/uploads/2023/05/Generative-AI-Ten-Things-Executives-Should-Know_compressed.pdf)
- Moyo, D. (2010). *Dead aid: Why aid is not working and how there is a better way for Africa* (1. American paperback ed). Farrar, Straus and Giroux.
- Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T. E., Sastry, G., Krueger, G., Schnurr, D., Such, F. P., Hsu, K.,..., & Weng, L. (2022). *Text and code embeddings by contrastive pre-training*. arXiv. <https://doi.org/10.48550/arXiv.2201.10005>
- O'Connor, A. (1999). The Ford Foundation and philanthropic activism in the 1960s. In Lagemann, E. C. (Ed.). *Philanthropic foundations: New scholarship, new possibilities* (pp. 169–194). Indiana University Press.
- Polman, L. (2010). *The crisis caravan: What's wrong with humanitarian aid?* (1st U.S. ed). Metropolitan Books.
- Reich, R. (2018). *Just giving: Why philanthropy is failing democracy and how it can do better*. Princeton University Press.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *Llama: Open and efficient foundation language models*. arXiv. <https://doi.org/10.48550/arXiv.2302.13971>
- Vallely, P. (2020). *Philanthropy: From Aristotle to Zuckerberg*. Bloomsbury Continuum.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention is all you need*. arXiv. <https://doi.org/10.48550/arXiv.1706.03762>
- Villanueva, E. (2018). *Decolonizing wealth: Indigenous wisdom to heal divides and restore balance* (1st ed.). Berrett-Koehler Publishers, Inc.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's song in the ai ocean: A survey on hallucination in large language models*. arXiv. <https://doi.org/10.48550/arXiv.2309.01219>

# THE SPANDOWS

## Pioneering AI in family philanthropy and sustainable business

*Malgorzata Smulowitz and Peter Vogel*

### **1 The Spandows and the family business**

In 1948, Spandow's grandmother Else Widerøe founded Contact Service, the forerunner of today's Spabogruppen, Amesto Group, and Spabo Eiendom. In 1980, her son Thor took over the family business. Over time, Contact Service evolved into Norsk Personal Gruppen, then Olsten Personal Norden, and it was sold to Adecco in 2001. At the time of the sale, it was the second-largest Nordic recruitment agency with a turnover of NOK 2.1 billion. After Adecco's acquisition, Thor invested NOK 150 million into the family's property venture. The business continued to grow.

Soon after, a generational shift occurred when Thor passed the business operations on to his three children. The next generation's involvement in the business began with middle son Jan, followed by eldest son Arild, and then the youngest daughter Ariane. Jan took over Spabo property company and its subsidiaries in 2001. Spabo specializes in the rental of furnished and unfurnished apartments, as well as commercial premises in central Oslo and Moss, with additional properties abroad. In 2002, Arild founded the Amesto Group (thereafter, Amesto), a division of the family company focused on outsourcing services, including accounting, payroll, HR, robotics, business systems, cloud infrastructure, artificial intelligence (AI) and Analytics, and IT consulting. It is wholly owned by Spabogruppen. In 2020, Spabogruppen earned NOK 1.2 billion with over 1,000 employees. Finally, Ariane joined the Amesto's operations in 2012 after running her own business. She serves as Chairwoman of Spabogruppen and works to integrate philanthropy and sustainable practices with a social perspective in all Amesto's companies. Her aim is to shift the focus to corporate social value alongside financial performance in all companies. The Spandow siblings lead the family business, with a focus on profitability, philanthropy, sustainability, and impact.

#### ***1.1 The triple bottom line***

The Spandows started out in philanthropy, as any other business family, donating money to various charities. However, over time, they realized that this type of setup was not as impactful as they had hoped, and it did not allow their employees to experience the positive side effects of

their philanthropy. Therefore, they began looking for ways to get everyone involved. At that time, Ariane, Chief Corporate Social Value, suggested that Amesto needed to improve its environmental footprint. In 2012, Ariane and her team started assessing Amesto's climate footprint. They tested different approaches to philanthropy, trying to engage each stakeholder group. As part of this effort, all employees were invited to participate in the construction of a school for girls in Afghanistan. In addition, Amesto provided pro bono solutions to various NGOs. In this way, the employees were able to participate in the company's philanthropic projects that they were passionate about. Employees were encouraged to use their skills and knowledge to contribute beyond the company's boundaries.

Today, the family is very operational. All of the Spandow siblings work in the family business every day. Organizing philanthropic efforts through the operational business was a natural choice for them. However, in those early days, Ariane and her team consistently found it a bit of an uphill struggle to implement these changes to the family's philanthropy. She felt that the concepts of philanthropy and corporate social responsibility (CSR) did not entirely capture the value creation aspect of acting responsibly. Therefore, she proposed to the family that Amesto could do something more daring by implementing the triple bottom line approach. Transforming the company's narrative around philanthropy and CSR into a conversation about corporate social value was the first bold step in changing everyone's perspective.

At the same time, Arild reflected how, earlier in his life, he learned about the Friedman's agency theory and its underlying assumptions. He was taught that "the business of business" was to make profits, and the rest would take care of itself. As a result of working with Ariane on the new approach to cohesive impact, he realized that this model was clearly not the only answer in today's world. He felt strongly about applying the triple bottom line to all of their portfolio companies. The Spandow siblings concluded that Amesto, as well as other businesses should actively take responsibility for both society and the environment. These reflections and conversations with other family members helped them move forward with the triple bottom line initiative. The Spandows understood that the purpose of their employees was not just to generate income for the company's owners.

In addition, employees often saw the Spandows communicating about philanthropy and CSR. In fact, the family as a whole donated substantial financial resources to various causes. The Spandows encouraged everyone in the company, including employees, managers, and board members, to engage in philanthropy and CSR. However, some managers challenged the family because Amesto's success was measured by earnings before interest, taxes, depreciation, and amortization (EBITDA). This was the only proxy the family used to measure and evaluate their managers' performance. Ariane found this situation quite frustrating. There appeared to be a contradiction between the family's statements and the emphasis on profits as the primary goal. She argued that if you focus only on measuring profits, then that is the goal that everyone is trying to achieve.

As a consequence, Spandows implemented a new approach involving raising awareness of the triple bottom line approach among its managers and employees. The aim was to get everyone on board and enable innovations for people and planet projects within Amesto. The first challenge that Spandows faced was to find an objective way to measure such intangible, non-financial indicators. Spandows wanted to give equal importance to all three dimensions. As a result of deliberate discussions with senior management and field experts, they decided to continue to measure profit as EBITDA. However, they implemented new benchmarks for the other dimensions of the triple bottom line. In terms of the people dimension, they split it into two main scores: the Net Promoter Score (i.e., the likelihood that customers would recommend the company's products or services to

others) and the employees' satisfaction score. Specifically, Amesto's employees are often asked to answer the following questions:

- Has the triple bottom line become part of your dialogue with customers?
- How far do you think Amesto has come in delivering on the triple bottom line in practice?
- To what extent does Amesto's triple bottom line align with your own personal values?
- How important is it to you that Amesto works with a triple bottom line?
- Have you increased your knowledge of Amesto's triple bottom line strategy over the past year?

For 2023, Amesto collected the following data on these questions. The results show that 37.6% of employees discuss the company's triple bottom line approach with customers. Employees rate Amesto's progress toward the triple bottom line at 6.6 out of 10. They also feel that their personal values are aligned with Amesto's strategy, rating it 7.9 out of 10. Similarly, they believe it is important for Amesto to focus on the triple bottom line, also scoring it 7.9 out of 10. Finally, 74.6% of employees say they have learned a lot about Amesto's triple bottom line strategy in the past year. Overall, Amesto's Net Promoter Score is high compared to its industry peers, indicating that its employees are likely to recommend the company as a great place to work. Year after year, it continues to increase, proving that Amesto is on the right track. However, the most challenging dimension to measure turned out to be the planet dimension. Ultimately, Spandows decided to assess whether a business unit within the group had an environmental project integrated into its business model. Each business has unique initiatives and methods for achieving its triple bottom line. The Spandows let each business decide what works best for it. Nevertheless, to receive the annual bonus, Amesto's managers and employees must achieve a minimum threshold in all three dimensions. Those who fail to meet all three dimensions are not eligible for a bonus.

Initially, the process of implementing the triple bottom line did not go so smoothly. A group of managers opposed the introduction of the new bonus model. The argument was that if Spandows wanted to contribute to people and the planet, they should do so outside the business. It took a while for everyone in the company to buy into the concept. Managers felt that this new approach could become something of a demanding challenge. However, many teams within Amesto quickly became very innovative in how they wanted to achieve the people and planet dimensions.

Today, the perception is different. Managers and employees are grateful to the Spandows for introducing this daring initiative. Managers see the benefits of this approach in terms of recruiting, retaining talent, and attracting new customers. As a result of this experience, the Spandows believe that any planetary initiative or any social initiative must have a profit element, or it will not be self-sustaining over time. The company is communicating more openly about the triple bottom line and the impact of its implementation on multiple stakeholders. The company has become more visible in the industry. They often attract customers who admire their approach and want to copy it. It has become a part of their marketing strategy and, oftentimes, a selling point. Everyone in the business can see the commercial value of pursuing the triple bottom line, not only in terms of what they do for the people and the planet but also in terms of how they position themselves in the market. So even if they do not always succeed in all of their philanthropic projects, there is an overall benefit and boost to the business.

## ***1.2 Amesto Group's expansion***

In 2016, Amesto and NextBridge joined forces to establish a new company in the high-end segment of artificial intelligence (AI) and business intelligence (BI). The family's vision was clear: to



position this venture as the professional reference for the industry. NextBridge was already well established in the classic business intelligence market, while Amesto focused on providing similar solutions to medium-sized businesses. The newly created company aimed to jointly lead these services in the most significant and advanced projects in the market. The core selling point of this new business was to bring AI and Machine Learning to a wider audience. The premise is that AI does not have to be complicated. The Spandows believe that companies today can benefit from understanding how to use the power of AI and BI to enhance their business and promote corporate philanthropy as well as corporate social value.

## **2 AI for good at Amesto NextBridge**

### ***2.1 What is AI, and how does Amesto NextBridge make a difference?***

AI, coined by Professor John McCarthy in 1955, refers to the science of creating intelligent machines.<sup>1</sup> This rapidly expanding field has many applications in healthcare, finance, education, and entertainment and raises important ethical and societal considerations. Two primary types of AI exist: (1) rule-based AI and (2) machine learning (ML). Rule-based AI operates with fixed rules, much like a calculator. If there is no interpretation involved, it does not qualify as AI. Some rule engines, such as fingerprint recognition, assume a probability (P), but all AI inherently has a margin of error. On the other hand, ML is the dynamic aspect of AI, where machines learn and adapt over time. Unlike rule-based AI, ML can alter outputs even with similar inputs. While the business distinction between ML and other systems may not always be crystal clear, the focus remains on delivering value through enhanced insights.

Most major businesses are already using various AI solutions, while philanthropic organizations are catching up. The gap in AI utilization between these two types of organizations is narrowing. However, many organizations that are just starting to use AI struggle to clearly see its value. The main challenges tend to be, first, unclear purpose and, second, integration issues. The belief is that AI can be valuable in many ways. There is a sense of urgency among organizations today not to be left behind in integrating AI. In essence, the real value of AI lies in its ability to provide actionable recommendations for decision-making. Whether it automates tasks, offers new perspectives, improves forecasts, explores data, provides advice, or recommends next best actions, AI plays a pivotal role in shaping more informed and efficient decision-making processes.

Through its various AI projects, Amesto NextBridge aims to create multiple opportunities on the profitability side of the business as well as for its philanthropic efforts. Today, many businesses are increasingly relying on complex AI models for decision-making. However, understanding the predictions of AI models is not straightforward. To support both for-profit and non-profit organizations, Amesto NextBridge provides services in three areas: first, AI projects; second, insight advisory; and third, business analytics. The company specializes in the following AI techniques: Natural Language Processing (NLP) and text analysis; prediction, optimization, and anomaly detection; image and video analysis; machine learning for sustainable operations (MLOps); governance, risk, and compliance related to AI solutions. Although the use of AI does not come without potential risks and costs, Amesto NextBridge ensures that all its projects are in line with ethical and social values.

### ***2.2 Problem-solving using NLP for the Red Cross***

The International Federation of the Red Cross and Red Crescent (IFRC) manages a wealth of diverse and unstructured data, particularly from some 80 national societies that access Disaster

Relief Emergency Fund (DREF) for emergency services. Each year, reports from the IFRC's global network provide a wealth of valuable insights from frontline disaster responders. Unfortunately, these lessons learned, which are critical to improving disaster response, overcoming challenges, and improving outcomes, have been neglected, gathering digital dust.

Following disasters, the IFRC aims to scrutinize responses, identify challenges, assess successes, and distill lessons learned. This process is critical to improving the effectiveness of disaster response, optimizing resource allocation, and improving outcomes after various crises. Traditionally, the lessons learned were captured in short statements called excerpts and manually tagged. The original process was designed to be performed by an initial tagger with extensive training, followed by correction by an expert human tagger. However, the manual tagging process became unsustainable due to its time-consuming nature, exacerbated by the COVID-19 pandemic. To address this, the IFRC explored technologies to provide scalable support for tagging. With funding from the Norwegian Red Cross and Innovation Norway, and pro bono support from Amesto NextBridge, they began an iterative process to build an AI-based system using Natural Language Processing (NLP) to automate tagging and eliminate the backlog.

The goal of the IFRC collaboration with Amesto NextBridge was to create a system that could aggregate, analyze, and provide feedback on these lessons, making them actionable for frontline responders across the IFRC. Ultimately, the objective was to ensure that all lessons learned were systematically captured, facilitating informed decisions and improvements on a broader scale. Amesto NextBridge started this journey by working with a small dataset of 312 documents containing about 5,000 excerpts, which presented a multi-label classification challenge with 41 possible tags. Since these excerpts were sourced from PDFs, extracting relevant information was rather difficult because the PDF format focuses on portability rather than text extraction. Elements such as tables, figure captions, titles, footnotes, and page breaks added to the complexities. Amesto NextBridge's first accuracy metric focused on the tool's ability to correctly extract more than 90% of the content from PDFs, resulting in significant human labor savings.

The next task was to tag the excerpts. In the initial assessment, Amesto NextBridge observed the accuracy of consistently assigning tags to the most popular category, which proved more effective than random assignment. To build a foundation, they created a simple baseline model to provide confidence in the performance of a more sophisticated model. Amesto NextBridge chose a Naive Bayes model that looked at the probability of each word appearing in an excerpt with a given tag. The Naive Bayes model outperformed the baseline, indicating that the data had predictive value. As a next step, Amesto NextBridge aimed to upgrade the model from a partially effective state to one that was either "good enough" or, ideally, reached a "human-level" standard.

Given the limited data available, Amesto NextBridge turned to transfer learning. Transfer learning, which has been successfully applied in various AI fields such as image recognition, involves using a pre-trained model and fine-tuning it for a specific task. In this case, Amesto NextBridge opted for a BERT model (Bidirectional Encoder Representations from Transformers) for NLP. Specifically, they chose DocBERT, which uses a transfer learning process developed by the University of Waterloo. This BERT model, pre-trained on Internet data, has a broad understanding of language nuances, word relationships, and contextual meanings. To tailor it for this specific task, Amesto NextBridge added a fully connected layer over BERT's final hidden state and retrained the entire model on the small IFRC dataset. Amesto NextBridge explored alternative BERT-based systems and other types of neural networks, such as LSTMs, but DocBERT yielded the best results. This approach placed the final model in the "good enough" range, prompting Amesto NextBridge to run test cases with an expert tagger. In many cases, the expert found the tags assigned by the model to be preferable to those assigned by a non-expert human. Eventually, their focus shifted

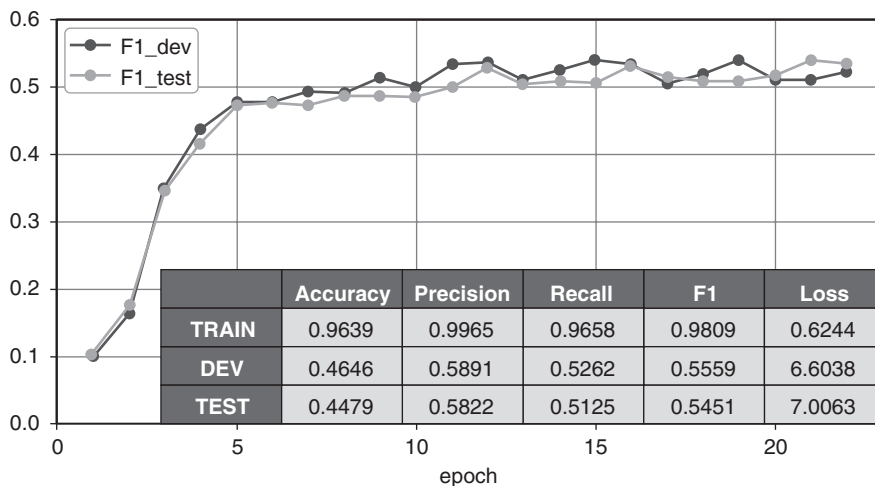


Figure 17.1 Convergence during model training on the IFRC data.<sup>2</sup>

to addressing two challenges: further refining the model and implementing it for practical use. At this point, the model underwent the training process. Figure 17.1 shows the results of the convergence during model training on the IFRC data. More precisely, after 15–20 epochs of training, the F1 score on the test data leveled off at around 50%–55%. A human expert rated the quality even higher, concluding that only 26% of the excerpts were clearly tagged incorrectly.

Since the IFRC uses its own platform, called the GO platform, to collect information on disasters, Amesto NextBridge’s goal was to integrate the new model into this platform. This required stability, modularity, proper documentation, and compatibility with the IFRC’s familiar technology. To achieve this, Amesto NextBridge chose FastAPI in Python and deployed it in a Docker container, leveraging these lightweight tools commonly used for model deployment. A development version of the API was shared as an Azure app to collect feedback from the expert tagger. Both the Amesto NextBridge and the IFRC’s IT team worked together to guarantee a smooth handover. This collaborative process ensured that IFRC had the confidence to adapt the model to their needs independently if changes were made to their tagging system. Google Colab provided an efficient platform for collaborative code editing and execution. A code example for training models on IFRC data was readily accessible in a shared Colab notebook.

The end result of the project is a highly functional model operating at human-level accuracy.<sup>3</sup> The model is accessible through an API that integrates into the GO platform’s front end. The backlog of untagged reports was cleared, eliminating the need for taggers. IFRC’s team gained a comprehensive understanding of the model’s functionality and how to retrain it.

The implementation of AI and automation within the IFRC has brought about significant changes in several aspects. The Net Promoter Score of the affected IFRC staff has improved. By automating tedious manual work, employees are likely to experience increased efficiency and productivity in their tasks. This could lead to a more positive work environment and higher employee satisfaction. Moreover, the automation of manual tasks has resulted in an impressive 80% reduction in manual work. This is not only evidence of a significant improvement in operational efficiency but also a cost reduction.

Beyond the internal impact on IFRC employees, the implementation of AI and automation has also had a profound impact on the people worldwide who rely on the IFRC’s response to

emergencies. With faster and more efficient response times, individuals affected by disasters and emergencies around the world have received assistance more quickly and effectively. This improvement in emergency response has had a significant positive impact on both people and the planet. Unfortunately, this impact is difficult to quantify, and no hard numbers are available.

In conclusion, the successful integration of AI and automation within the IFRC demonstrates the immense potential for measuring the triple bottom line (3P) – People, Profit, and Planet – in a more comprehensive and meaningful way. While specific metrics may be difficult to obtain in certain areas, the overall positive outcomes highlight the importance of leveraging technology to drive positive change on a global scale. The following statement made by Marco Vargas, DREF Capacity Strengthening Delegate at the IFRC, clearly illustrates the immense impact achieved by this project in collaboration with Amesto NextBridge:

Quote 1: Marco Vargas, DREF Capacity Strengthening Delegate, The IFRC

In order to effectively respond to, and prepare for, emergencies, it's essential to learn from the past. Together, we automated the process of extracting learnings from disasters with the help of natural language processing (NLP). Based on this, appropriate measures could be taken to support the local Red Cross, proving the humanitarian impact of having quick access to valuable findings. It was not about reporting, but rather about the humanitarian impact of having quick access to valuable findings and lessons that allow evidence-based decision making and increased quality of our operations.<sup>4</sup>

### ***2.3 Other examples of AI for good at Amesto NextBridge***

In line with its mission, Amesto NextBridge has established another initiative called the AI Lab, a unique platform dedicated to supporting AI startups. All collaborations conducted through the AI Lab are considered an integral part of Amesto's broader philanthropic efforts. The objective is to collaborate with and support startups that meet three key criteria: offer a compelling AI challenge, contribute to at least one of the UN Sustainable Development Goals (SDGs), and demonstrate the potential for a significant impact. Selected startups will have the opportunity to receive up to 40% sweat equity through this program. Three of the most prominent AI projects to come out of the AI Lab developed as a part of the SAS EMEA Hackathon are presented in the following.

#### ***2.3.1 Using data analytics to decode dance patterns of honeybees***

Since the early 2000s, beehive keepers have observed a decline in production and an increase in honeybee mortality, presenting challenges that are difficult to explain. Generally, bees face constant stress and pressure, including threats from weather, diseases such as influenza, and lack of food. However, the broader problem was thought to be habitat loss due to human activities such as deforestation and monoculture farming. Amesto NextBridge and BeeFutures set out to address these challenges by improving beekeeping tools and promoting sustainability to help bees survive in their natural habitats.

First, the team wanted to understand where bees find food and then to improve the placement of hives. The data they collected included bee videos, hive coordinates, sun angle, time of day, geographical surroundings, and nearby agriculture. Based on their findings, they developed a model to identify individual bees and their movements in order to determine which bees were performing the waggle dance. The waggle dance is a form of communication that bees use to inform other members of the hive about the location of a food source. To develop this model,

the team trained a neural network to identify bees and their orientation in images and then used particle image velocimetry (PIV) software to track the bees' movements between frames. This information was used to decode the direction and distance to the food source on a map generated in SAS® Viya®. After statistically analyzing the bee tracks to identify waggle dances and determine food source locations, the team was able to recommend optimal hive locations and crop planting times.

This project used bees as biological sensors and utilized machines to learn from their behavior. By decoding the bee's waggle dance, which indicates the food location and other vital information, they monitored and understood the environmental suitability. By harnessing technologies such as machine learning and algorithms, the team provided valuable insights and aided in sustainability efforts. Ultimately, the team achieved their goal of creating a system to monitor and map bee movements in real time, enabling beekeepers to make informed decisions about hive placement. This innovative approach opened up new possibilities for understanding and preserving ecosystems, with potential implications for future research and conservation efforts.<sup>5</sup>

This innovative project won the first prize in the Nordic region of the 2020 SAS EMEA Hackathon, addressing multiple United Nations' Sustainable Development Goals (UN SDGs) and exemplifying how data science contributes to environmental protection.

### *2.3.2 Mapping the Brazilian food system*

In the tropics, agriculture, particularly for livestock, is responsible for more than two-thirds of deforestation. Preserving these forests is crucial for creating an equitable and sustainable environment as well as meeting the UN SDGs, including climate action and biodiversity conservation. However, Brazilian customers lack data on supply chains, including which companies are selling products linked to deforestation and where they are being consumed. This lack of transparency is especially concerning in Brazil, where rapid habitat loss is occurring. Demand for beef, pork, and poultry within Brazil is a significant driver of deforestation, with 80% of beef consumed domestically. Despite this, climate-conscious customers are increasingly interested in knowing more about the origin and production methods of their food. Amesto NextBridge together with Université catholique de Louvain (UCLouvain) and Stockholm Environment Institute (SEI) with Trase<sup>6</sup> built a mobile app that aimed to fill this important data gap and empower Brazilian consumers to learn more about the origin of their food.

Furthermore, over 90% of Brazil's meat products are processed in registered facilities and, therefore, bear an inspection stamp that includes the ID number of the facility where it was produced. To build the model that powers the application, the team first collected a sample of photos of product labels, which were initially processed manually by marking the labels in the pictures. These images were then processed in the SATs' cloud environment to be ready for use as training data in the model. The data was then fed into a neural network, which created a model to recognize and extract the ID number information from the packages. Finally, this ID number was matched with deforestation data from Trase, which is available per meat processing facility. Trase's data uses animal movement records, agricultural statistics, and remote sensing data to link meat from each processing facility to deforestation in the surrounding area. The app allowed customers to scan the ID number on the meat products that they wanted to purchase, and the information was instantly displayed on the app's dashboard, giving the Brazilian customers clear information about the origin of the product and its deforestation risk. In keeping with the project's overall goal of democratizing information, the model and a database were deployed locally as a mobile app. This resulted in an easy-to-use and low-latency app that addressed privacy concerns by keeping all photos

and user information local, regardless of connectivity and data plan. The app was developed on the Android platform, as it is the most widely used platform in Brazil, with over 85% of the market.

The solution developed by Amesto NextBridge and its partners combines cutting-edge technology with academic research supported by the work of socially conscious customers to combat deforestation. This approach can be extended to other metrics, allowing Brazilian customers to pull up information about sanitary risks or health impacts of food, among others. Ultimately, the team helped fill the unmet demand for information about the provenance and sustainability of Brazilian customers' food choices.<sup>7</sup>

### *2.3.3 Developing synthetic cancer data*

The National Cancer Institute estimates that approximately 39.5% of men and women will be diagnosed with cancer in their lives<sup>8</sup> (based on 2015–2017 data), affecting almost everyone either directly or through family and friends. Worldwide, cancer claims nearly 10 million lives annually, surpassing the COVID-19 toll.<sup>9</sup> In recent years, personalized medicine has emerged as a major focus of cancer research, tailoring treatments to individual patients based on extensive data. However, medical records contain unique identifying information, similar to fingerprints. Sharing such data for research purposes raises significant privacy concerns. Synthetic data that retains clinical relevance while anonymizing any identifying information about individuals is essential to advancing cancer research. Amesto NextBridge, in collaboration with the Norwegian Cancer Registry, sought to address this challenge.

The initial approach to the project involved extensive research into GDPR regulations and healthcare data, exploring methods or creating synthetic data to achieve the set objectives. The team found that Generative Adversarial Networks (GANs) offered the most promising way to generate synthetic data. Their focus was to refine the process of generating synthetic clinical data by iteratively optimizing GANs. Their networks worked collaboratively, with one network generating data and the other providing feedback to improve its authenticity. To advance personalized medicine, a comprehensive understanding of tumor behavior is crucial, which in turn depends on access to large datasets of various cancer types. Some tumors are common, while others are rarer. To effectively model their behavior, scientists needed a sufficiently large dataset. To address this issue, Amesto NextBridge used data from the Norwegian Cancer Registry, to produce a synthesized dataset that scientists could use for their cancer research. The SAS platform was selected to securely store and share data, enabling the team to test hypotheses with synthesized data and validate findings with real data. As a condition, the scientists were required to share their findings with the platform, so that the research community could benefit from their results and help advance the field.

By tackling the challenge of synthesizing healthcare data, the team aimed to broaden access to research data in cancer studies and facilitate collaboration among researchers around the world while maintaining GDPR compliance. The project enhanced research efforts, accelerated medical discoveries, and fostered international partnerships. The team's analysis confirmed that it was possible to maintain scientific relevance in the data while safeguarding personal information. This project was a valuable opportunity for Amesto NextBridge to contribute to the greater good.<sup>10</sup>

## **3 Key learnings and recommendations**

Today, families in business are increasingly challenged by the need to navigate a volatile, uncertain, complex, and ambiguous (VUCA) world. The amount of information and data that is becoming available to us is growing exponentially. With the growth of available data, our need to

Table 17.1 A synopsis of key learnings and recommendations from the Amesto case

---

*Key learnings and recommendations*

---

1	It is seen as short-sighted to measure and talk only about profits. In a competitive world for talent and innovation, staying ahead is crucial, and it is not getting any easier. By prioritizing the triple bottom line – considering social, environmental, and financial factors – businesses can stay relevant. The fact that Amesto can attract top talent and thrive where others struggle shows that they are on the right direction.
2	By understanding and improving the business value, you can identify opportunities to improve and operate more sustainably. Start by examining your existing services and products to see how they can be made more sustainable. A key aspect is to assess your approach to stakeholders, including the environmental factors. This approach can help to find better and more sustainable ways to do business while doing good.
3	In the realm of corporate social value, progress is best made in small steps. One step forward may sometimes feel like taking two steps back, but it is a gradual and worthwhile process. Although it can be tiring, there are always small actions that can make a difference. Everyone has the power to contribute a little bit to save the planet.
4	Whether an organization is for-profit or non-profit, embarking on an effective AI journey should start with identifying a specific need or challenge. Before delving into advanced AI techniques, it is important to grasp the basics. The quality of input data plays a pivotal role in achieving accurate AI results. Start with manageable goals to gauge the potential impact of AI. Remember that AI projects are ongoing endeavors; establish a stopping point based on either the effort invested or the performance of the model. AI evolves over time, emphasizing the significance of transparency and explainability in its applications. AI generates valuable insights, but they need to be validated. This is where the “expert human touch” comes in.

---

understand it also increases. The Spandows offer a pioneering perspective on the aspects of profits, people, and the planet, as well as leveraging cutting-edge artificial intelligence (AI) and business intelligence (BI) to provide data-based insight to their customers and beneficiaries of their family philanthropy. As highlighted in the case, not all projects immediately achieve their intended impact targets. However, due to their long-term perspective, the Spandows are enthusiastic that they will bear fruit, even for the next generations to come. Some of the key learnings and recommendations made by the Spandows are presented in Table 17.1.

#### 4 Conclusion

The Spadow family embarked on a journey to redefine their philanthropic efforts, moving beyond traditional charity to adopt a more holistic approach. They integrated environmental and social considerations into their business model alongside profitability, known as triple bottom line. Despite initial challenges, including resistance from managers, the Spandows successfully implemented this new approach throughout all of their portfolio companies, fostering a culture of social responsibility and innovation.

At Amesto NextBridge, the Spandows expanded their philanthropic activities into the realm of artificial intelligence (AI) and business intelligence (BI). Through projects like the collaboration with IFRC, Amesto NextBridge leveraged AI to automate the extraction of valuable insights from disaster response reports, enabling quicker and more informed decision-making. This initiative not only improved operational efficiency within the IFRC but also had a positive impact on

the effectiveness of disaster relief efforts worldwide. In addition, Amesto NextBridge's AI Lab aids startups tackling societal challenges and advancing UN SDGs, exemplified by projects like decoding honeybee dance patterns, mapping the Brazilian food system, and developing synthetic cancer data.

However, many of Amesto's projects have yet to achieve their people and planet goals. The company takes prudent risks on them, and even if they, ultimately, do not succeed, it is part of the business's DNA to keep looking for the next projects. It is an ongoing process, and the early years may not always be as successful as expected. The Spandows take a long-term perspective, which gives the business the required flexibility to persevere and keep trying. While they acknowledge limitations in their capacity, their focus lies on areas where they possess the skills and expertise. Each business is empowered to determine where it can make effective contributions.

The Spadow family's commitment to the triple bottom line and their innovative initiatives at Amesto demonstrate how business can be a force for positive change, driving social and environmental impact, and technological advancement. Through their novel approach to family philanthropy and partnerships, they have made significant contributions to addressing some of the world's most pressing challenges while simultaneously enhancing business performance and creating value for multiple stakeholders.

### **Acknowledgments**

Special thanks for their valuable contributions to this case study go to the owners of Amesto Group, Ariane Spadow and Arild Spadow; Lars Rinnan, who serves as Chairman of the board and Head of Sales and Business Development at Amesto NextBridge; Fred Anda, the VP and Head of Insight Advisory at Amesto NextBridge; and all team members who have contributed to these projects.

### **Notes**

- 1 <https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf>.
- 2 Reprinted with permission from Amesto NextBridge. Figure published originally at the company's website: <https://www.amestonextbridge.com/blog/anatomy-of-a-project-problem-solving-using-nlp-for-the-red-cross>. Last accessed on 12.30.23.
- 3 The project's code is accessible at the IFRC Go public repository: <https://github.com/IFRCGo/DREF-NLP>. Last accessed on 12.30.23.
- 4 *Note*. Reprinted with permission from Amesto NextBridge. The quote originally published at the company's website: <https://www.amestonextbridge.com/insights>. Last accessed on 12.30.23.
- 5 For more information visit: [https://www.sas.com/da\\_dk/customers/beefutures.html](https://www.sas.com/da_dk/customers/beefutures.html).
- 6 Trase is a not-for-profit initiative founded in 2015 by the Stockholm Environment Institute (SEI) and Global Canopy to bring transparency to deforestation and the agricultural commodity trade. For more information visit: <https://trase.earth/about>.
- 7 For more information visit: [https://www.sas.com/sv\\_se/customers/amesto-nextbridge.html](https://www.sas.com/sv_se/customers/amesto-nextbridge.html).
- 8 <https://www.cancer.gov/about-cancer/understanding/statistics>.
- 9 <https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality>.
- 10 For more information visit: <https://communities.sas.com/t5/SAS-Hackathon-Team-Profiles-Past/A-quest-to-solve-the-mysteries-of-cancer-with-AI-and-synthetic/ta-p/729989>.

### **Bibliography**

Anda, F., 2022. How to add value with AI [WWW Document]. Amesto NextBridge. URL <https://www.amestonextbridge.com/blog/how-to-add-value-with-ai> (last accessed 12.30.23).



- Cropp, B., 2022. Opening the black box: Explainable AI [WWW Document]. Amesto NextBridge. URL <https://www.amestonextbridge.com/blog/opening-the-black-box-explainable-ai> (last accessed 12.30.23).
- Cropp, B., Shantsev, D., 2022. Anatomy of a project: Problem solving using NLP for the Red Cross [WWW Document]. Amesto NextBridge. URL <https://www.amestonextbridge.com/blog/anatomy-of-a-project-problem-solving-using-nlp-for-the-red-cross> (last accessed 12.30.23).
- IFRC, 2021. Opening the black box – Building systems which helps us learn together [WWW Document]. IFRC.
- IFRCGo/DREF-NLP [WWW Document], 2022. GitHub. URL <https://github.com/IFRCGo/DREF-NLP> (last accessed 12.30.23).
- Smulowitz, M., 2023a. Leveraging family philanthropy to create corporate social value [WWW Document]. IbyIMD. URL <https://www.imd.org/ibyimd/live-events/leveraging-family-philanthropy-to-create-corporate-social-value/> (last accessed 12.30.23).
- Smulowitz, M., 2023b. How the triple bottom line can add impact to your family firm’s philanthropy [WWW Document]. IbyIMD. URL <https://www.imd.org/ibyimd/videos/leading-in-turbulent-times/leveraging-family-philanthropy-to-create-corporate-social-value/> (last accessed 12.30.23).
- Smulowitz, M., 2023c. Do you know how to measure your impact beyond profits? [WWW Document]. IbyIMD. URL <https://www.imd.org/ibyimd/brain-circuits/do-you-know-how-to-measure-your-impact-beyond-profits/> (last accessed 12.30.23).
- Vogel, P., Smulowitz, M., 2022. *Navigating your family’s philanthropic future across generations*. Lausanne: IMD Publishing. <https://www.imd.org/research-knowledge/family-business/reports/navigating-your-familys-philanthropic-future-across-generations/>

# DIGITAL STUNT PHILANTHROPY

## Mechanisms, impact, and ethics of using social media influencing for the greater good

*Monica Lea and Lucia Gomez*

### 1 Introduction

As the world becomes increasingly interconnected through technology, the integration of Artificial Intelligence (AI)-powered tools has emerged as a breakthrough force shaping the strategies and effectiveness of philanthropic endeavors (Henriksen & Richey, 2022). In the past years, traditional philanthropy has undergone a gradual transformation, and AI technology has started to be used for data-driven decision-making, to improve efficiency of internal operations, or to produce engaging and impactful communications. Social media platforms are especially powerful tools since they provide philanthropists and organizations with unparalleled opportunities to amplify their missions and mobilize support, largely due to their sophisticated AI-powered recommendation systems.

In this chapter, we tackle the landscape of AI-learning social media algorithms and how they can be utilized for philanthropic good. We explore the relationship between AI-learning algorithms and the promotion of philanthropic content on social media platforms, including several dimensions that underlie audience engagement with social media content that align with philanthropy-like content.

To further understand the properties of online philanthropic content, we allude to the concept of Digital Stunt Philanthropy (DSP) (Lea, 2023), a framework that aims to better understand the phenomenon of philanthropy influencers utilizing AI-powered social media platforms to develop philanthropic-focused content backed by private sponsors and nonprofit partners. Although social media trends may be a fleeting spectacle, this chapter uses the DSP framework to understand the mechanisms that drive philanthropic social media content and how those are reacted to and enhanced by AI-driven algorithms that respond to and amplify content.

There is much we do not know about the impact of AI on philanthropy, including its influence on social media platforms and their algorithmic learning. While DSP offers a new model for understanding social media-based philanthropy, it is unclear whether philanthropy influencers are engaging people to become more active in charitable causes or whether they remain a means of raising awareness about nonprofits and their causes. Furthermore, tech companies like Google (which owns YouTube), TikTok, and Meta, keep many facets of their algorithms confidential, so we do not know exactly how AI systems promote (reward) or suppress (punish) philanthropy-focused

content. As a result of this uncertainty, this chapter aims to shed light on how general philanthropists can benefit from utilizing AI-driven platforms as a means to showcase the impact of their philanthropy and further raise awareness of their causes.

To this end, we address the current trends in social media surrounding the philanthropy-based content of the popular YouTuber MrBeast on the social media video-sharing website YouTube. First, we provide an overview of the role of online platforms and AI in promoting philanthropy-based content, followed by a more thorough summary of DSP as a developing theoretical perspective to understand more about the aspects of user engagement and AI learning that boost philanthropy-centered content on social media platforms. Second, we examine the case study of YouTuber MrBeast and the success of his YouTube platforms including the primary MrBeast channel and philanthropic-centered YouTube channel, *Beast Philanthropy*. Through Natural Language Processing analysis of the content of these two channels, we discuss their comparative impact to assess what aspects beyond MrBeast's influencer/celebrity status drive engagement with the philanthropy-centric content. Third, we provide a discussion of the broader implications and ethical considerations at the heart of philanthropic-centered content being used for social media content creation and entertainment. Finally, we offer some recommendations for philanthropists who wish to engage social media users in their philanthropic causes. We hope that readers of this chapter will gain valuable insights into the significance of AI-driven social media algorithms in supporting philanthropy-centered content online.

## 2 The role of online platforms and AI in boosting philanthropy

In recent years, the online presence of philanthropy has continuously increased, revolutionizing the way individuals engage in charitable causes and organizations communicate about their initiatives (Taylor et al., 2012). In contrast with traditional channels for philanthropy such as personal connections, events, or non-web-based mass appeals to action, the dawn of the digital age has allowed online platforms to become powerful catalysts for social change, offering an unprecedented scale of impact from local stakeholders to a global audience.

Among other digital platforms used to showcase philanthropic causes, social media stands out as a major driver of the democratization of philanthropy (Soder, 2009). Specifically, social media's ability to connect people globally and in real time allows users to see other users as voices for philanthropy. Social media has thus become a fertile ground for grassroots movements and spontaneous acts of generosity, in part, due to AI-driven algorithms' ability to connect users who share similar interests (Zhou et al., 2012).

The diversity and quantity of social media platforms continue to flourish, demonstrating a transformative change in societies and communication patterns, including TikTok, Instagram, Facebook, LinkedIn, WhatsApp, and X/Twitter. Among these sites, YouTube occupies a unique position as the second most used social media app globally after Facebook and serves as one of the largest video-sharing platforms for users to share video content (Statista, 2024). YouTube experienced a meteoric growth in its onset activity (from 30,000 daily views in May 2005 to 25,000,000 daily views in January 2006), and the website's acquisition by Google in October 2006 was a "game changer" in the social media landscape. This is, in part, due to their significant strategic incorporation of AI-based technology for video recommendation and curation (Covington et al., 2016).

At the core of YouTube's AI recommendation system lies a vast database of videos along with associated metadata that describes them (description, tags, etc.) as well as the users that interact with them (engagement metrics, demographics, etc.). With thousands of hours of content being

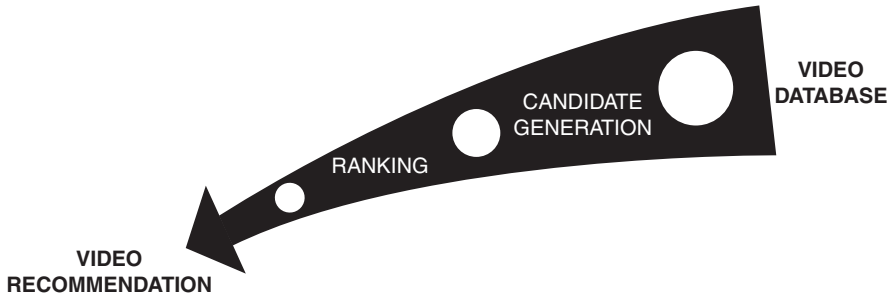


Figure 18.1 Schematic depicting the process of video selection for recommendation.

uploaded by users every day, this rich data repository serves as the foundation for the two fundamental algorithmic components that are essential to delivering the high-value and user-tailored content that YouTube audiences experience: *candidate generation* and *ranking networks* (see Figure 18.1).

The YouTube AI recommendation system begins with *candidate generation*, where the algorithm meticulously sifts through the extensive database and compares the user’s historical activity with that of users with similar interests and viewing behavior to identify a curated selection of candidate videos that are likely to match a given user’s preferences. Next, the *ranking networks* aspect of the YouTube AI system comes into play to further refine and prioritize this initial selection. This is done by evaluating the performance and appeal potential of each candidate’s video based on users’ behavioral metrics such as watch time, likes, comments, shares, or click-through rates. This evaluation results in behaviorally informed video ranking, placing those with higher scores at the top of the list for recommendation to a given user.

Together, the *candidate generation* and *ranking networks* algorithms measure overall content performance and engagement by examining users’ interaction with videos and, indirectly, provide a proxy measure of content quality. Through this procedure, the YouTube AI system aims to identify and recommend specific videos that match each user’s preferences and characteristics, and that have demonstrated a high level of overall user satisfaction and interaction with the broader YouTube community. As a result, recommendations are not only personalized (two different users might thus be recommended with very different content) but also reflect the most engaging and impactful content available on the platform for each user behavioral type.

As discussed, AI algorithms used by YouTube do not analyze video content directly but rather rely on users’ interactions with metadata-tagged content. Consistently, the YouTube creators platform states: “Our algorithm doesn’t pay attention to videos, it pays attention to viewers. So, rather than trying to make videos that’ll make an algorithm happy, focus on making videos that make your viewers happy” (YouTube, 2024). For instance, while the visual appeal of a video thumbnail may significantly impact viewer engagement and subsequently influence performance, YouTube AI learns from temporal human behavior rather than directly analyzing the thumbnails themselves. This distinction is fundamental to understanding that the foundation of YouTube’s AI recommendation system is rooted in human behavior, and therefore built on the premise that user engagement is an indicator of the social relevance and quality of content (Airoldi et al., 2016; Bendersky et al., 2014; Cheng et al., 2008; Davidson et al., 2010).

However, this reliance on human behavior as the primary determinant of content recommendations also introduces the potential for biases to emerge within the AI system. These often manifest

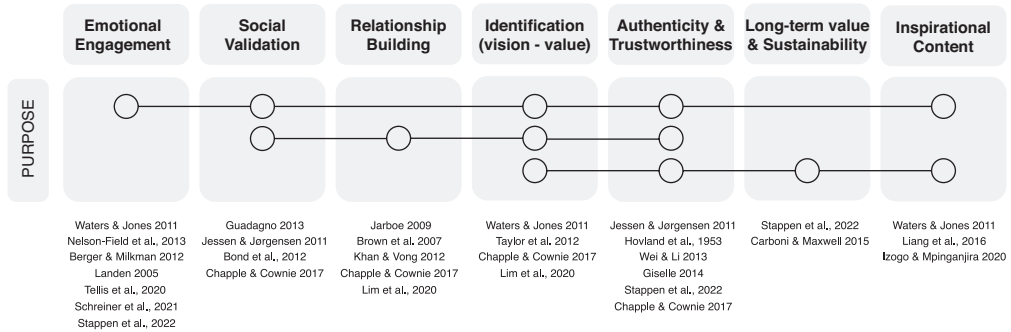


Figure 18.2 Figure summarizing seven dimensions underlying audience engagement with social media content that aligns with philanthropy-like content.

as systematic patterns observed in mass consumption habits, where certain types of content are disproportionately favored or promoted based on prevailing trends and virality. This way, videos featuring sensationalist or clickbait-style content may garner higher levels of engagement due to their ability to pique curiosity or provoke strong emotional reactions, leading the algorithm to prioritize similar content in recommendations. Ultimately, content creators who adhere to successful trends and cater to dominant demographics in their content creation may enjoy greater visibility and success on the platform, perpetuating existing biases and limiting the diversity of recommended content.

The aforementioned propensity for content to generate high levels of engagement through eliciting an emotional reaction in viewers holds particular relevance in the context of philanthropic content. Philanthropic activities’ content, which typically addresses pressing social and/or environmental challenges, has a high potential to resonate deeply with viewers and to captivate and mobilize audiences. In this context and considering the continued rise of social media use globally, we explore the proven correlates of attractive content, with a particular focus on how they characterize philanthropic content popularity. Although many of these dimensions are related to the content itself, such as *richness*, *originality*, and *length* (Figueiredo et al., 2014; Sabate et al., 2014), or to the behavior of the content creator, such as *posting frequency* (Salazar, 2017) or *collaborative posting* (Koch et al., 2018), we pay particular attention to content characteristics related to audience engagement as a means to identify the key factors that turn philanthropic content engaging (see Figure 18.2).

## 2.1 Emotional engagement

Emotional engagement has been repeatedly identified as a strong correlate of attractiveness in YouTube content (Tellis et al., 2020). Additionally, Waters and Jones (2011) found that the success of YouTube videos of philanthropic organizations was strengthened by their emotional engagement component. The showcase of highly emotive scenes and/or stories and the sentiment experienced by the viewer were reported to be the primary drivers of video sharing and, thus, virality (Nelson-Field et al., 2013). In addition, content with a positive valence of high arousal has been found to be the most successful (Berger & Milkman, 2012; Schreiner et al., 2021). This is not new to philanthropy, which strategically appeals to positive emotions for their strategic video communication (Landen, 2005).

## **2.2 Social validation**

Social validation is a two-way phenomenon that contributes to the popularity of content (Guadagno et al., 2013). On the one hand, videos that receive the most likes and shares are signaled as socially validated and thus promoted through AI recommendation (Jessen & Jørgensen, 2011). On the other hand, social media consumers tend to gravitate toward content that highlights socially valued traits such as empathy or compassion, which are omnipresent in philanthropy-related content. Overall, social validation mechanisms operate similarly in online and offline contexts (Bond et al., 2012), and content creators capitalize on this mechanism by which “for good” causes gain approval and connect with viewers willing to identify with them.

## **2.3 Relationship building**

Long recognized by philanthropists through call-to-action purposed campaigns, *relationship building*-based marketing also predicts YouTube audience engagement and retention (Jarboe, 2009). Indeed, YouTube and other social media platforms offer the opportunity to create a sense of community and exchange among users beyond punctual content, which strongly aligns with long-term audience growth and engagement (Brown et al., 2007; Feroz Khan & Vong, 2014). In addition, the formation of parasocial relationships between content creators and audiences is a phenomenon that enhances long-term user engagement, as the viewer progressively perceives the creator as a friend (Lim et al., 2020). Similarly, philanthropic organizations and supporters often build long-term meaningful relationships that could also be cultivated through social media by fostering familiarity and connection. Thus, the development of social media-fueled relationships through content and mission sharing, as well as digital exchange, can provide benefits for philanthropic branding.

## **2.4 Identification through vision and value alignment**

Research analyzing YouTube videos produced by nonprofit organizations reveals that the most popular content tends to focus on articulating a clear philanthropic purpose, vision, and alignment with core values (Waters & Jones, 2011). Notably, when the shared vision and values of the organization resonate with those of its audience, it fosters a *sense of identity* and connection, increasing the likelihood that viewers will share the content and subsequently receive algorithmic recommendations (Taylor et al., 2012). This evidence underscores the importance of crafting content that effectively communicates both the organization’s vision and values, as this type of content also establishes a meaningful connection with the audience, ultimately driving engagement and extending the reach of philanthropic initiatives.

## **2.5 Authenticity and trustworthiness**

Conveying authenticity and trust is and has been historically of clear importance for philanthropists to engage donors in their missions (Konstantinou & Jones, 2022; Martin, 1994), a strategy supported by the Social Credibility Theory (Hovland et al., 1953). Moreover, the degree of influence of a message is generally based on the believability of the source (Kok Wei & Li, 2013), which is also correlated with the transparency of the message (Auger, 2014). In support of this, evidence shows that philanthropic videos showcasing exemplary programs and services from organizations are more popular on YouTube (Waters & Jones, 2011). Content creators who promote transparent content increase their credibility and build stronger and longer-term engagement under

loyalty branding, partially through the formation of parasocial relationships (Chapple & Cownie, 2017). Credibility and trust are also modulated by social reaction to content, and thus, this dimension is highly related to the social validation described above, which ultimately also shapes AI recommendation (Jessen & Jørgensen, 2011).

## **2.6 Long-term value and sustainability**

Channels that have a history of constant value creation through high-quality regular publishing are more likely to succeed with AI-based recommendations, as well as being “secure recommendations.” Philanthropic organizations established to sustain long-term missions are candidates to benefit from this competitive advantage. In this line, past research analyzing user engagement with the content of nonprofits’ Facebook pages has shown that those profiles with a longer past history of value creation through user engagement are significantly more likely to succeed in terms of audience responsiveness for future publications (Carboni & Maxwell, 2015).

## **2.7 Inspirational content**

Inspirational content thrives on YouTube. It taps into deep-seated human needs by offering hope, optimism, and belief in a better future while at the same time fostering a sense of growth, shared purpose, and belonging. Philanthropic missions, by nature, promote actions for the common good and awakening engagement. Indeed, inspirational content has been found to be present in most popular videos released by philanthropic organizations (Waters & Jones, 2011), and the use of inspirational content has also been shown to be a powerful driver of charitable behavior (Zhao et al., 2023).

In essence, the use of digital content via platforms such as YouTube represents a potent strategy for philanthropic organizations to cultivate connections with their audiences. By creating genuine and credible videos that highlight their mission, values, and impactful initiatives, these organizations can effectively captivate donors and foster unwavering loyalty. Furthermore, by maintaining a consistent stream of high-caliber content over time, philanthropic entities can forge a reputable image and earn the trust of their viewers. Overall, integrating inspirational narratives and inclusive messaging serves to deepen the emotional bond and foster a collective sense of purpose between the audience and the philanthropic entity. Utilizing social media platforms like YouTube offers philanthropic organizations a distinct avenue to extend their reach and engage a broader demographic. By leveraging the persuasive force of storytelling, emotional resonance, and inclusive communication, philanthropic entities can motivate viewers to actively support their endeavors. Through strategic deployment of digital content and leveraging the functionalities of social media platforms, philanthropic organizations can proficiently convey their mission, galvanize action, and cultivate enduring partnerships with their audience.

While the use of digital content and social media publishing and interaction proves beneficial in many cases, past surveys on nonprofits’ usage of online platforms show that up to 74% of organizations use social media for one-way<sup>1</sup> communication instead of two-way<sup>2</sup> engaging exchanges (Carboni & Maxwell, 2015). Factors such as the lack of human power to keep a constant flow of communication may hinder the power of online relationship building for philanthropy. However, one-way communication can also lead to successful connection with nonprofit actions, through the creation of entertaining and engaging content. Through this lens, DSP appears as a novel manner to attract audiences to emotion-eliciting causes that are ultimately rewarded by social media AI algorithms.

### **3 Digital stunt philanthropy**

There have been several successful initiatives, including the widely publicized 2014 ALS (amyotrophic lateral sclerosis) Ice Bucket Challenge, which transformed and upscaled social engagement with charitable causes, and numerous other endeavors have followed (Pressgrove et al., 2018). In this trend, individuals who participated in the challenge recorded short videos of participants having a bucket of icy water poured over their heads. These clips were then shared across various social media platforms, with participants encouraging others to both replicate the challenge and contribute to the ALS Association. Overall, the campaign raised an estimated \$115 million for ALS-related research in 2014 and over 17 million people participated in the Ice Bucket Challenge. What aspects of the ALS Ice Bucket Challenge captured the attention of social media users to share, replicate, and, perhaps most importantly, donate? Pressgrove et al. (2018) found that content that promoted an arousal of positive emotions, including admiration, awe, and humor, was more frequent than negative emotive content. Consequently, philanthropic-focused algorithmically engaging content should promote positive sentiments in viewers as social media users' preferences lean toward positive philanthropic content.

#### **3.1 What is digital stunt philanthropy?**

DSP provides a framework for better understanding the relationship between AI-driven social media platforms and the promotion of philanthropic-focused content. The phrase “stunt philanthropy” is relatively new to the cultural zeitgeist (Khan, 2023), and research on the impact of stunt philanthropy is completely absent from academic literature despite its growing popularity. For the purposes of this chapter, DSP can be defined as an online content creator in partnership with a private sector sponsor, who engages an individual, group of people, or charitable organization in the giving of money, material goods, and/or services in an entertainment setting that is documented and uploaded online for the exposure and engagement with viewers.

There are several components of DSP, including: (1) the focus of the content is reliant on a form of charitable giving; (2) the philanthropic content is shared publicly to entertain in part; (3) there is space and encouragement for audience engagement; (4) there is a call to action to the audience within the content to raise awareness or also participate in philanthropic giving; (5) the content is deemed algorithm-friendly and will be supported by the online platform to which it is uploaded; and (6) the content presented is deemed sponsor-friendly by private partners who provide financial support to the content creator, either through direct partnerships or through platform-facilitated support. The following paragraphs provide a more in-depth examination of these features along with Table 18.1, which provides a summary of these facets of DSP and their associated definitions.

While the focus on philanthropic giving is largely self-explanatory, it deserves some attention in how this form of giving looks like directly online. The philanthropic activity of DSP posted online can range from donating a large sum of money to an organization to simply giving a person in need a few dollars, food, or some other charitable good or service. What makes stunt philanthropy unique from other forms of charity is its direct relationship to performance or “the stunt.” Stunts have become a standard part of the social media landscape. An online stunt refers to a planned, attention-grabbing action or event conducted primarily through social media, with the intention of garnering widespread attention, engagement, or viral spread online. Online stunts are often designed to be unconventional, surprising, or provocative in nature, in order to capture the interest and imagination of internet users, generate buzz, and increase visibility for a brand, individual, or cause. Online stunts are often crafted with the intention of spreading rapidly across social media



Table 18.1 Key components of digital stunt philanthropy

<i>Facet</i>	<i>Definition</i>
Charitable giving	The act of voluntarily giving money, goods, or services to those in need or to an NGO/nonprofit organization that serves the public good.
Public-facing entertainment	The content is created to be engaging, to consume and provide enjoyment, pleasure, or amusement to the viewing audience.
Audience engagement	The interactions, reactions, and participation of users with digital content or platforms. It encompasses various activities and metrics that indicate how actively involved and interested an audience is with online content.
Call to action	A prompt or directive designed to encourage an immediate response or specific action from the audience. E.g., donating, sharing, subscribing, liking, commenting.
Algorithm-friendly	The content is designed, structured, or formatted in a way that facilitates its recognition, interpretation, and processing by algorithms effectively and efficiently.
Sponsor-friendly	The content is suitable or attractive for sponsorship by businesses, organizations, or individuals. In the context of events, content creation, or projects, being sponsor-friendly means that the activity or initiative offers opportunities for sponsors to align their brand, products, or services with the target audience and goals of the endeavor.

platforms and online communities, leveraging the power of sharing, retweeting, and reposting to reach a wide audience in order to achieve goals of virality.

The “ice bucket challenge” was a pre-existing online trend in which participants filmed themselves dumping ice water over their heads and nominated others to do the same; however, it was only after the challenge reached ALS patient and former Boston College baseball captain Pete Frates and his network that the challenge evolved toward donations to ALS research (Wicks, 2014). In addition to leveraging engaging content for audiences to learn about charitable causes, these stunts also provide entertainment for social media users. Online stunts can take various forms, including elaborate pranks, publicity stunts, social media challenges, flash mobs, interactive campaigns, or other forms of digital spectacle. When executed effectively, online stunts can generate significant visibility for charitable causes and contribute to the broader conversation both online and offline.

### ***3.2 Stunt, celebrity influencer, and use of spectacle online***

Social media content creators, also referred to as influencers when they achieve an impactful status or follower count on an online platform, have begun to use this influence to raise awareness and engage in various forms of social media-driven charitable activities. Social media content featuring popular influencers or celebrities tends to have a greater potential to trend or go viral, and there have been documented cases and studies exploring this phenomenon in connection with philanthropy and activist causes (Bennett, 2014). Due to their ability to influence fans and culture, celebrity philanthropy has emerged as its own subfield of philanthropic study (Hassid & Jeffreys, 2015). However, research has found that celebrity-based philanthropy, while engaging, poses its own unique set of challenges for nonprofit organizations and their celebrity status advocates

(Jeffreys & Allatson, 2015). Others argue that instead of mobilizing audiences, celebrity-endorsed philanthropic efforts encourage fans to be passive donators and product consumers rather than active participants in the collective efforts at the heart of philanthropy (Kapoor, 2012). Social media influencers, like MrBeast, while celebrity-like in their status and attention from followers, utilize their platforms in a unique way that is specifically catered to social media platforms and audience engagement. Therefore, it is necessary to analyze what aspects of this phenomenon may be beneficial to current and future philanthropists who aspire to effectively engage in philanthropy-focused social media initiatives.

### **3.3 Audience engagement on social media**

One aspect of social media platforms compared to traditional media is the space for content creators and audience engagement through the use of public comment sections and the ability to share videos and posts publicly and privately with others. With the ability for viewers to engage, the stunt for the audience becomes even more significant for the content creator as they receive direct audience feedback and engagement. Measurements for social media engagement developed by AI-driven algorithms involve a number of activities and metrics that indicate how actively involved and interested an audience is with online content. Some common forms of audience engagement online include:

- **Likes, shares, and comments:** Users react to content by expressing their opinions, sharing it with others, or engaging in discussions in the comment section;
- **Click-through rates (CTR):** The percentage of users who click on a link or call to action within content, indicating their interest and engagement;
- **Time spent on page:** The duration users spend consuming content on a webpage, indicating their level of engagement and interest;
- **Social media interactions:** Interactions such as retweets, mentions, direct messages, and follows on social media platforms that indicate engagement with content or profiles.

Overall, online audience engagement is crucial for content creators and organizations as it indicates the desirability of their content to the algorithm and helps to further promote and interact with online communities. It is important for stunt philanthropy to be engaging and provide some entertainment value to social media users. In addition to engaging the audience, content creators can also include a call to action within their content to bolster their own charitable giving and/or to increase awareness of a specific issue depicted in the video.

The last two features of stunt philanthropy are necessary for it to thrive in today's digital age, where content is seen as *algorithm-friendly* and *sponsor-friendly*. These are two separate concepts, but both must be successfully navigated for content to thrive and remain relevant in the current digital space. As O'Brien (2022) outlines in plain terms, "algorithms are used on social media to sort content in a user's feed. With so much content available, it's a way for social networks to prioritize content they think a user will like based on a number of factors" (para. 9). *Algorithm-friendly* content is designed to be effectively and efficiently recognized and processed by algorithms, taking into account aspects that inform AI learning on social media websites. On YouTube, specifically, these factors can include the frequency with which a content creator posts, the length of the video posted, and the appeal of a thumbnail (the image that previews a YouTube video), which

encourages users to click on a video and boost CTR. Other examples of algorithm-friendly optimization determinants for DSP include:

- **Search engine optimization (SEO):** Algorithm-friendly content is crafted with keywords, metadata, and formatting that search engine algorithms favor, making it more likely to rank higher in search engine results pages (SERPs);
- **Social media:** Algorithm-friendly posts or content is tailored to meet the criteria set by social media algorithms for visibility and engagement, such as using relevant hashtags, posting at optimal times, or generating high levels of interaction;
- **Machine learning:** Algorithm-friendly data sets and features are prepared and pre-processed to enhance the performance of machine learning models, ensuring that algorithms can effectively learn patterns and make accurate predictions;

As a result, algorithm-friendly content is digital content that maximizes platform criteria; however, remaining algorithm-friendly remains a moving goal post for individual creators and organizations, who must constantly reevaluate strategies as the metrics measured by algorithms and AI learning also drive change.

*Sponsor-friendly* content is similar to algorithm-friendly content in that creators must also keep in mind how their content is being perceived by the algorithm for access to monetization and by private companies that may become sponsors of users' content. For social media platforms, being sponsor-friendly means that the activity or initiative offers opportunities for private company sponsors to align their brand with a creator's audience. Platforms like YouTube provide creators with the opportunity to monetize content through the process of AdSense, where advertisers allow ads for their brands to be displayed before, after, or during a YouTube video. This AdSense is then turned into income for content creators, with higher AdSense being generated by videos with higher views and watch times. In order to be deemed sponsor-friendly through the use of AdSense, content creators must be algorithm-friendly by default to even be considered. Furthermore, private companies will also seek out content creators (and vice versa) to provide sponsors for the direct content of the video. This results in the creator providing an ad that is read directly into the video for the audience to demonstrate their support and alignment with a particular brand. It is this relationship between AI-driven algorithms on social media platforms, digital monetization by private entities, and content creators with ever-growing spheres of influence on platforms that is driving the existing phenomenon of DSP.

## 4 The case of MrBeast and Beast Philanthropy

### 4.1 Background

Perhaps no one remains more influential to this DSP phenomenon and trend in content creation than Jimmy Donaldson, better known as YouTube influencer and social media personality MrBeast, who has become the leading creator in this genre and is one of the most popular content creators on the platform with over 237 million subscribers as of February 2024 (MrBeast, n.d.). He holds many titles including YouTuber, philanthropist, entrepreneur, and content creator. He has gained widespread fame for his attention-grabbing stunt videos, which often involve large-scale challenges, acts of kindness, or entertaining experiments. MrBeast has become widely known for his attention-grabbing videos on his multiple YouTube channels that

share his moniker, including *MrBeast* and *Beast Philanthropy*, where his content often focuses on large-scale stunts, challenges, and philanthropic efforts. MrBeast has created several online philanthropic initiatives, including #TeamTrees, #TeamSeas, and a YouTube channel dedicated to his 501(c)3 nonprofit organization *Beast Philanthropy* (Beast Philanthropy, n.d.; TeamSeas, n.d.; TeamTrees, n.d.).

Created in September 2020, the YouTube channel *Beast Philanthropy* serves as a branch of MrBeast's content, where he engages in various acts of charity and philanthropy on an extravagant scale that encapsulate the spirit of DSP. As stated on the About page, "100% of the profits from my ad revenue, merch sales, and sponsorships will go towards making the world a better place!" (Beast Philanthropy, n.d.). By creating effective sponsor-friendly content, MrBeast utilizes collaborations with private corporate sponsors to provide financial backing for the acts of charitable giving shared on the *Beast Philanthropy* channel. Currently, Donaldson represents an apex in the larger history of charitable online content. MrBeast's philanthropic efforts are indisputably linked and driven by algorithmic success. As Donaldson states himself, "In our case, we reinvest it all. So year over year, whatever we make, we just spend it on videos and the next year is higher. And I just keep doing it and I just pray it keeps working" (Kennedy, 2021, para, 22). However, *how* Donaldson was able to acquire popularity and support from sponsors to provide significant contributions for philanthropic-centered digital content is important to answer, especially for those unacquainted with social media platforms.

In 2019, MrBeast and YouTuber Mark Rober launched #TeamTrees, a collaborative stunt philanthropy effort in partnership with the Arbor Day Foundation to raise funds to plant trees around the world (#TeamTrees, n.d.). The challenge was to raise \$20 million to plant 20 million trees. With widespread support from other popular YouTube creators, celebrities, and various online communities, the initiative exceeded expectations, raising over \$24 million and garnering over 100,000,000 views for their first video.<sup>3</sup> MrBeast himself made a substantial donation to kick-start the initiative, contributing \$100,000 of his own money. He then launched a series of videos on his YouTube channel, where he collaborated with other creators and undertook various challenges and fundraising activities to raise additional funds for #TeamTrees. #TeamTrees partnered with the Arbor Day Foundation, the world's largest membership-based nonprofit dedicated to planting trees with over 50 years of experience (Arbor Day Foundation, n.d.). The success of this initiative was facilitated, in part, by direct donation mechanisms available to social media users to respond to the call to action to match a dollar for a tree. The campaign gained significant traction on social media platforms, with supporters using the hashtag #TeamTrees to share the initiative and encourage others to contribute. Contributions made through TeamTrees.org or the YouTube donation button were given to the Arbor Day Foundation to support #TeamTrees' tree-planting efforts. #TeamTrees exemplifies how digital platforms can be leveraged to make a significant impact on global issues.

Following the success of #TeamTrees, MrBeast and Mark Rober launched #TeamSeas as a charitable initiative aimed at trash removal with the goal of removing 30 million pounds of plastic and trash from oceans, rivers, and beaches by the end of 2022. Similar to #TeamTrees, MrBeast leveraged both his network of popular peers and his social media audience in partnership with two nonprofit organizations, Ocean Conservancy and The Ocean Cleanup, to remove one pound of trash for every dollar donated. The hashtag #TeamSeas was used across various social media platforms to track and promote the initiative and encourage further sharing in line with the call to action put to viewers. Like #TeamTrees, #TeamSeas demonstrates the power of online communities and social media influencers to mobilize support for important environmental causes.

## 4.2 Methods

To understand the content that characterizes the *MrBeast* and *Beast Philanthropy* YouTube channels and to decipher their specific and differentiated offerings, we applied an AI-based analytical pipeline. Specifically, we used the YouTube Data v3 API to gather all of their video descriptions and titles ( $n = 777$  for MrBeast and  $n = 39$  for Beast Philanthropy videos, up to February 20, 2024). We also collected video statistics to report on video history and performance per channel. We used the open source BGE-M3 large language model (LLM) (Chen et al., 2024) for text embedding via the python implementation of FlagEmbedding, followed by UMAP dimensionality reduction (McInnes et al., 2020) in 2D to visualize the similarity space of the analyzed videos. This UMAP network-based reduction was then used for video clustering with the Leiden optimization algorithm (see Figure 18.3). Finally, to make the LLM+UMAP output explainable, we deconstructed and mapped the word weights (Term Frequency – Inverse Document Frequency; TF-IDF) (Sparck Jones, 1972) per video description and calculated their log-fold change count per cluster (see Figure 18.4).

Comparing the MrBeast and Beast Philanthropy YouTube channels, our analysis provides further insight into popular philanthropy-related content. While both the MrBeast and Beast Philanthropy YouTube channels share a commitment to philanthropy, they differ in their approach, content focus, and audience engagement strategies. MrBeast’s main channel offers a diverse range of content, including philanthropy, entertainment, and challenges, while Beast Philanthropy is solely dedicated to showcasing charitable activities and making a positive impact on society. Figure 18.3 provides a visual of this cluster analysis. The results indicated that MrBeast and Beast Philanthropy videos are largely different, as shown by the little overlap in the NLP coordinate space (UMAP), with only a subset of their content being similar, grouped in cluster 3 (gray-shaded area, Figure 18.3). The NLP subspace including videos from the two analyzed groups of videos (circles cluster) has a well-balanced proportion of content from the two channels (44 videos for MrBeast and 38 for Beast Philanthropy). Conversely, videos from Beast Philanthropy were absent in other clusters, those that were characterized by MrBeast-specific content. This NLP scenario is, thus, suitable for contrasting the content of the entertainment-centered MrBeast videos with the philanthropy-centered Beast Philanthropy content (see Figure 18.4).

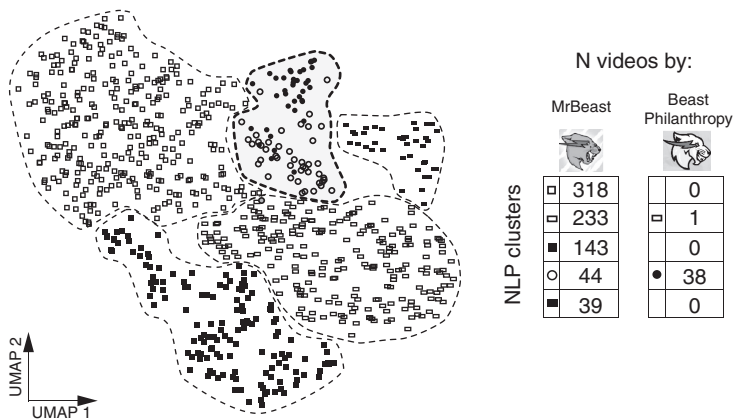


Figure 18.3 Natural Language Processing indicates that little overlap exists between MrBeast and Beast Philanthropy YouTube channels’ content.

Differential content - Philanthropy vs Entertainment

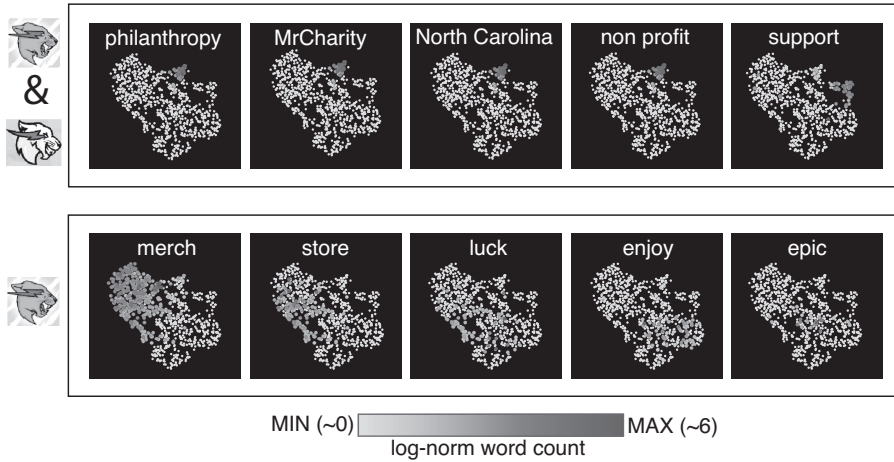


Figure 18.4 Top differential content keywords reveal a common philanthropy content and a differing use of entertainment and marketing.

As argued previously, when analytically comparing the content focus of each channel, we here further reveal that the MrBeast channel includes a variety of content highly entertaining as well including philanthropy. Instead, the Beast Philanthropy channel focuses predominantly on philanthropic activities.

We discovered that cluster 3 (purple in Figure 18.3; top row in Figure 18.4) videos’ content largely differed from the content of videos populating other clusters (all other colors in Figure 18.3, bottom row in Figure 18.4). Specifically, we found that cluster 3 videos’ content is very philanthropy centered and characterized by differential words such as “philanthropy,” “MrCharity,” or “non profit.” In contrast, content from other clusters is differential by either their marketing and commercial messages (keywords such as “merch” or “store”) or the stunt component of entertainment (keywords such as “enjoy” or “epic”). These observations indicate that while MrBeast’s content is diverse and combines the three key ingredients of philanthropy, entertainment, and commercial orientation, Beast Philanthropy primarily utilizes only the first of those.

With access to this widespread dissemination, YouTube channels such as MrBeast and Beast Philanthropy enjoy an expansive viewership base, often numbering in the millions. This extensive reach facilitates the dissemination of philanthropic content to diverse and geographically dispersed audiences, thereby enhancing the potential for influence and impact.

Finally, we evaluated the general performance of both channels, showing in Figure 18.5 the main statistics per video: publication date, number of views, number of likes, and number of comments. In addition to the content, these two channels differ in one main parameter: their lifetime. It is worth noting that the MrBeast channel is much older and holds much more content than Beast Philanthropy, which in itself is a competitive advantage for both audience engagement and AI recommendation. Other reported metrics are not different between the channels, suggesting that these two channels generate similar community engagement in terms of consumption (n views), user satisfaction (n likes), and interaction (n comments).

To be successful on digital platforms, creators must tailor their content to the preferences of the digital platform they are posting to. This can include creating *eye-catching thumbnails*, engaging

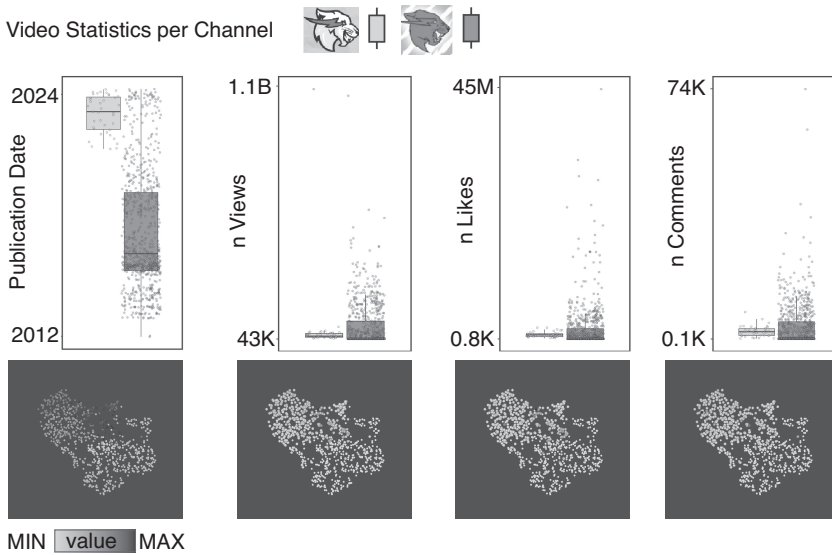


Figure 18.5 Video Statistics for MrBeast and Beast Philanthropy YouTube channels.

in *trending topics*, and *family-friendly branding* strategies, to increase viewers' click rate and the *probability of sponsorship*. MrBeast has largely perfected this practice, as demonstrated by his high viewership. There are several elements to MrBeast's content creation that allow him to achieve consistent virality, in contrast to the more fleeting virality of specific trends. First, MrBeast invests a significant amount of capital and time in the production of his videos, including millions of dollars and multi-shoot days scheduled per video.<sup>4</sup> Although his videos are approximately 10–20 minutes long, the amount of production value placed in each video increases its overall entertainment value. Second, Donaldson does not work alone. Instead, he is the head of a company and a team of people who help him produce his videos and the MrBeast brand (Cacich, 2021). Having a team dedicated to producing viral content is an advantage not all content creators or philanthropic organizations have to maximize production and ultimately profit. It is therefore the partnership of content creators, nonprofit partners, and private company sponsors that produces the greatest success.

Looking directly at the role of algorithm-friendly strategies, MrBeast credits several things for maximizing the metrics that the algorithm currently favors for the benefit of his videos. One metric is the value YouTube places on the video CTR and viewer retention once the video is clicked. Videos with high CTRs and high viewer retention are promoted higher in the algorithm. In addition to a fast-paced editing style, to keep audience members engaged, MrBeast provides a small glimpse of the final product the video focuses on to entice viewers to watch the entire video.<sup>5</sup> Lastly, MrBeast has a specific strategy that involves optimizing the thumbnails of his YouTube videos (Perelli, 2021). This typically includes a photoshopped image of Donaldson himself to help cue to the viewer that this is a MrBeast video, along with a stylized image of the video's topic. The backgrounds are minimal in order to enhance the colorful photoshopped images in the forefront of the thumbnail. Donaldson is clear that these thumbnails do not fall into the category of "clickbait" thumbnails, where creators intentionally promise something in the thumbnail that the viewer will not see in the actual video. Instead, he lures viewers in with his grand stunts and then delivers them

in the video. It is some of these characteristics that have made MrBeast algorithmically popular and continue to bolster his success.

Due to the significant number of views of stunt philanthropy content, corporations are enticed to sponsor videos in order to gain access to this mass market share of millions of viewers as well as gain a positive reputation among video viewers. As Miller and Hogg (2023) discuss this aspect of the MrBeast case study,

in what has proven to be a virtuous circle, these large prizes and giveaways attract increasingly larger audiences, creating even more extreme amounts of revenue to give away in subsequent videos, which in turn creates even larger audiences and thus more revenue.

In this way, the DSP cycle can continue to prosper and grow.

## **5 Impact**

Characterized by attention-grabbing online initiatives to promote philanthropic causes, DSP has had a significant impact on the success of charitable endeavors in several ways. By leveraging social media platforms' AI-learning algorithms and recommendations, DSP harnesses the power of social media to showcase philanthropic activities, effectively reaching a wider audience and increasing visibility for charitable causes. By sharing compelling content and engaging stories, philanthropists can effectively communicate their mission and impact to a global audience. By incorporating elements of entertainment into philanthropic initiatives, creators are able to help attract and engage a larger audience. By making philanthropic content entertaining and engaging, DSP can increase its reach and effectiveness in raising awareness and driving action for charitable causes.

There are medium-specific benefits to using AI for promotion, engagement, and monetization. AI technology can be harnessed to promote philanthropic content to the right audience, maximizing its impact and reach. Digital platforms offer opportunities for monetization through advertisements, sponsorships, and partnerships, allowing philanthropists to generate revenue to support their causes (Kopf, 2020). AI algorithms can identify and promote philanthropic content that resonates with audiences, ensuring continued engagement and support for charitable causes (Pressgrove et al., 2016). Online platforms democratize access to information and opportunities, allowing individuals from diverse backgrounds to participate in and contribute to charitable causes virtually (Song et al., 2015).

When philanthropists use digital platforms to showcase their actions and impact, they can inspire others to get involved, creating a snowball effect of support. Based on the significant viewership of MrBeast's philanthropic videos, philanthropic social media content has the potential to motivate and inspire audience members to become philanthropists themselves. As stated in the About section of *Beast Philanthropy's* website, "MrBeast is ready to inspire the next generation of philanthropists."<sup>6</sup> By showcasing the impact of charitable giving and the joy of giving back, DSP initiatives can expand the philanthropic sector and encourage more people to contribute to positive social change.

## **6 Ethical issues and implications**

While DSP has the potential to create a positive change within the philanthropic sector, there are several ethical issues to consider. As there is a notable power imbalance between philanthropists, who have the resources and platform to initiate large-scale campaigns, and the beneficiaries or



communities they aim to support, there is a need to consider the influence of power dynamics, the role of private actors shaping the public good, and accountability (Capone, 2023). There is a risk of beneficiaries being exploited for online content without their full consent to the implications of their participation. This raises questions about the ethics of using beneficiaries' stories and struggles for entertainment or promotional purposes. It is, therefore, essential for creators and nonprofit partners to be transparent in order to promote trust and informed consent.

This imbalance can further influence the dynamics of decision-making, representation, and impact. While online spaces are theoretically democratic, the algorithms used in social media platforms tend to emphasize content that generates engagement which can amplify biases and contribute to polarization (Arora et al., 2022). Existing literature acknowledges gender and racial biases that can be further exacerbated by social media algorithms (Fosch-Villaronga et al., 2021).

There are also questions about the long-term sustainability of DSP initiatives. While they can generate significant attention and support in the short term, there is a need to ensure that these efforts lead to meaningful and lasting change, rather than just fleeting moments of online engagement. While the ALS Ice Bucket challenge was an unforeseen success, by September of 2014, donations to the ALS Association had returned to pre-viral levels (Sohn, 2017). There is no one-size-fits-all approach to digital philanthropy, and the literature on virtual altruism continues to grow in understanding how social causes gain attention and generate larger collective support (Van Der Linden, 2017). While social media has the potential to amplify philanthropic efforts, many nonprofit organizations lack the resources and expertise to effectively utilize these platforms. It is therefore important for philanthropists to keep in mind the resources available to their organizations and networks to prepare for a sudden influx of attention.

It is important to consider the role of the private actors engaged in DSP (e.g., content creators, corporate sponsors, philanthropists, nonprofit organizations), as these actors wield considerable influence in shaping the landscape of digital philanthropic endeavors and the implications of this influence. The involvement of private actors can significantly impact the direction, focus, and outcomes of philanthropic initiatives, often reflecting their own priorities, values, and interests. Corporate sponsors, driven by branding objectives and corporate social responsibility (CSR) goals, may choose to align their philanthropic investments with issues that resonate with their target market or enhance their brand image (Sanzo et al., 2015). As a result, certain social causes or initiatives that closely align with corporate interests may receive disproportionate attention and funding, while others are sidelined.

Due to this stake given to private donors' interests rather than democratically decided issues, some argue that the influence of private actors in shaping philanthropic priorities can exacerbate existing economic inequalities and power imbalances (Van Dyk & Fourie, 2015). Wealthy individuals and private corporations with significant financial resources have the capacity to steer philanthropic efforts toward issues that benefit their own interests (Maclean et al., 2021). This is apparent even in MrBeast's work, which originated from his philanthropic work in his home state of North Carolina to provide relief during the COVID-19 pandemic in 2020.<sup>7</sup>

## **7 Conclusion**

The landscape of philanthropic giving is continuing to shift in the 21st century in ways we may not have predicted. Philanthropic-centered content has gained a niche and growing presence on social media platforms and among stakeholders engaged in stunt philanthropy. The DSP framework provides a structured approach to understanding the various components involved in creating impactful philanthropic content online. From leveraging spectacle and shock value to harnessing

the power of social media virality, this framework illuminates the strategies employed by creators like MrBeast to maximize their philanthropic reach and impact.

The rise of MrBeast and his unique brand of philanthropy sheds light on the multifaceted dynamics at play in the realm of online altruism. By leveraging his massive following and engaging in large-scale philanthropic stunt video projects, MrBeast has created a platform for raising funds and addressing societal challenges such as hunger, homelessness, and unemployment. The reliance on algorithm-friendly and sponsor-friendly content highlights the intricate relationship between digital platforms, private company sponsorships, and content creators. With the additional support of algorithmic-driven success and corporate partners, this form of online philanthropy presents a way in which individuals, regardless of their relationship with traditional models of philanthropy, can easily and readily engage in the performance of charitable giving should they have the digital savvy to do so.

However, amid the excitement and admiration for digital stunt philanthropy, it is crucial to critically examine the ethical implications associated with creating content that revolves around acts of charity. From concerns about exploitation and sensationalism to questions surrounding the long-term sustainability and effectiveness of such initiatives, ethical considerations must be at the forefront of any discussion surrounding digital philanthropy. As we continue to navigate the evolving landscape of online altruism, it is imperative that creators, platforms, and audiences alike remain vigilant in their efforts to uphold ethical standards and ensure that the impact of digital philanthropy extends beyond mere spectacle to meaningful, lasting change in the lives of those in need. Only through thoughtful reflection, responsible action, and a commitment to transparency and accountability can we harness the full potential of digital media to drive positive social change in our increasingly interconnected world.

### Notes

- 1 One-way: digital content is used by organizations to communicate about their actions, but not to engage the audience in discussions or actions.
- 2 Two-way: posting digital content serves as an anchor for the audience's engagement in discussions or actions.
- 3 [https://www.youtube.com/watch?v=HPJKxAhLw5I&ab\\_channel=MrBeast](https://www.youtube.com/watch?v=HPJKxAhLw5I&ab_channel=MrBeast)
- 4 <https://en.wikipedia.org/wiki/Coffeezilla>
- 5 Coffeezilla. (2021, December 29). Mr. Beast's secret formula for going viral [Video]. YouTube. <https://youtu.be/6pMhBaG81MI>
- 6 <https://www.beastphilanthropy.org/about>
- 7 <https://www.beastphilanthropy.org/about>

### References

- Airoldi, M., Beraldo, D., & Gandini, A. (2016). Follow the algorithm: An exploratory investigation of music on YouTube. *Poetics*, 57, 1–13. <https://doi.org/10.1016/j.poetic.2016.05.001>
- Arbor Day Foundation (n.d.). Mission. <https://Www.Arborday.Org/About/>
- Arora, S. D., Singh, G. P., Chakraborty, A., & Maity, M. (2022). Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183, 121942.
- Auger, G. A. (2014). Trust me, Trust me not: An experimental analysis of the effect of transparency on organizations. *Journal of Public Relations Research*, 26(4), 325–343. <https://doi.org/10.1080/1062726X.2014.908722>
- Beast Philanthropy (n.d.). [YouTube Channel]. Retrieved February 2, 2024, from <https://www.youtube.com/@BeastPhilanthropy>
- Bendersky, M., Garcia-Pueyo, L., Harmsen, J., Josifovski, V., & Lepikhin, D. (2014). Up next: Retrieval methods for large scale related video suggestion. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1769–1778. <https://doi.org/10.1145/2623330.2623344>

- Bennett, L. (2014). 'If we stick together we can do anything': Lady Gaga fandom, philanthropy and activism through social media. *Celebrity Studies*, 5(1–2), 138–152. <https://doi.org/10.1080/19392397.2013.813778>
- Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192–205. <https://doi.org/10.1509/jmr.10.0353>
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415), 295–298. <https://doi.org/10.1038/nature11421>
- Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, 21(3), 2–20. <https://doi.org/10.1002/dir.20082>
- Cacich, A. (2021). YouTuber MrBeast reached 30 million subscribers with a little help from his friends. *Distractify*. <https://www.distractify.com/p/mrbeast-crew>
- Capone, E. (2023). Addressing power dynamics in philanthropy through counter-storytelling. *Philanthropy & Education*, 6(2), 1–16.
- Carboni, J. L., & Maxwell, S. P. (2015). Effective social media engagement for nonprofits: What matters? *Journal of Public and Nonprofit Affairs*, 1(1), 18–28. <https://doi.org/10.20899/jpna.1.1.18-28>
- Chapple, C., & Cownie, F. (2017). An investigation into viewers' trust in and response towards disclosed paid-for endorsements by YouTube lifestyle Vloggers. *Journal of Promotional Communications*, 5(2). <https://promotionalcommunications.org/index.php/pc/article/view/95>
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). *BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation* (arXiv: 2402.03216). arXiv. <http://arxiv.org/abs/2402.03216>
- Cheng, X., Dale, C., & Liu, J. (2008). Statistics and social network of YouTube videos. *2008 16<sup>th</sup> International Workshop on Quality of Service*, 229–238. <https://doi.org/10.1109/IWQOS.2008.32>
- Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. <https://doi.org/10.1145/2959100.2959190>
- Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., & Sampath, D. (2010). The YouTube video recommendation system. *Proceedings of the Fourth ACM Conference on Recommender Systems*, 293–296. <https://doi.org/10.1145/1864708.1864770>
- Feroz Khan, G., & Vong, S. (2014). Virality over YouTube: An empirical analysis. *Internet Research*, 24(5), 629–647. <https://doi.org/10.1108/IntR-05-2013-0085>
- Figueiredo, F., Almeida, J. M., Benevenuto, F., & Gummadi, K. P. (2014). Does content determine information popularity in social media?: A case study of YouTube videos' content and their popularity. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 979–982. <https://doi.org/10.1145/2556288.2557285>
- Fosch-Villaronga, E., Poulsen, A., Søråa, R. A., & Custers, B. H. M. (2021). A little bird told me your gender: Gender inferences in social media. *Information Processing & Management*, 58(3), 102541.
- Guadagno, R. E., Muscanell, N. L., Rice, L. M., & Roberts, N. (2013). Social influence online: The impact of social validation and likability on compliance. *Psychology of Popular Media Culture*, 2(1), 51–60. <https://doi.org/10.1037/a0030592>
- Hassid, J., & Jeffreys, E. (2015). Doing good or doing nothing? Celebrity, media and philanthropy in China. *Third World Quarterly*, 36, 75–93.
- Henriksen, S. E., & Richey, L. A. (2022). Google's tech philanthropy: Capitalism and humanitarianism in the digital age. *Public Anthropologist*, 4(1), 21–50. <https://doi.org/10.1163/25891715-bja10030>
- Hovland, C. L., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change*. Yale University Press.
- Jarboe, G. (2009). *YouTube and video marketing: An hour a day* (1st ed). Wiley Technology Pub.
- Jeffreys, E., & Allatson, P. (Eds.) (2015). *Celebrity philanthropy*. Intellect.
- Jessen, J., & Jørgensen, A. H. (2011). Aggregated trustworthiness: Redefining online credibility through social validation. *First Monday*. <https://doi.org/10.5210/fm.v17i1.3731>
- Kapoor, I. (2012). *Celebrity humanitarianism: The ideology of global charity*. Routledge.
- Kennedy, M. (2021, December 16). YouTube star MrBeast rethinks old notions of philanthropy. *Associated Press*. <https://apnews.com/article/entertainment-technology-lifestyle-business-greenville-6a9751477b7376afd3962371ff11213e>

- Khan, K. (2023). Mr. Beast's stunt philanthropy takes fundraising to new heights – A game changer. *Impact Wealth*. <https://impactwealth.org/mr-beasts-stunt-philanthropy-takes-fundraising-to-new-heights-a-game-changer/>
- Koch, C., Lode, M., Stohr, D., Rizk, A., & Steinmetz, R. (2018). Collaborations on YouTube: From unsupervised detection to the impact on video and channel popularity. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4), 1–23. <https://doi.org/10.1145/3241054>
- Kok Wei, K., & Li, W., You. (2013). Measuring the impact of celebrity endorsement on consumer behavioural intentions: A study of Malaysian consumers. *International Journal of Sports Marketing and Sponsorship*, 14(3), 2–22. <https://doi.org/10.1108/IJSMS-14-03-2013-B002>
- Konstantinou, I., & Jones, K. (2022). Investigating GEN Z attitudes to charitable giving and donation behaviour: Social media, peers and authenticity. *Journal of Philanthropy and Marketing*, 27(3), e1764. <https://doi.org/10.1002/nvsm.1764>
- Kopf, S. (2020). “Rewarding good creators”: Corporate social media discourse on monetization schemes for content creators. *Social Media+ Society*, 6(4). <https://doi.org/10.1177/2056305120969877>
- Landen, H. M. (2005). *Marketing with digital video: How to create a winning video for your small business or non-profit* (2nd ed. completely rev). Oak Tree Press.
- Lea, M. (2023, May 30). After the ALS ice bucket challenge and the rise of MrBeast, stunt philanthropy might be here to stay. *The Conversation*. <https://theconversation.com/after-the-als-ice-bucket-challenge-and-the-rise-of-mrbeast-stunt-philanthropy-might-be-here-to-stay-205096>
- Lim, J. S., Choe, M.-J., Zhang, J., & Noh, G.-Y. (2020). The role of wishful identification, emotional engagement, and parasocial relationships in repeated viewing of live-streaming games: A social cognitive theory perspective. *Computers in Human Behavior*, 108, 106327. <https://doi.org/10.1016/j.chb.2020.106327>
- Macleane, M., Harvey, C., Yang, R., & Mueller, F. (2021). Elite philanthropy in the United States and United Kingdom in the new age of inequalities. *International Journal of Management Reviews*, 23(3), 330–352.
- Martin, M. W. (1994). *Virtuous giving: Philanthropy, voluntary service, and caring*. Indiana University Press. <https://books.google.fr/books?id=QIRHAAAAMAAJ>
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform manifold approximation and projection for dimension reduction* (arXiv: 1802.03426). arXiv. <http://arxiv.org/abs/1802.03426>
- Miller, V., & Hogg, E. (2023). ‘If you press this, I’ll pay’: MrBeast, YouTube, and the mobilisation of the audience commodity in the name of charity. *Convergence: The International Journal of Research into New Media Technologies*, 29(4), 997–1014. <https://doi.org/10.1177/13548565231161810>
- MrBeast. (n.d.). MrBeast [YouTube Channel]. YouTube. Retrieved February 2, 2024, from <https://www.youtube.com/@MrBeast/featured>
- Nelson-Field, K., Riebe, E., & Newstead, K. (2013). The emotions that drive viral video. *Australasian Marketing Journal*, 21(4), 205–211. <https://doi.org/10.1016/j.ausmj.2013.07.003>
- O’Brien, C. (2022, January 19). How do social media algorithms work. *Digital Marketing Institute*. <https://digitalmarketinginstitute.com/blog/how-do-social-media-algorithms-work>
- Perelli, A. (2021). YouTube star MrBeast breaks down how he makes eye-catching thumbnails and why he’d pay \$10,000 for the best possible one. *Business Insider*. <https://www.businessinsider.com/mrbeast-youtube-thumbnail-strategy-advice-for-creators-2021-10?r=US&IR=T>
- Pressgrove, G., McKeever, B. W., & Jang, S. M. (2018). What is contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge. *International Journal of Nonprofit and Voluntary Sector Marketing*, 23(1), e1586. <https://doi.org/10.1002/nvsm.1586>
- Sabate, F., Berbegal-Mirabent, J., Cañabate, A., & Lebherz, P. R. (2014). Factors influencing popularity of branded content in Facebook fan pages. *European Management Journal*, 32(6), 1001–1011. <https://doi.org/10.1016/j.emj.2014.05.001>
- Salazar, J. M. R. (2017). Inverted u-shaped impact of social media posting frequency on engagement and sentiment ratio. *Empirical Quests for Management Essences*, 1(1), 1–15.
- Sanzo, M. J., Álvarez, L. I., Rey, M., & García, N. (2015). Business–nonprofit partnerships: Do their effects extend beyond the charitable donor–recipient model? *Nonprofit and Voluntary Sector Quarterly*, 44(2), 379–400. <https://doi.org/10.1177/0899764013517770>
- Schreiner, M., Fischer, T., & Riedl, R. (2021). Impact of content characteristics and emotion on behavioral engagement in social media: Literature review and research agenda. *Electronic Commerce Research*, 21(2), 329–345. <https://doi.org/10.1007/s10660-019-09353-8>
- Soder, C. (2009). Social media become key tool for nonprofits. *Crain's Cleveland Business*, 30(45), 5.

- Sohn, E. (2017). Fundraising: The Ice Bucket Challenge delivers. *Nature*, 550(7676), S113–S114. <https://doi.org/10.1038/550S113a>
- Song, A., Lee, H. I., Ko, M., & Lee, U. (2015, April). Every little helps: Understanding donor behavior in a crowdfunding platform for non-profits. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1103–1108).
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Statista. (2024). [Most popular social networks worldwide as of January 2024, ranked by number of monthly active users] (Statista). <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- Taylor, D. G., Strutton, D., & Thompson, K. (2012). Self-enhancement as a motivation for sharing online advertising. *Journal of Interactive Advertising*, 12(2), 13–28. <https://doi.org/10.1080/15252019.2012.10722193>
- TeamSeas. (n.d.). #TeamSeas [YouTube Channel]. YouTube. Retrieved February 2, 2024, from <https://www.youtube.com/@teamseas>
- TeamTrees. (n.d.). #TeamTrees [YouTube Channel]. YouTube. Retrieved February 2, 2024, from <https://www.youtube.com/@teamtrees/featured#TeamTrees>. (n.d.). FAQ. Retrieved February 2, 2024, from <https://teamtrees.org/>
- Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2020). What drives virality (sharing, spread) of YouTube video ads: Emotion vs brand prominence and information. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3583613>
- Van Der Linden, S. (2017). *The nature of viral altruism and how to make it stick*. <https://doi.org/10.17863/CAM.26206>
- Van Dyk, L., & Fourie, L. (2015). Challenges in donor–NPO relationships in the context of corporate social investment. *Communicatio*, 41(1), 108–130.
- Waters, R. D., & Jones, P. M. (2011). Using video to build an organization’s identity and brand: A content analysis of nonprofit organizations’ YouTube videos. *Journal of Nonprofit & Public Sector Marketing*, 23(3), 248–268. <https://doi.org/10.1080/10495142.2011.594779>
- Wicks, P. (2014). The ALS Ice Bucket Challenge – Can a splash of water reinvigorate a field? *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 15(7–8), 479–480. <https://doi.org/10.3109/21678421.2014.984725>
- YouTube. (2024). Growing your channel. *YouTube Creators*. <https://www.youtube.com/creators/how-things-work/content-creation-strategy/>
- Zhao, X., Chen, B., & Jin, P. (2023). Inspired to donate: How donors’ social class impacts charitable donations. *Journal of Consumer Behaviour*, 22(1), 3–13. <https://doi.org/10.1002/cb.2042>
- Zhou, X., Xu, Y., Li, Y., Josang, A., & Cox, C. (2012). The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review*, 37(2), 119–132. <https://doi.org/10.1007/s10462-011-9222-1>

## **PART III**

# Philanthropy for AI development and regulation



# Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# NAVIGATING RISK COMPLEXITY ASSOCIATED WITH DATA PHILANTHROPY FOR AI

*Rahul Jha*

## 1 Introduction

The intersection of AI and philanthropy represents a dynamic and evolving field of study with significant implications for both sectors. This chapter contributes to this burgeoning discourse by exploring the transformative potential of POs as catalysts for the Ethical and Inclusive AI (EIAI) revolution, as underscored in the book's overarching theme. This exploration is particularly pertinent given POs' unique position at the nexus of industry, government, and academia, allowing them to leverage their influence to shape the trajectory of AI development and implementation. Building upon the foundational work of George et al. (2019a, 2019b), which emphasizes the need to examine variations in DP activities and their impact on donor firms, this chapter explores the nuances of these variations through a case study. These contextual examinations provide insights into how POs can use AI to enhance governance, organizational development, and data-driven decision-making, effectively contributing to the development of impactful AI strategies.

Furthermore, the chapter posits that AI regulatory and risk management frameworks are pivotal in shaping the certainty surrounding DP and its intended impact. This assertion aligns with the handbook's emphasis on how the philanthropic sector can utilize AI for impact assessment strategies. It underscores the need for robust frameworks that govern AI development and ensure its ethical and inclusive deployment.

This chapter addresses the 'data invisible' issue that results from the digital divide and argues for the critical role of DP as a tool to bridge this gap. This argument resonates with the book's objective to provide comprehensive insights on leveraging AI within the philanthropic sector to improve the inclusion and representation of marginalized communities in the digital landscape.

The chapter's abductive reasoning approach, enriched by a case study and grounded in existing literature, aims to make a substantial contribution to the sparse academic knowledge in this field. It endeavors to offer a well-rounded perspective that can guide POs in their journey to embrace AI and actively participate in promoting EIAI practices (see Table 19.1).



Table 19.1 Chapter objectives and research questions

	<i>Objectives</i>	<i>Research questions</i>
1	To define and understand the concept of data philanthropy, provide a brief literature review of this emerging field and identify stakeholders of DP.	How can the concept of DP be integrated within the broader framework of corporate philanthropy, and what are the implications and modalities of its implementation in the context of ethical AI practices?
2	To define and understand the concept of ethical and inclusive AI (EIAI) and responsible innovation and its application in DP.	What are the key drivers and governance frameworks that facilitate or hinder the practice of data philanthropy among organizations?
3	To engage with data ownership in the context of DP.	How do legal challenges and intellectual property rights, such as licenses, impact data access, reuse, and resharing for philanthropic purposes, particularly in fostering ethical and inclusive AI systems?
4	To explore a case study on IO's role in DP illustrating the practical applications and challenges of DP and open data.	<i>Case Study:</i> How can an IO optimize its role as an orchestrator in DP to enhance access to firm data while balancing the interests of diverse stakeholders and maintaining data quality?

## 2 Conceptual foundations of DP

This section defines DP, traces its emergence and evolution, and distinguishes it from related concepts. By establishing a clear conceptual framework, it contributes to the chapter's objective by grounding the discussion in a solid academic foundation and clarifying the scope of DP for POs and other stakeholders. DP is a relatively new and evolving concept within data science, corporate social responsibility, and philanthropy. It refers to private sector companies sharing their data for the public good, often by partnering with nonprofit organizations, research institutions, or government agencies. The data shared can help address social issues, improve public services, and contribute to significant advancements in fields as diverse as health, education, and disaster response. The Covid-19 crisis saw an explosion of such collaborations to address public health challenges.

### 2.1 Defining and understanding DP

The existing literature on DP does not converge on its definition. DP encompasses several interpretations as mentioned in the UNDP working paper (*Data Philanthropy, International Organizations and Development Policy: Ethical Issues to Consider*, 2020) which include: (1) The act of private entities donating their commercially sensitive data to support charitable causes (Wu, 2015), (2) Collaborative endeavors where businesses contribute their data to advance the public interest (United Nations, 2018), (3) The practice of corporations distributing their exclusive datasets to aid societal welfare initiatives (*Big Data Philanthropy for Humanitarian Response*, 2012), (4) Definition through the combination of three elements: (a) unpaid for sharing of or access to (b) privately held data or proprietary data insights for (c) the greater good (Lev Aretz, 2019).

George et al. (2019a, 2019b) refer to DP according to the UN Global Pulse definition: DP is defined as firm activities that include one of the following means of sharing data: (1) Distributing combined and processed datasets for scrutiny under confidentiality agreements; (2) Permitting scholars to examine data within the company's internal network; (3) Participating in live data repositories that consolidate data from various companies in the same sector to maintain competitive

edge; (4) Extracting data within the company's security barriers and sharing key metrics; and/or (5) Accumulating and offering data science knowledge and capabilities.

It is unclear from the literature review whether IOs are considered DP actors (as opposed to recipients of DP). Furthermore, individuals, governments, and IOs all act in harmony in the DP ecosystem. Still, they have either not been acknowledged as actors in DP or another term exists to describe their actions in this space of data exploitation for the greater good. This chapter aims to put a spotlight on IOs in DP ecosystem.

## **2.2 Emergence and evolution of DP**

DP has emerged as a response to the increasing recognition of the value of big data in solving complex societal challenges. Initial discussions in the literature highlight its potential to improve the effectiveness of humanitarian efforts and public policy development.

Some of the literature focuses on ethical considerations, particularly with respect to privacy and data protection (Taddeo, 2016). Scholars and practitioners stress the importance of establishing frameworks and guidelines to ensure that data is shared and used responsibly and that individuals' privacy is protected.

DP has two modalities: First, *Private Data Donation*, which entails granting specific entities access to designated datasets within a restricted environment for particular objectives; and second, *Open Data Donation*, which involves disseminating designated datasets to the public under open licenses, either through public repositories or via open APIs for general or specific purposes. Of the two, private data donation is more prevalent in practice.

## **2.3 Integration with corporate philanthropy**

Corporate philanthropy involves the donation of resources or funds by companies to charitable causes (Peterson et al., 2021). Godfrey (2005) highlights two essential qualifiers of corporate philanthropy: first, the voluntary and second, the non-reciprocal giving of corporate resources to benefit the community welfare. Corporate philanthropy can be further classified into corporate giving, corporate volunteering, and corporate foundations (Gautier & Pache, 2015). These practices reflect the various ways in which corporations can engage in philanthropy, each with its own strategies and impacts on societal welfare.

DP uniquely cuts across all three categories of practicing corporate philanthropy: corporate giving, corporate volunteering, and corporate foundations. It complements corporate giving by adding a data-driven dimension to financial contributions, enhancing the impact of charitable initiatives. In corporate volunteering, employees with data science skills can offer their expertise to social causes, aligning individual efforts with the company's broader DP goals. Lastly, corporate foundations can integrate DP into their long-term projects, creating more comprehensive and impactful social good programs. This multifaceted approach allows DP to amplify traditional corporate social responsibility efforts. It is worth noting that Lev Aretz (2019) mentions that for data philanthropy to be considered as such, these philanthropic efforts must be outside of the core business model of the firms.

## **2.4 IOs as a stakeholder in DP**

IOs have access to private sector data through their membership, as government counterparts in relevant ministries and regulatory bodies can help mediate IO's access to private sector data. Such access, however, depends on individual circumstances. Table 19.2 summarizes how different cases of DP can be facilitated by IOs.

Table 19.2 DP – outline of key DP scenarios, detailing stakeholder roles and how IOs facilitate these efforts (George et al., 2019a)

<i>Case of DP</i>	<i>Stakeholders and their role</i>	<i>Potential role of IOs</i>
Distributing combined and processed datasets under confidentiality agreements	Firms: Provide datasets; Researchers: Analyze data under confidentiality	Facilitate agreements; Ensure data privacy and security; Provide a platform for data sharing
Permitting scholars to examine data within the company’s internal network	Companies: Grant access to internal data networks; Scholars: Conduct in-depth data analysis	Mediate access permissions; Establish ethical standards for data use; Support collaborative research initiatives
Participating in live data repositories consolidating data from various companies	Companies: Share and pool data; Repository administrators: Manage and maintain the data repository	Coordinate and oversee data repositories; Standardize data formats and protocols; Promote data sharing among companies
Extracting data within company’s security barriers and sharing key metrics	Companies: Process and share essential data indicators; Data users: Utilize shared data for analysis and decision-making	Guide data processing and sharing protocols; Ensure data integrity and relevance; Facilitate partnerships between companies and data users
Accumulating and offering data science knowledge and capabilities	Data science experts: Provide expertise and training; Beneficiary organizations: Apply learned skills to data-driven projects	Organize training and capacity-building workshops; Foster networks of data science experts and organizations; Support the application of data science in public and NGO sectors

### 3 Ethical and Inclusive AI (EIAI)

This literature review explores the emerging field of DP, with a specific focus on its potential applications in developing Ethical and Inclusive AI (EIAI) systems. We will examine the challenges POs face in adopting AI technologies and the role of IOs in facilitating DP initiatives. Key concepts such as corporate philanthropy, data as a strategic resource, and the intricacies of data ownership will be discussed. The review also distinguishes between DP, open data, and other related concepts within computational social science. This sets the stage for an analysis of the regulatory requirements for DP and the need for responsible innovation frameworks to ensure EIAI development.

For this chapter, AI is a reference to the AI system as a machine-engineered setup capable of generating outputs such as forecasts, suggestions, or choices that have an impact on real or virtual settings. These AI systems are engineered to function with varying degrees of independence (*AI Risk Management Framework | NIST, 2021*). As of November 2023, the OECD has updated the definition of an AI system to align it with the EU AI Act, and the new definition now reads as

An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in autonomy and adaptiveness after deployment.

IOs are well positioned to bring together the various frameworks emerging organically from the deployment of AI and its concerns in both the Global South and the developed nations. On November 25, 2021, UNESCO came up with an ethical framework aimed to harnessing the positive potential of AI while mitigating its risks (*UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence*, 2023). It serves as the first universal set of ethical norms and assigns the onus of its implementation to individual states. UNESCO oversees its application and requires periodic progress reports from its member states.

*Core Principles of the first draft of the recommendation on the ethics of AI (First Draft of the Recommendation on the Ethics of Artificial Intelligence, 2020)*

The framework focuses on ensuring that AI technologies are developed and deployed in a manner that is aligned with human rights and contributes to the achievement of the Sustainable Development Goals. It addresses key ethical concerns such as transparency, accountability, and privacy, and provides actionable policy directives in areas such as data governance, education, culture, labor, healthcare, and the economy. From these overarching principles, the recommendation delves into specific areas of concern, such as data protection, prohibitions on harmful uses of AI, and the imperative of inclusiveness, each of which addresses key facets of ethical AI implementation.

### **3.1 Data protection**

The guidelines advocate enhanced data protection measures that go beyond current practices by tech companies and governments. They emphasize the need for transparency and individual control over personal data, including the right to access and erase one's own data records.

### **3.2 Prohibitions**

The framework explicitly prohibits the use of AI for social scoring and mass surveillance, citing their invasive nature and infringement on human rights. It insists that ultimate accountability rests with humans, not AI systems.

### **3.3 Inclusiveness in the AI ethics**

Clause 19 of the framework highlights inclusiveness and diversity, emphasizing that these principles should be upheld throughout the entire life cycle of AI systems. This should align with international human rights laws, norms, and values, and reflect the diversity in culture, gender, and social background. The clause advocates active inclusion of individuals from various groups based on factors such as sex, gender, language, religion, political views, nationality, ethnicity, sexual orientation, and other identifying characteristics such as disability and age. It also warns against tendencies toward uniformity, suggesting that such inclinations should be closely monitored and rectified. An excellent example of a dataset generated by people with disabilities to support accessibility is IncluSet, which requires only metadata to point to any information in the dataset (Kacorri et al., 2020).

### **3.4 Deployment of AI systems for POs**

AI continues to be integrated into the information technology architecture of organizations across diverse industries. However, POs have been left behind in the adoption race as corporations play a significant role in the research, development, and deployment of AI systems (Cihon et al., 2021).

It has been highlighted that POs, as discussed by Voida (2014), require a broader approach beyond just information technology design to adequately support the wide range of philanthropic activities and the diverse stakeholders involved in this sector. This approach should address the multifaceted needs of philanthropy and its stakeholders, highlighting the necessity for POs to adopt AI in a manner that supports their diverse activities and the broad spectrum of participants involved. As Juni (2023) highlights the lack of AI integration among Swiss POs, key areas where AI is underutilized but seen as potentially beneficial include donor matching, reporting, and impact evaluation. Few interesting case studies have been explored where AI is leveraged by POs by Adrienne (2023). Here, the principles of DP can be leveraged by the POs to gain access to the valuable dataset in the sustainable development goal domains in which they operate. DP can support POs by bridging the digital divide through the adoption of digital tools, developing ethical frameworks for AI usage, enhancing their digital capabilities through capacity building, and fostering collaborative research with institutions to promote ethical AI usage.

### 3.5 Key terminologies in DP

A distinction must be made between peripheral concepts around data management. *Data collaborative* is another term that is used interchangeably with DP, as defined by Susha and Gil-Garcia (2019). Lev Aretz (2019) points out how “collaborative” is a misnomer – for there is barely any collaboration involved among stakeholders. *Data donation* refers to the active consent of an individual to donate their personal data for research purposes (*Data Donation Lab*, 2023). *Digital philanthropy* is the concept of utilizing modern digital tools and platforms by individual users to enhance charitable activities and is further classified into three digital philanthropy products: (1) event-based products, (2) issue-based products, and (3) gamified mini-programs (Song et al., 2022). Open data refers to making data freely available to the public without restricting access, usage, or redistribution. It involves releasing data, usually by public organizations or governments, which can be used by individuals, businesses, researchers, and other stakeholders for various purposes. *Open Data* (OD) has been further supported by Open Government Data (OGD) and has gained traction across the world’s democracies. It promotes transparency and accountability in governance by allowing citizens to access and scrutinize information about public services, policies, and decision-making processes (Janssen et al., 2012). As mentioned by Heimstädt et al. (2014), it is important to acknowledge that OGD is a subset of OD, as the data could be sourced from outside of government, such as corporations and NGOs. The “open” in OD describes the technical and legal attributes that allow the sharing and reusing of the dataset across the digital economy. The legal openness of the data, however, needs to be complemented by the technical openness of the data (file formats), as described in Berners-Lee’s model of Linked Open Data (LOD) (Bizer et al., 2009) to transform “data on the web” into “the web of data” by linking different datasets scattered across the internet and hosted in other websites, repositories, and archives.

## 4 Discussion

The discussion explores the dynamics between AI and its ethical, inclusive application. It acknowledges AI’s transformative potential while confronting the challenges and risks associated with its unchecked deployment. This dichotomy sets the stage for a deeper examination of specific areas within AI ethics, such as the need for ethical and inclusive AI, the pressing call for AI regulation, and the pivotal role of responsible innovation in DP.

#### **4.1 Data as a strategic asset**

The Resource-based View of the firm (RBV) focuses on four key parameters that characterize strategic resources, which are tied to firm performance and organizational efficacy (Crook et al., 2008). These four parameters are (1) Valuable; (2) Rare; (3) Inimitable; and (4) Non-substitutable; this theory is better known by the acronym VRIN (George et al., 2019a, 2019b). In their paper, they establish data as a strategic VRIN resource of the firm. The disparity in digital literacy and connectivity underscores the inherent bias in data, emphasizing the need to better understand data as a strategic resource that is not neutral but shaped by various external factors. Therefore, bias in data is unavoidable as certain users have a larger digital footprint than others, based on the extremely high variance of digital literacy and connectivity across economies and regions as reported by the International Telecommunication Union, where the number of people in the world not connected to the internet is a whopping 2.3 billion (*Press release, 2023*).

#### **4.2 The need for Ethical and Inclusive AI (EIAI)**

While there are many guidelines in place to make traditional software safer, AI systems bring their own unique risks. For instance, AI systems use data that can change significantly over time, making them act unpredictable. These systems are often complicated and work in complex settings, making them difficult to fix when things go wrong. They are also influenced by social factors and human behavior. The risks and benefits of AI come from its technical features, how people use it, who uses it, and the social setting in which it is used. This makes AI particularly challenging for organizations and society to handle safely. If not managed well, AI can exacerbate unjust or harmful situations. If managed correctly, AI can help make things fairer and safer. Currently, AI systems lag in these two areas: (1) *Explainability/Interpretability* and (2) *Transparency*, although progress is being made on all these fronts, albeit slowly.

##### *4.2.1 Explainability*

Often referred to as interpretability in AI, this refers to the study of how to understand the decisions of AI systems and how to design systems whose decisions are easily understood or interpretable (Rudner & Toner, 2021). Even the most accomplished designers and scientists of Large Language Models cannot explain the enhanced capabilities of AI.

##### *4.2.2 Transparency*

Stanford had designed a metric called the foundation model transparency index. The highest score achieved was 54% by Llama 2 from Meta (Xiong & Zhang, 2023). This highlights the need for the AI community to unite to enhance transparency – for policy frameworks can only proceed with transparency of the foundation models. Despite the growing number of AI initiatives, three key challenges persist: (1) Underrepresentation of developing nations in global AI dialogues, which hinders their economic and social development. UNESCO’s work on AI ethics exemplifies how IOs can foster inclusivity. (2) Fragmentation in AI initiatives, with over 200 AI ethics and governance frameworks (Corrêa et al., 2023) but no unified platform, limits accessibility for countries and stakeholders outside existing alliances. (3) More public sector capacity and expertise in AI is needed, particularly in understanding its role in advancing Sustainable Development Goals (SDGs). While the AI for Good Global Summit, organized by an IO, aims to bridge this gap, a more systematic approach is needed to fully harness AI’s potential benefits and mitigate its risks (Guterres, 2020).

### 4.3 Collaboration governance framework and DP

Collaboration Governance is defined as the processes and structures of public policy decision-making and management that engage people constructively across the boundaries of public agencies, levels of government, and/or the public, private, and civic spheres to reach a goal that could not otherwise be accomplished (Emerson et al., 2011). DP involves collaboration with diverse stakeholders in the data ecosystem. Academic literature on collaboration includes a framework proposed as a Collaboration Governance Framework (CGR): The CGR framework suggests that collaboration emerges when at least one of the following catalysts exists: (1) guidance from leaders, (2) significant motivators (such as challenge, emergency, risk, or chance), (3) mutual reliance (where organizations cannot achieve an objective independently), and (4) unpredictability (such as in addressing a social issue). These four catalysts are pertinent to DP (Emerson et al., 2011; see Figure 19.1).

Applying this framework to the data collaborative context, Susha and Gil-Garcia (2019) identified resources (primarily financial) and incentives as two additional drivers of data collaborative. Building on their work and the author’s attempt to apply this framework to DP-related collaboration governance, multidimensional cost reduction is proposed as an additional driver for achieving collaboration DP.

#### 4.3.1 Reducing the cost of liability risks

Researchers have found that the decreasing transparency of foundation models is based on the legal underpinnings of current cyber law arrangements. Companies are liable for using data they did not have explicit permission for (*Art. 89 GDPR – Safeguards and Derogations Relating to Processing for Archiving Purposes in the Public Interest, Scientific or Historical Research Purposes or Statistical Purposes – General Data Protection Regulation (GDPR)*, 2016). Therefore, firms choose not to share the details of the data used as a training dataset in order to minimize their risk of liability due to the lack of explicit purpose associated with collecting training dataset. Lazer et al. (2020) noted that the sharing of data and open data may have reduced in the wake of laws such as GDPR to protect privacy. This lack of transparency about the dataset on which the model was trained could also imply that companies overestimate the risks of sharing the data, while underestimating the benefits of addressing the wicked problems (Verhulst, 2023).

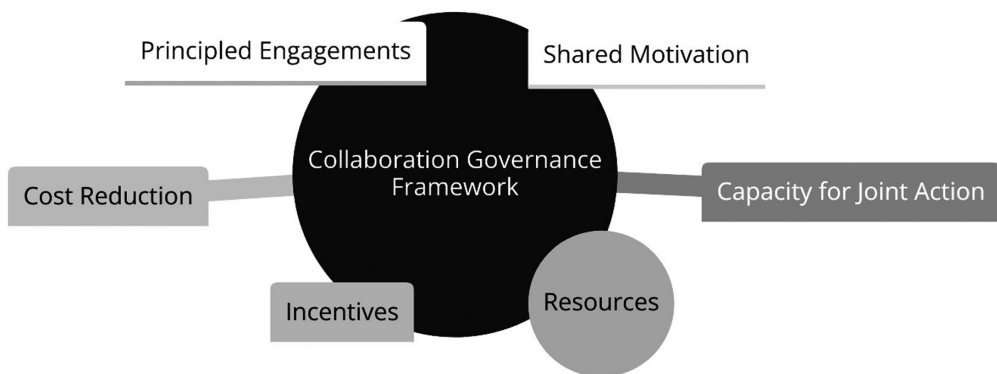


Figure 19.1 Collaboration governance framework.

Source: Adapted by author from Susha and Gil-Garcia.

Further elaborating on this context, Susha and Gil-Garcia (2019) added financial resources and incentives as additional drivers of data collaboration. Expanding on this and integrating the author's suggestions for applying the CGR framework in diverse governance contexts, multidimensional cost reductions of (1) legal and (2) compliance are proposed as key drivers of DP. Researchers have pointed out that the opacity of foundation models based on current cyber law frameworks makes companies reluctant to share data. To avoid liability from using data without explicit consent, firms tend to be reticent to disclose the datasets used to train AI models. Lazer et al. (2020) observed that data sharing and openness have decreased post-GDPR, which was designed to protect privacy. This reticence not only raises questions about transparency but also indicates a potential overestimation of risks and underestimation of the benefits in tackling complex societal problems, as suggested by Verhulst (2023).

#### *4.3.2 Reducing the cost of compliance to adhere to AI principles*

When using DP to develop AI systems, it is crucial to align the initiative with the core ethical principles outlined in the AI ethics framework. Table 19.3 summarizes exhaustive dimensions of AI ethics principles to be adhered to by firms engaging in DP, according to UNESCO's Uniform Ethical Framework.

This may increase the compliance and coordination overhead for firms engaging in DP, and efforts must be made to ensure that the coordination cost of compliance does not exceed the overall perceived benefit and/or impact to the POs. As indicated in findings by Morley et al. (2023), incurring additional costs, slowing down innovation, and draining resources were identified as the top disadvantages of developing AI ethically ( $n = 54$ ).

Based on Table 19.3, adhering to AI ethics principles in DP activities undoubtedly introduces additional layers of complexity, which can lead to increased costs and heightened compliance requirements. The need to provide data that supports harm prevention and addresses social and environmental issues requires rigorous data selection and validation processes. Ensuring that datasets are ethically sourced and relevant requires thorough vetting procedures and potentially the development of new sourcing strategies. The enhancement of safety and security features, alongside the commitment to diversity and non-discrimination, calls for sophisticated data analysis tools and expertise to identify and mitigate biases, which may not be readily available without significant investment.

Furthermore, the obligation to respect privacy and adhere to data protection standards necessitates the implementation of robust data management and security protocols. This often involves advanced technologies and specialized personnel to anonymize data and maintain privacy by design. The clear oversight and accountability measures required for ethical data use in AI demand meticulous documentation and reporting systems, which can be resource-intensive to establish and maintain.

Advocating for transparency and demanding explanations of AI outcomes necessitate the creation of communication channels and explanatory frameworks that are accessible to a non-technical audience, adding to operational costs. Engaging a wide range of stakeholders to ensure diverse perspectives in AI development not only increases administrative overhead but also requires the facilitation of stakeholder engagement and consensus-building activities.

In summary, while the integration of AI ethics principles into DP is essential for fostering trust and ensuring responsible AI development, it comes with a significant increase in both direct and indirect costs associated with compliance, personnel training, system upgrades, and ongoing management of DP initiatives.



Table 19.3 UNESCO ethics of AI principles (*Recommendation on the Ethics of Artificial Intelligence*, 2022) mapping to EIAI aspects

	<i>Principle</i>	<i>Summary</i>	<i>Relevance to DP for EIAI</i>
1	Proportionality and do no harm	AI should aim for legitimate aims without harm.	Provides data to prevent harm and address social issues.
2	Justification of AI use	AI must be appropriate, respect rights, and be rigorous.	Ensures ethical data sourcing for beneficial AI.
3	Safety and security	AI systems should address safety and security risks.	Enhances AI safety with secure data sharing.
4	Fairness and non-discrimination	AI should prevent discrimination and promote justice.	Offers diverse data to reduce AI bias.
5	Minimizing discrimination	Efforts to avoid AI biases and discrimination are vital.	Contributes representative datasets for fairness.
6	Sustainability	AI's impact on sustainability should be assessed.	Prevents duplication in data acquisition efforts to save energy.
7	Right to privacy and data protection	Privacy must be protected throughout AI's life cycle.	Ensures data sharing, respects privacy standards.
8	Data protection frameworks	Establish data protection governance.	Participates in frameworks for responsible sharing.
9	Privacy by design	AI must protect personal information.	Anonymizing data before sharing.
10	Human oversight and determination	Persons/entities should be responsible for AI systems.	Conducts data sharing with clear accountability.
11	Human decision in AI reliance	Humans should control AI systems.	Supports AI that enhances human decision-making.
12	Transparency and explainability	AI should be transparent and explainable.	Advocates for clarity on donated data usage in AI.
13	Balancing transparency	Balance transparency with privacy and security.	Navigates data sharing between transparency and privacy.
14	Transparency for trust	Transparency contributes to trust in AI.	Fosters trust by being clear about data usage.
15	Explainability of AI systems	AI should be understandable and insightful.	Demands explanations on data impact on AI.
16	Responsibility and accountability	AI actors should assume responsibility for AI impact.	Holds AI actors accountable for ethical data use.
17	Accountability mechanisms	Develop AI oversight and audit mechanisms.	Includes DP in accountability tracking.
18	Awareness and literacy	Promote understanding of AI.	Plays a role in AI education and engagement.
19	Learning about AI impact	Educate on AI's societal impact.	Supports research on AI's effects on society.
20	Multi-stakeholder governance	Respect laws in data use and regulation.	Engages with stakeholders for respectful data use.
21	Stakeholder participation	Diverse participation for inclusive AI governance.	Involves various stakeholders for diverse AI input.

#### **4.4 Legal challenges in accessing, reuse, and resharing of data**

In the context of DP, “ex post” refers to the practice of using data that has already been collected, often without prior explicit consent for specific uses, to train AI systems. This approach can expedite the availability of data but raises ethical concerns regarding consent and data privacy. Conversely, “ex ante” involves obtaining explicit consent before collecting data, ensuring that all data usage complies with ethical standards and legal requirements from the outset. While this method enhances transparency and trust, it can be more costly and time-consuming. Today’s AI systems are trained on historical data, which is collected and utilized retrospectively, or ex post. This causes ambiguity about what historical data can be used and by whom, which raises an interesting set of copyright and data ownership challenges in the AI era. This chapter endorses Lev Aretz’s (2019) call for a legal academic discussion of DP, although it must be complemented by technical discussions as well, particularly around non-legal nature of an important metadata governing website data crawlability called Robot.txt (Pierce, 2024). Ex ante data collection is also facilitated by augmenting user engagement with privacy policies, and new digital tools for contextual privacy policies (CPPs) that embed relevant privacy information directly into their corresponding contexts must be explored to overcome users avoiding engaging with privacy policies (Windl et al., 2022).

Data is recognized as an asset with potential benefits for various stakeholders. It can drive scientific collaboration, enhance market efficiency, and increase transparency in governance and corporate sectors (Gallagher et al., 2020). However, the sharing and use of data are often complicated by legal considerations. The more restrictive the licenses, the lesser the uses and reuses that are possible, and the more likely that an incompatibility with other licenses may arise in the life cycle of the data (Malone, 2023). Open licensing is one of the most suitable options for philanthropic organizations (Gray, 2009). In the context of promoting data philanthropy, the standardization of data licenses has been proposed as a crucial step toward facilitating the reuse of data for philanthropic purposes (Benjamin et al., 2019). Additionally, the licensing of FAIR (Findable, Accessible, Interoperable, and Reusable) data has been highlighted to promote the reuse of data for philanthropic endeavors (Labastida & Margoni, 2020). There is an increasing degree of awareness among POs about open licensing options such as Creative Commons or General Public License (GPL). Creative Commons (CC) licenses and sui generis database rights are important legal tools in data management. CC licenses manage copyright restrictions and can be applied to databases, while sui generis database rights protect the investment made in creating a database. POs must secure the necessary rights and identify the elements covered by CC licenses when making databases publicly available. Text and data mining activities on CC-licensed databases are permissible, but compliance with license terms is essential. Understanding these aspects of data management enables organizations to navigate the complexities and unlock the full potential of data for societal benefit. It’s important to flag here that software codes are not suitable artifacts where CC can be applied, and CC recommends referring to open-source licensing model (*Copyright & Licences – Research Data – UNIGE*, 2017). As can be inferred from this, the legal landscape is still working on the heterogeneity of different licensing provisions governing different elements of AI systems. An example of a license selector to help navigate licensing applicability is *License Selector* (2015).

Creative Commons (CC) licenses serve as a standardized legal framework that facilitates the public utilization of copyrighted works by delineating explicit permissions and restrictions (*About*

CC Licenses, 2020). The suite consists of six primary licenses, each of which varies in the degree of permissiveness it offers. These range from the most permissive, CC BY, which allows both commercial and non-commercial adaptations as long as the creator is credited, to the most restrictive, CC BY-NC-ND, which permits only non-commercial distribution of the work in its original form. Additionally, the Creative Commons Public Domain Dedication (CC0) enables creators to voluntarily relinquish their copyright, thereby contributing their works to the public domain. The application of these licenses is irrevocable and necessitates prior ownership or explicit permission to license the work. By leveraging the CC licenses, creators and institutions can contribute to an increasingly open and collaborative digital ecosystem (see Table 19.4).

DP ecosystem has many actors, and relying on legal job function exclusively to classify the applicability of specific licenses could be a bottleneck. Therefore, Creative Commons have started their certification program through which it is reported that the successful participants are well

Table 19.4 Spectrum of copyright levels through Creative Commons suite of licenses (*About CC Licenses, 2020*)

<i>License type</i>	<i>Permissiveness level (1–6)</i>	<i>Permissions granted</i>	<i>Restrictions</i>	<i>Commercial use</i>	<i>Adaptation of work</i>
CC BY	Most permissive (6)	Distribution, adaptation	Must attribute to the creator	Yes	Yes
CC BY-SA	5	Distribution, adaptation	Must attribute and share alike	Yes	Yes
CC BY-ND	4	Distribution	Must attribute; no adaptations	Yes	No
CC BY-NC	3	Distribution, adaptation	Must attribute; non-commercial	No	Yes
CC BY-NC-SA	2	Distribution, adaptation	Must attribute; non-commercial and share alike	No	Yes
CC BY-NC-ND	Most restrictive (1)	Distribution	Must attribute; non-commercial; no adaptations	No	No
CC0 (Public Domain)	Special case	Distribution, adaptation	No restrictions; voluntary copyright relinquishment	Yes	Yes

equipped to advise their institutions on open licensing best practices (Training, 2022). Another countermeasure against copyright issues arising in generative AI space is the support by the firms to pay for the legal fees for any copyright infringement (Novak, 2023). Under this copyright shield mechanism, firms agree to pay some of their users the cost incurred from legal claims around copyright infringement.

## **5 Case study – IOs as an orchestrator of DP activities**

### **5.1 Background**

This international organization (IO) is the United Nations specialized agency for information and communication technologies – ICTs. Its mandate ranges from standardizing ICT technologies to facilitating an enabling environment for improving connectivity infrastructure, as well as allocating radio spectrum and satellite orbits. It also collects and disseminates a vast array of data through its data portal called DataHub, which is instrumental in global ICT development (About, 2023). The IO’s transition from the World Telecommunication/ICT Indicators Database (WTID) to the Data Hub is an illustrative case of adapting to the demands of the digital age by making ICT data more accessible and user-friendly. The Data is licensed to the public through a Creative Commons Attribution-Non-Commercial-ShareAlike 3.0 IGO license. This strategic shift not only reflects an organizational commitment to improving data dissemination but also signifies a response to the evolving needs of data users worldwide. By making datasets available free of charge for non-commercial purposes, the IO is expanding the reach of valuable ICT data and potentially fostering greater innovation and research in the field.

### **5.2 The problem**

Despite the wealth of data available, there are challenges to using this information for social good. These include obtaining data from the firms that may be reluctant to share it, ensuring data quality, aligning the interests of different stakeholders, and uptake of the data for analysis and exploitation. In some cases, the member states may not have a dedicated national statistics office to follow up on the survey questionnaires. The key questions that this case study aims to answer are listed in Table 19.5.

### **5.3 Analysis**

IO acts as a central hub for data collection, validation, and dissemination. It engages with national governments and private entities to gather a comprehensive set of ICT data. This data is critical to understanding and improving access to and use of ICTs by households and individuals. IO’s role

*Table 19.5* Case study questions

---

<i>Key questions</i>	
1	How can IO effectively orchestrate DP activities?
2	What mechanisms can ensure data quality and availability?
3	How can IO align the interests of governments and private sector companies, particularly in telecommunications?

---

in standardizing data collection and ensuring international comparability is vital for the credibility and utility of the data. The digital divide between economies is evident in the ability of governments to respond to the questionnaires and survey forms that are floated among the member states.

The member states that are lagging in the digital transformation of their ICT ecosystem do not have the necessary reporting mechanism, or the presence of the relevant coordinating agency to provide the ICT-related data. For those economies that do not reply to the questionnaire, price data are collected directly from telecommunication operators' websites and/or through direct correspondence with the operator, and the data is cross-checked by the IO. For example, the source of data on retail prices is the advertised prices of selected services for residential customers, effective at the time of data collection, from operators with the largest market share in an economy, measured by the number of subscriptions (*ITU Data Collection*, 2013).

In addition, the IO also supports the capacity development of member states to facilitate data acquisition through the provisioning of a manual. For example, a manual has been prepared to help countries produce high-quality and internationally comparable data on ICT access and use by households and individuals. The manual focuses on the partnership's core list of indicators on measuring ICT access and use by households and individuals. It can be used as basic reference material when preparing, designing, and implementing ICT household surveys, and serves as the basis for IO's training course on ICT household statistics.

The data collected through diverse stakeholders and thorough analysis is then subject to copyright laws that do not permit commercial use. The data is licensed to the public through a Creative Commons Attribution-Non-Commercial-ShareAlike 3.0 IGO license. Under the terms of this license, one may copy, redistribute, and adapt the data for non-commercial purposes, provided the work is appropriately cited. The lack of commercial licensing terms may redirect the users of such data to other organizations, some of which offer their dataset under different licensing arrangements. For example, the World Bank offers its data under Creative Commons Attribution 4.0 (CC BY 4.0), which includes commercial use (*Data Access and Licensing*, 2023).

In transitioning from the WTID to the Data Hub, the IO has embraced a progressive approach toward data sharing, reflecting a keen understanding of the digital landscape's evolving demands and playing its part as an orchestrator in the data philanthropy landscape. However, this transition also presents areas for improvement, particularly in ensuring data quality, updating frequency, and supporting commercial use. By critically evaluating these aspects, the IO can refine its strategies, ensuring the Data Hub serves not only as a repository of ICT data but also as a catalyst for innovation and policymaking.

#### **5.4 Summary**

IO's structured approach to data collection and its collaboration with various stakeholders position it as a key orchestrator in DP. Its efforts contribute significantly to informed policymaking and the advancement of global ICT access. However, the restriction of commercial use may hinder its at-scale adoption by the industry body. Some recommendations based on thematic focus of this chapter are listed in Table 19.6.

#### **5.5 Limitations**

While the chapter outlines potential strategies for POs to navigate regulatory landscapes and promote ethical AI, it is limited by assumptions on the universal availability of international organizations (IOs) and standardized frameworks, the challenge of adapting to swiftly changing regulatory

Table 19.6 Recommendations based on applying DP principles to the case study

<i>Recommendations</i>	
1	IO could continue strengthening partnerships with private sector companies, leveraging their data for the public good while respecting commercial confidentiality and individual privacy.
2	It could enhance its data validation processes to ensure the highest data quality.
3	IO should advocate for and support the development of clear guidelines and frameworks for DP to ensure ethical and effective use of data.
4	The manual developed for the economies to be able to report back to IO's survey questionnaire can be complemented with capacity-building workshops and training aimed at the uptake of the economy's ability to respond with relevant data.
5	The royalty associated with the commercial use of the data by for-profit firms can be integrated in the platform itself through tiered API access. Alternatively, the royalty model can be replaced by a licensing option that allows commercial use by transitioning to CC BY 4.0.

environments, and the potential overlook of operational hurdles in data privacy and ownership. Furthermore, the reliance on case studies, while enriching, may not capture the broad spectrum of DP practices. Additionally, the proposed solutions may not fully consider the diverse technological capabilities across POs or the complexities of multi-stakeholder collaboration and their implications for data integrity and use.

Lastly, data privacy and ownership concerns: the chapter acknowledges the complexity of data ownership and privacy but may not fully address the operational challenges POs face in managing these concerns, especially considering varying international data protection regulations. These limitations suggest a need for ongoing adaptation and broader empirical validation to ensure the relevance and effectiveness of the outlined approaches in the dynamic landscape of DP and AI.

### **5.6 Recommendations and future research**

Organizations aiming to leverage AI should be aware of coordination and compliance overheads before engaging in DP activities. Adhering to AI principles may increase compliance costs, which must be budgeted for if DP activities include building an AI system.

Academic literature tends to focus on private sector firms as practitioners of DP. Using the case study of an IO, it is evident that DP is a large part of IO's operations and, as such, should be included in the reference to the main stakeholders engaged in the DP.

Understanding the impact of generative AI – and making policy decisions around it – requires new interdisciplinary scientific inquiry into culture, economics, algorithms, and the interaction of technology and creativity (Epstein et al., 2023). This implies DP, where the provision of diverse stakeholders in addition to data scientists must take place if the development of an AI system is a goal.

While academic literature focuses primarily on private sector firms as practitioners of DP, it is essential to explore the role of the IOs in DP ecosystem. The impact of risk management frameworks on firms varies depending on their capability to comply with evolving privacy regulations. Further investigation could be done based on the work of Ghosh (2019) to explore the organizational design options for managing the challenges associated with developing AI systems within the PO context.

The impact of risk management framework on firms will vary depending on the ability of the firm to comply with the evolving privacy regulations. However, the paradox of privacy and

transparency means that a balance must be struck by optimizing the privacy needs with the transparency needs of the AI system. There is an ongoing call in this space for independent third-party auditors who can perform the audits for the developed AI systems against the above benchmarks (*Radical Proposal: Third-Party Auditor Access for AI Accountability*, 2021). An enabling environment for such auditors to thrive in an ecosystem will drive down costs in the long run. This could be excellent phenomenon-based research to look out for in the research community.

Lastly, POs should also look out for an emerging and evolving concept of Digital Public Infrastructure (DPI). DPI is a digital network that allows governments to deliver economic opportunities and social services to all of its residents with efficacy (Hong, 2023). The concept of DPI encompasses (1) Networked Open Technology Standards, (2) Enabling Governance, which includes frameworks and policies that support the effective deployment and management of DPI, and (3) Innovative Community and Market Players – DPI thrives on the contributions of a diverse community that includes governments, POs, private sector participants, and civil society (*Digital Public Infrastructure*, 2024). This community drives innovation, particularly in public programs, through a competitive and collaborative approach. POs such as Bill and Melinda Gates Foundation have made significant investments in promoting DPI as a tool to end poverty (*Statement from Gates Foundation CEO Mark Suzman: Why We Need Digital Infrastructure*, 2022). The next wave of AI-facilitated innovation that the POs will drive will, in turn, shape the future of DPI deployment and scaling across the world.

## 6 Conclusion

In conclusion, this exploration into the realm of DP and AI has unveiled the dynamics at play between technological innovation, ethical considerations, and the transformative potential of POs. Through a comprehensive analysis grounded in the concepts of ethical and inclusive AI, responsible innovation, and the pivotal role of IOs in orchestrating DP efforts, this chapter has illuminated the path forward for POs seeking to leverage AI for social good. The discussions underscore the importance of navigating the complex regulatory landscape, highlighting the need to adhere to ethical principles and the fostering of collaborative governance frameworks to ensure the responsible deployment of AI technologies.

The case study on the IO's role as a DP orchestrator further highlights the critical function of such entities in enhancing global ICT access and facilitating the ethical use of data for AI development. This underscores the broader implications of DP for societal advancement, beyond the confines of traditional philanthropic activities. The recommendations provided serve as a blueprint not only for POs but also for policymakers, private sector firms, and IOs, advocating for a concerted effort to create a more inclusive, ethical, and innovative AI-driven future.

The evolving landscape of DP and AI requires focused research on how private data donation can propel open innovation within firms. This exploratory research should examine the intersection of DP and open innovation, with a particular focus on cases where sharing private datasets underpins innovation and business operations. Exploring this synergy offers a promising avenue for understanding how DP can expand beyond its philanthropic roots to foster an ecosystem of shared innovation. The findings could offer actionable strategies for companies looking to leverage their data assets responsibly and innovatively, aligning with the principles of responsible innovation and collaboration governance outlined by Emerson et al. (2011) and furthering the understanding of DP's potential beyond traditional philanthropic activities.

## References

- About (2023). ITU. <https://www.itu.int/en/about/Pages/default.aspx>
- About CC Licenses (2020, May 22). Creative Commons. <https://creativecommons.org/share-your-work/cclicenses/>
- Adrienne, P. (2023). *Case Studies Showcasing How Nonprofits Are Utilizing AI to Enhance Operations and Fundraising Efforts*. <https://www.linkedin.com/pulse/case-studies-showcasing-how-nonprofits-utilizing-ai-enhance-phillips/?trackingId=%2BxobWvLsRA%2BmFgXYS%2BsReg%3D%3D>
- AI Risk Management Framework | NIST (2021). <https://www.nist.gov/itl/ai-risk-management-framework>
- Art. 89 GDPR – Safeguards and Derogations Relating to Processing for Archiving Purposes in the Public Interest, Scientific or Historical Research Purposes or Statistical Purposes – General Data Protection Regulation (GDPR) (2016). General Data Protection Regulation (GDPR). <https://gdpr-info.eu/art-89-gdpr/>
- Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., & Shee, A. (2019). Towards standardization of data licenses: The Montreal Data License. In *arXiv [cs.CY]*. arXiv. <http://arxiv.org/abs/1903.12262>
- Big Data Philanthropy for Humanitarian Response (2012, June 5). IRevolutions. <https://irevolutions.org/2012/06/04/big-data-philanthropy-for-humanitarian-response/>
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data. *International Journal on Semantic Web and Information Systems*, 5, 33. [tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf](http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf)
- Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate governance of artificial intelligence in the public interest. *Information. An International Interdisciplinary Journal*, 12(7), 275. <https://doi.org/10.3390/info12070275>
- Copyright & Licences – Research Data – UNIGE (2017, January 31). <https://www.unige.ch/researchdata/en/share/rights/>
- Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns (New York, N.Y.)*, 4(10), 100857. <https://doi.org/10.1016/j.patter.2023.100857>
- Crook, T. R., Ketchen, D. J., Jr, Combs, J. G., & Todd, S. Y. (2008). Strategic resources and performance: A meta-analysis. *Strategic Management Journal*, 29(11), 1141–1154. <https://doi.org/10.1002/smj.703>
- Data Access and Licensing (2023). <https://datacatalog.worldbank.org/public-licenses>
- Data Donation Lab (2023). <https://www.ikmz.uzh.ch/en/research/divisions/media-use-and-effects/projects/datadonation-lab.html>
- Data Philanthropy, International Organizations and Development Policy: Ethical Issues to Consider (2020). UNDP. <https://www.undp.org/publications/data-philanthropy-international-organizations-and-development-policy-ethical-issues-consider>
- Digital Public Infrastructure (2024). UNDP. <https://www.undp.org/digital/digital-public-infrastructure>
- Emerson, K., Nabatchi, T., & Balogh, S. (2011). An integrative framework for collaborative governance. *Journal of Public Administration Research and Theory*, 22(1), 1–29. <https://doi.org/10.1093/jopart/mur011>
- Epstein, Z., Hertzmann, A., Investigators of Human Creativity, Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., Mahari, R., Pentland, A. S., Russakovsky, O., Schroeder, H., & Smith, A. (2023). Art and the science of generative AI. *Science*, 380(6650), 1110–1111. <https://doi.org/10.1126/science.adh4451>
- First Draft of the Recommendation on the Ethics of Artificial Intelligence (2020). <https://unesdoc.unesco.org/ark:/48223/pf0000373434>
- Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., Schneider, F. D., Kattge, J., Poelen, J. H., Madin, J. S., Ankenbrand, M. J., Penone, C., Feng, X., Adams, V. M., Alroy, J., Andrew, S. C., Balk, M. A., Bland, L. M., Boyle, B. L., ..., & Enquist, B. J. (2020). Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution*, 4(3), 294–303. <https://doi.org/10.1038/s41559-020-1109-6>
- Gautier, A., & Pache, A.-C. (2015). Research on corporate philanthropy: A review and assessment. *Journal of Business Ethics: JBE*, 126(3), 343–369. <https://doi.org/10.1007/s10551-013-1969-7>
- George, J., Yan, J. (kevin), & Leidner, D. (2019a). Data philanthropy: An explorative study. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. <https://doi.org/10.24251/hicss.2019.707>



- George, J., Yan, J. (kevin), & Leidner, D. (2019b). Data philanthropy: An explorative study. *Hawaii International Conference on System Sciences 2019 (HICSS-52)*. [https://aisel.aisnet.org/hicss-52/os/org\\_issues\\_in\\_business\\_intelligence/3/](https://aisel.aisnet.org/hicss-52/os/org_issues_in_business_intelligence/3/)
- Ghosh, S. (2019). Organising for Artificial Intelligence (AI) technologies. *Japan Social Innovation Journal*, 8(1), 1–19. <https://doi.org/10.12668/jsij.8.1>
- Godfrey, P. C. (2005). The relationship between corporate philanthropy and shareholder wealth: A risk management perspective. *AMRO*, 30(4), 777–798. <https://doi.org/10.5465/amr.2005.18378878>
- Gray, J. (2009, September 1). *Open licensing for philanthropic foundations – “Why not?”* Open Knowledge Foundation Blog. <https://blog.okfn.org/2009/09/01/open-licensing-for-philanthropic-foundations-why-not/>
- Guterres, A. (2020). Roadmap for digital cooperation. *United Nations*. June. [https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Update\\_on\\_Roadmap\\_implementation\\_April\\_2021.pdf](https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Update_on_Roadmap_implementation_April_2021.pdf)
- Heimstädt, M., Saunderson, F., & Heath, T. (2014). From toddler to teen: Growth of an open data ecosystem. *JeDEM – EJournal of EDemocracy and Open Government*, 6(2), 123–135. <https://doi.org/10.29379/jedem.v6i2.330>
- Hong, T. (2023, August 16). *Why digital public infrastructure matters*. Bill & Melinda Gates Foundation. <https://www.gatesfoundation.org/ideas/articles/what-is-digital-public-infrastructure>
- ITU Data Collection (2013). <https://www.itu.int/ITU-D/ict/datacollection/>
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258–268. <https://doi.org/10.1080/10580530.2012.716740>
- Juni, 14. (2023, June 14). *Current and potential AI use in Swiss philanthropic organizations – survey results*. SwissFoundations. <https://www.swissfoundations.ch/aktuell/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/>
- Kacorri, H., Dwivedi, U., Amancherla, S., Jha, M. K., & Chanduka, R. (2020). IncluSet: A data surfacing repository for accessibility datasets. *ASSETS. ACM Conference on Assistive Technologies*, 72. <https://doi.org/10.1145/3373625.3418026>
- Labastida, I., & Margoni, T. (2020). Licensing FAIR data for reuse. *Data Intelligence*, 2(1–2), 199–207. [https://doi.org/10.1162/dint\\_a\\_00042](https://doi.org/10.1162/dint_a_00042)
- Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060–1062. <https://doi.org/10.1126/science.aaz8170>
- Lev Aretz, Y. (2019). Data philanthropy. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3320798>
- License Selector (2015). <https://ufal.github.io/public-license-selector/>
- Malone, P. (2023, July 31). *An Evaluation of Private Foundation Copyright Licensing Policies, Practices and Opportunities*. Berkman Klein Center. [https://cyber.harvard.edu/publications/2009/Open\\_Content\\_Licensing\\_for\\_Foundations](https://cyber.harvard.edu/publications/2009/Open_Content_Licensing_for_Foundations)
- Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & SOCIETY*, 38(1), 411–423. <https://doi.org/10.1007/s00146-021-01308-8>
- Novak, M. (2023, November 6). OpenAI to pay legal fees of business users hit with copyright lawsuits. *Forbes Magazine*. <https://www.forbes.com/sites/mattnovak/2023/11/06/openai-to-pay-legal-fees-of-business-users-hit-with-copyright-lawsuits/>
- Peterson, D. K., Cathryn, V. L., & Pham, C. (2021). Motives for corporate philanthropy and charitable causes supported. *Journal of Strategy and Management*, 14(4), 397–412. <https://doi.org/10.1108/JSMA-09-2020-0241>
- Pierce, D. (2024, February 14). *With the rise of AI, web crawlers are suddenly controversial*. The Verge. <https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders>
- Press Release. (2023). 33(3–4), 242–242. <https://doi.org/10.1007/bf00383954>
- Radical Proposal: Third-Party Auditor Access for AI Accountability (2021). Stanford HAI. <https://hai.stanford.edu/news/radical-proposal-third-party-auditor-access-ai-accountability>
- Recommendation on the Ethics of Artificial Intelligence (2022). <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- Rudner, T. G. J., & Toner, H. (2021, March 17). *Key Concepts in AI Safety: Interpretability in Machine Learning*. Center for Security and Emerging Technology. <https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning/>

- Song, Q., Lee, C., & Han, L. (2022). The platformization of digital philanthropy in China: State, tech companies, and philanthropy engineering. *China Information*, 0920203X2211439. <https://doi.org/10.1177/0920203x221143940>
- Statement from Gates Foundation CEO Mark Suzman: Why We Need Digital Infrastructure*. (2022, December 1). Bill & Melinda Gates Foundation. <https://www.gatesfoundation.org/ideas/media-center/press-releases/2022/12/digital-public-infrastructure>
- Susha, I., & Gil-Garcia, J. R. (2019). A collaborative governance approach to partnerships addressing public problems with private data. <https://scholarspace.manoa.hawaii.edu/handle/10125/59726>
- Taddeo, M. (2016). Data philanthropy and the design of the infraethics for information societies. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 374(2083), 20160113. <https://doi.org/10.1098/rsta.2016.0113>
- Training*. (2022, October 3). Creative Commons. <https://creativecommons.org/about/training-and-consulting/training/>
- UNESCO Member States Adopt the First Ever Global Agreement on the Ethics of Artificial Intelligence* (2023, April 20). <https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence>
- United Nations (2018). *A Decade of Leveraging Big Data for Sustainable Development | United Nations*. <https://www.un.org/en/un-chronicle/decade-leveraging-big-data-sustainable-development>
- Verhulst, S. (2023). Computational social science for the public good: Towards a taxonomy of governance and policy challenges. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4338028>
- Voida, A. (2014). A case for philanthropic informatics. In S. Saeed (Ed.), *User-Centric Technology Design for Nonprofit and Civic Engagements* (pp. 3–13). Springer International Publishing. [https://doi.org/10.1007/978-3-319-05963-1\\_1](https://doi.org/10.1007/978-3-319-05963-1_1)
- Windl, M., Henze, N., Schmidt, A., & Feger, S. S. (2022, April 27). Automating contextual privacy policies: Design and evaluation of a production tool for digital consumer privacy awareness. *CHI Conference on Human Factors in Computing Systems*. CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans LA USA. <https://doi.org/10.1145/3491102.3517688>
- Wu, J. (2015). *Big Data Philanthropy: The Social Impact of Donating Data*. <https://www.linkedin.com/pulse/data-philanthropy-social-impact-donating-june-wu/>
- Xiong, B., & Zhang, D. (2023). *Introducing The Foundation Model Transparency Index*. Stanford HAI. <https://hai.stanford.edu/news/introducing-foundation-model-transparency-index>

# FROM MARGIN TO MAINSTREAM

## Moving philanthropy to reshape our AI-enabled future

*Yolanda Botti-Lodovico and Vilas Dhar*

### 1 Introduction

The digital economy came with a monumental promise: technology would generate new opportunities, greater efficiency, and “one of the most effective solutions to poverty” (Nielsen 2014; Lynch, 2018). But this promise never came to fruition; even as global GDP grew and technology advanced, inequality between and within countries increased, aggravated by climate change, global conflict, and the pandemic. As a result, the wealthiest 10% now enjoy 76% of global wealth, while over 160 million people have fallen into poverty since the COVID-19 pandemic (*Global Inequality Rises Again*, 2023).

The world is now stepping into another power transformation – spurred by the emergence of artificial intelligence (AI) – from which the philanthropic sector has been largely absent. This absence reverberates on civil society and communities at the frontlines of vulnerability. Without philanthropic leadership as a bridge between power and civil society, the social sector will miss out on critical opportunities to create human-centered AI to solve local and global challenges, including climate change, hunger, health inequities, and beyond. Across the globe, civil society will remain excluded from policy development despite their key role in shaping previous transitions of power throughout history – while private sector actors determine the future trajectory of our AI-enabled world.

As institutions of public trust, foundations must serve humanity – and AI has the potential to revitalize and reinforce this mission. The following sections frame challenges, opportunities, and the critical pathway to fulfilling this obligation and realizing AI’s potential. In the first section, we explore the structural and organizational obstacles the philanthropic sector will need to overcome to secure its place in the global AI agenda. Second, we analyze how foundations can help restore civil society’s voice in the ongoing transformation of power through innovative mechanisms of support and active partnership. Finally, we present a roadmap for organizations looking to invest in digital transformation – including practical approaches to AI and data solutions, responsible data and AI practices, and galvanizing global frameworks for ethical AI. Ultimately, by becoming technology-forward, foundations can propel the digital transformation process across the social sector and empower civil society to advance a shared mission of equity, justice, and human dignity for all.

## **2 Structural and organizational barriers to AI adoption in philanthropy**

Since the start of the digital revolution, philanthropy demonstrated a hesitancy to lead in the adoption of technology. The sector's risk aversion to digital transformation – relative to private corporations and businesses – stems partly from structural and organizational barriers such as cost, internal capacity, infrastructure, leadership, and trust (Lester, 2014). However, as technologies like AI and data solutions play an increasingly prominent role in society – and global challenges become ever more prevalent – digital transformation will prove critical to philanthropy's success as institutions of not merely public capital but public trust, serving humanity at scale.

### **2.1 Cost of digital transformation**

Across the philanthropic sector, internal investments toward digital transformation are limited. A 2022 report from the Technology Association of Grantmakers (TAG) revealed that half of all foundations surveyed allotted a mere 1%–5% of their entire operational budget to the purchase and adoption of technology, while only 15% allotted above 10% to technology (*2022 State of Philanthropy Tech*, 2022). In private sector companies, technology spending across industries receives around four times the amount invested in the social sector (*Roadmap for Funders*, 2020).

Low technology spending within foundations often stems from priorities and resource constraints. Some organizations prioritize grantmaking, direct service delivery, and programmatic outcomes over revamping internal operations. As a result, fewer resources have gone toward enabling their internal digital transformation process and setting up foundational infrastructure to enable future efficiencies and innovation (*Roadmap for Funders*, 2020).

Foundations are beginning to understand what the private sector grasped at the start of the digital revolution. At scale, investments in technological innovation can play a transformative role in nonprofits and civil society organizations. Emerging technologies like AI have already helped civil society changemakers effectively leverage data for policy change, improve the delivery of humanitarian aid, better understand and meet global health needs, and much more. However, without leadership and support from philanthropy, the social sector may miss the opportunity to fully harness the power of technology for good.

### **2.2 Low internal digital capacity**

Staffing and retention challenges often constrain a foundation's internal capacity for digital transformation. While technical staffing numbers have improved, foundations with lower asset sizes need additional support to catch up to their larger counterparts. Disparities worsen based on the type of the foundation. In 2022, private foundations reported nine technical employees to every one nontechnical employee, whereas public foundations reported only one technical employee to every 28 employees (*2022 State of Philanthropy Tech*, 2022). Without in-house technical experts or commitments to organization-wide digital learning, many foundations still need a solid basis for operational optimization, sector-wide learning, enhanced programmatic impact, and better collaboration. They also need to take advantage of critical opportunities to provide training and support to their grantees and boost their overall impact. Currently, few foundations are engaging in external capacity development to fill this gap; in 2022, only 39% of community foundations provided technical support and training to grantees, compared to 33% of private foundations (*2022 State of Philanthropy Tech*, 2022). Reticence to technology adoption across the sector impedes philanthropy's capacity to advance social impact and reduce costs in the future.

### 2.2.1. Limited data infrastructure

Data can help foundations optimize their giving practices, better target organizations in need, and compile best practices and insights for the social sector. However, too many foundations lack the infrastructure needed to collect, safely store, share, and analyze existing data. This challenge has persisted over time; in 2014, the United Nations assessed that “too many people, organizations, and governments” were excluded from the benefits of data use for sustainable development as a result of knowledge, capacity, and resource constraints (Nielsen, 2014). Four years later, a 2018 analysis from Alliance Magazine found that smaller foundations had not yet leveraged the power of data-driven giving due to resource and capacity constraints (Kassatly, 2018). Moreover, a 2023 report from the Dorothy A. Johnson Center for Philanthropy deemed the philanthropic sector “lightyears behind its peers” in data availability (Abalo et al., 2023).

Constructing a robust internal data infrastructure begins with quality data. Much of the data available across the philanthropic sector relies on independent stakeholders, whose efforts lack redundancy (Abalo et al., 2023). Foundations do not regularly or systematically collect data regarding their beneficiaries or constituents, ongoing needs, or complementary work by colleagues in the space (Anderson, 2014). Most existing data sources focus on grantmaking rather than impact, and the social sector at large is not bound by any legal obligations to share that data (Kassatly, 2018).

Further, foundations lacking digital maturity are often unable or unwilling to support grantees’ costly infrastructural investments. Overall, under 1% of all global donations go toward developing critical infrastructure to support technology adoption and use, reducing the social sector’s impact at large (*Roadmap for Funders*, 2020). For nonprofits, data and the ability to derive insights from that data are both critical to demonstrating impact for continued philanthropic support and better understanding the needs of the populations they serve. When foundations prioritize creating their own robust data infrastructure – complete with systematized metrics for data gathering, replicable standards for data sharing and privacy, and open-source tools to analyze and leverage insights from that data – they can help empower their partners to do the same. Foundations can also help generate critical information to guide decisions on both sides of the grantmaking process.

### 2.2.2 Willingness and engagement from leadership

On average, most foundations lack internal forcing functions around digital transformation. For those who have already taken the leap or plan to leap, the motivation to disrupt the status quo often comes from those at the top – driven by visionary leaders who understand the critical importance of investing in innovation and digitization throughout their organization. In 2012, a TAG and Grants Managers Network (GMN) joint report revealed that a mere 23% of foundation executives fully grasped technology’s benefits, while 66% of executives were “supportive but not very knowledgeable” (Lester, 2014). As AI entered the scene, the field of philanthropy fell further behind despite a relative increase in more tech-forward foundations. Still today, foundations without close and ongoing ties to the tech industry tend to maintain a more skeptical approach to AI (Dervishi, 2023).

Again, a lack of buy-in to innovate within a foundation often spills over into their engagement with grantees and partners. As a result, the institutions and organizations they serve cannot receive the support and collaboration they need to adopt technologies and leverage them for impact. A report from 2020 revealed that a major factor impeding digital transformation in philanthropy is “willingness,” that is, willingness to invest in digital infrastructure – which inhibits their capacity as civil society partners to drive sustainable solutions (*Roadmap for Funders*, 2020).

### **2.3 Trust and buy-in from communities**

The growing dissemination of emerging technologies like AI into the public sphere, as well as increasing outcries around the risks and dangers, create a new set of obstacles to adopting technologies across the philanthropic sector. Communities, civil society, and data ethicists have expressed rightful concerns about AI perpetuating unwanted biases, the lack of equitable representation in data sets that drive funding decisions, and the dangers of further exploiting already marginalized populations who too often miss out on the benefits of global giving and progress. For digital transformation to be successful, foundations, their grantees, and the communities they serve must fully trust in the ethical development and use of technologies.

The homogeneity of individuals and organizations participating in important conversations around the development, use, and ownership of technologies – not only within foundations but within society at large – further complicates building a culture of trust. For example, a mere 22% of AI professionals globally identify as women, based on statistics from the United Nations Educational, Scientific and Cultural Organization (UNESCO) (*International Women's Day: New Factsheet Highlights Gender Disparities in Innovation and Technology*, 2023). In 2019, a report from Bloomberg revealed that under 2% of Google and Facebook's technical employees were Black. These disparities, rooted in systemic inequities and exclusionary practices, impact technology design, leading to greater risks of harm in their deployment: from faulty facial recognition software, to algorithmic biases that withhold social security benefits from individuals with disabilities and beyond (Firth-Butterfield et al., 2022).

The present state of diversity, equity, and inclusion (DEI) efforts within foundations – either via in-house DEI experts or training programs – creates further challenges. The lack of mature DEI outcomes further reduces the internal capacity for culturally responsive and informed digital approaches. While a 2022 survey of foundations revealed improvements in several internal technical teams undergoing DEI training, 30% of all private foundations surveyed had not yet provided DEI training to their tech teams (*2022 State of Philanthropy Tech*, 2022). Survey respondents shared that of the existing DEI training programs available, almost no updates or changes had been made since 2020 – indicating an intention to build awareness but little progress toward implementation. In the age of AI, foundations are responsible for prioritizing DEI principles by including historically marginalized voices in their digital transformation process, fostering diverse AI and technical talent across the social sector, and co-creating a digital future with communities that elicit trust and courageous buy-in from everyone.

In sum, philanthropy is uniquely positioned to collaborate with other sectors in addressing the significant challenges faced by humanity, – from climate change to food insecurity, to wide-scale labor market disruption. As a bridge between traditional mechanisms of power and civil society, foundations have transitioned from pure grantmaking engines to institutions of public capital and public trust, poised to make the world a better place for all people. To do so, they need internal buy-in, broad-based participation, and ongoing commitments from their organization and the organizations they serve – to help build equitable and robust solutions driven by civil society. In the following section, this chapter will map out how philanthropy can leverage its access to power and resources to propel the social sector forward in the age of AI.

## **3 Driving social sector transformation at scale**

Public awareness and engagement with the possibilities of AI tools have triggered the start of a transformation of power, one that has the potential to shift the course of human history. Presently, only a subset of stakeholders, including Big Tech and wealthy governments in the Global North,

have monopolized decisions around how technologies are designed, used, and shared across sectors, geographies, and communities. This form of concentrated decision-making, by its very structure, serves to perpetuate the status quo and further excludes many in the Global Majority, as well as marginalized communities across the Global North.

By engaging early and intentionally with AI, philanthropy has an opportunity to not merely optimize its internal operations but, more importantly, create new systems of power driven by our shared human values. Through innovative partnership mechanisms, philanthropy can help build civil society capacity at the frontlines of vulnerability to better understand problems through data and create uniquely responsive AI tools to solve them. In partnership with governments and civil society, philanthropy can help foster broad-based digital and AI literacy that empowers communities to fully participate in this transformation. At scale, philanthropy can help ensure that civil society plays a key role in bending the arc of the AI revolution toward justice and equity – so that our newfound capabilities serve both people and the planet.

### ***3.1 Optimizing philanthropic operations***

Overcoming reticence to AI adoption in philanthropy opens the doorway to optimized internal processes and accelerated external impact. On an organizational level, digital transformation requires that every member of the organization commit to developing a foundational understanding of data and AI (Dhar & Firth-Butterfield, 2021). Leadership is responsible for creating that space – to encourage individual learning, set goals and benchmarks, and recruit in-house technical expertise. On a sectoral level, digital transformation requires collaboration, regular sharing of best practices, and ongoing accountability across organizations to infuse the social sector at scale with human-centered AI and digital technologies.

Artificial intelligence can help foundations facilitate engagement, restructure financial diligence, streamline internal processes, and generate new strategic frameworks for sector-wide progress. In terms of grantee-funder interactions, several AI tools exist to help increase efficiencies on both the donor and grantee sides and improve pathways for lower-resourced organizations to prepare and review application materials. For example, nonprofits might turn to ChatGPT to craft a grant proposal or their mid-term and final reports. They might use DALL-E to incorporate unique visuals into their donor presentations. For changemakers with disabilities, limited literacy, or language barriers, AI tools like voice-to-text can help build pathways to more effective storytelling and capturing reports from the field. AI-powered chatbots can produce a “conversational interface” for real-time engagement between donors and grantees – allowing for quick feedback and alignment on goals and visions. AI can also reduce bureaucratic obstacles and the risk of errors throughout the funding lifecycle – empowering understaffed nonprofits from the application stage to the final report (Beasley, 2023).

With funding secured, AI can also help streamline due diligence and other internal processes. It can help match financial patterns and reveal trends, build financial models, automate reporting and expense management, and create real-time summaries, analytics, and visual dashboards around grantmaking activities. Foundations can also use AI to reduce the workload of their programmatic staff and free space for more direct engagement with grantees. New, creative ways of partnership may include capacity building, joint problem-solving, thought leadership, content sharing, and field-based learning.

At scale, AI can improve equity and inclusivity across the philanthropic sector. AI-driven platforms may contribute insights on grantmaking trends and patterns, identify areas for collaboration

and improvement, and help devise solutions to complex challenges while encouraging accountability across organizations. For example, using AI to understand the racial and socioeconomic demographics of grantees can help root out issues of bias or exclusion in funding decisions. Technology can also help facilitate more open and equitable application processes that better align priorities between parties and also connect nonprofits to additional grant opportunities.

### ***3.2 Pioneering new models of philanthropic partnership***

Foundations play an important role in ensuring that civil society is prepared and empowered to harness the potential of AI for advancing human dignity, equity, and justice across the globe. Tech-forward foundations are already developing visionary approaches to empower AI adoption across the social sector. Leveraging their proximity to capital, power, and public trust, foundations have partnered with civil society to build AI-enabled resources to advance organizations of all sizes and missions.

Three innovations have emerged as core elements of this future-ready partnership model, which foundations are uniquely positioned to scale. These innovations include in-house teams dedicated to building grantee capacity in data management and the applications of AI; collaborative data stores for multi-sectoral collaboration on complex issues like climate change; and ongoing initiatives for fostering diverse technical talent and expanding AI literacy, both in organizations and in society at large.

#### ***3.2.1 In-house capacity building and product development***

The first innovation shifts away from the traditional model of giving to prioritize pathways for ongoing and active collaboration through capacity building, support, and mechanisms for mutual learning. In addition to ordinary grant giving, the Patrick J. McGovern Foundation incorporates an in-house team of data scientists, software engineers, product managers, and critically, user experience design experts to work directly with grantees as they tackle a range of data and AI challenges. These partnerships have not only advanced the overall data fluency of participating organizations but have also helped surface ideas for new products, raised common pain points and opportunities along the digital transformation journey, and driven the creation of collaborative solutions.

For example, the Patrick J. McGovern Foundation's Products and Services team fills a critical gap in data management and analytical capacity among social sector organizations through high-touch partnerships, informative webinars, peer learning events, and training programs. Past iterations of the Foundation's Accelerator program have focused on topics such as how to use data responsibly to safeguard human rights and drive climate action, a collaborative design studio to support cross-organizational collaboration at the nexus of climate and health, and content creation around how to both develop and scale effective data strategies and stewardship. Participating organizations of diverse backgrounds, sizes, and experience levels receive hands-on technical guidance and partnership to overcome data challenges, build sustainable data environments that support robust AI applications, and forge partnerships with like-minded organizations working toward similar goals – all while developing internal capacity to seek, receive, and use grant funding to build products and tools that advance their mission.

Alongside developing partner capacity, this in-house team of scientists and engineers works closely with stakeholders ranging from peer funders, to government agencies, to private technology companies, and grant partners to directly translate civil society insights into practice.



With regular input from these communities, they co-design and build products that directly respond to community needs that are not served by market-oriented products. Their work starts by identifying a challenge often revealed through lived experience, direct service delivery, and/or data from an organization. Next, the teams conduct research and learn from communities, other organizations, and product developers who have worked on addressing similar challenges. They then devise a “solution plan” and product roadmap, and after undergoing relevant reviews and approvals, they build the product for deployment in the field. Through collaborative brainstorming, curiosity, and diverse expertise, the team has created a range of valuable products – including a Large Language Model (LLM)-backed product designed to assist resource-constrained journalists in better engaging with communities on social media. Products in development include a financial diligence tool to help foundations mitigate financial risks among grantees and an AI-powered mobile application that generates financial and safety impact predictions for climate disasters across the globe. These activities not only produce usable products but also build a library of experience-driven best practices for deploying AI and data solutions. These best practices, designed to ensure responsible and ethical outcomes, can then be shared with partners and private sector technology companies to guide broader trust and safety initiatives in technology development.

### 3.2.2 Collaborative data stores and standards

Beyond capacity reserves within individual organizations, the social impact sector often faces a collective action problem. Despite the abundant data on various global issues, organizations too often struggle to leverage it for insight and impact. Part of this challenge stems from the limited capacity within individual organizations to manage data properly and leverage it for shared insights. Other challenges include the unmet need for shared and interoperable standards for aggregating and analyzing data across sectors and geographies. For example, an IMF blog from 2021 noted that over 200 standards and frameworks guide the climate disclosure and sustainability reporting practices of 40 different countries (Ferreira et al., 2021). The social sector needs a new collaborative, standardized approach to inspire sustainable data transformation across issue areas.

As organizations become more comfortable managing their data, foundations can provide much-needed support in empowering them to share that data and access additional data from other stakeholders across the broader landscape. Collaborative data stores – populated by civil society and sustained by philanthropy – can improve our understanding of evolving needs and align objectives and priorities across sectors and geographies. Global reporting standards for different issues can also help improve data interoperability for better insights. At scale, these mechanisms can revolutionize our ability to solve global challenges.

In the context of climate change, philanthropy is already starting to support the creation of collaborative data stores and standards to improve shared efforts toward mitigation, adaptation, and resilience. A partner of the Patrick J. McGovern Foundation, Creative Commons is working with multi-stakeholder partners to support the open sharing of extensive, standardized climate data sets through its Open Climate Data project. Creative Commons also facilitates more open sharing of research outputs as the norm in climate science via its Open Climate Campaign (Creative Commons, 2022; *Open Climate Data*, 2023). Organizations like Climate TRACE are working with philanthropy and global partners to improve inventories on greenhouse gas emissions through AI (*Climate TRACE*, n.d.). Adapting these models of data partnership can help tackle challenges beyond climate change – from health crises to poverty to rebuilding infrastructure post-conflict.

### *3.2.3 Fostering technical talent and AI literacy*

As AI creates new opportunities for profit and economic participation, a small subset of the global population has reaped the benefits. Those holding the most significant advantage include Big Tech and employees of Big Tech, predominantly hailing from the Global North. The philanthropic sector has an opportunity to re-engineer this dynamic, ensuring that more communities in the Global South and marginalized communities across the Global North share in the process and fruits of technological innovation. In doing so, they can help infuse both our AI creation and our emerging societal infrastructure with principles of equity and justice.

Professional development programs – jointly architected by civil society and philanthropy – are one pathway to intentionally fostering a more diverse and inclusive tech workforce. For example, AkiraChix is a nonprofit organization that works with philanthropy to build the world’s leading source of African female technological talent. Started by female software engineers in Kenya, the organization ensures that young women from underserved communities have the resources and tools to both prosper in a technical career and live fulfilling lives (*AkiraChix*, n.d.). Other organizations like Per Scholas are working with philanthropy and employers to bridge the digital skills gap in the United States. Their robust technical education programs for professionals are easily accessible and tuition-free and contribute to creating a more diverse workforce in tech. With over 40% of the learners being women and another 85% people of color, their efforts provide a replicable model for enabling a more inclusive and representative technology sector (*Per Scholas Expands Tech Training to Houston*, 2023).

Including more communities in the process and benefits of tech creation will also require broader AI literacy in society at large. At present, AI literacy rates across the globe remain low; for many, improvements are unattainable without improved access to digital tools. A 2021 survey cited by the World Economic Forum revealed that 84% of Americans lack literacy in AI (Boyle, 2021). The situation is even worse in low- and middle-income countries, bringing the global total of people offline to 2.6 billion (International Telecommunication Union, 2023). Women are disproportionately excluded (*CSW67 Opening Statement: Digital Rights Are Women’s Rights*, n.d.). With 60% of our world GDP driven by digital infrastructure, these disparities have rippling effects across society, perpetuating legacies of racial, social, and gender-based inequalities (Azagury, 2023).

In response, organizations like Team4Tech, a nonprofit impact accelerator, are working to ignite transformative change. Through integrating digital and AI education across a spectrum of programs – including teacher training initiatives on a global scale – Team4Tech is cultivating community-level capacity in both the utilitarian application of AI and its iterative improvement (*Team4Tech*, 2023). For Team4Tech’s program participants, success in an AI-driven future is at reach. However, creating not just AI-literate but AI-empowered societies at scale demands broad multi-sectoral support and participation – from philanthropy, governments, the private sector, and civil society alike.

### *3.3 Amplifying the civil society voice and impact*

AI and the digital revolution will transform how citizens engage with the political, economic, and social structures shaping their lives. In similar events throughout human history, civil society played a foundational role in protecting rights and holding new leaders accountable. Civil society actors helped shape transitions in leadership after the fall of autocratic and colonial regimes. They secured the right to vote for women and other marginalized communities. Still today, civil society

works to counter the human and environmental impacts of transnational corporations' activities, fighting to protect the rights of communities across the globe. Civil society's unique proximity to and even experience with the struggles of humanity empower them to tackle the vulnerabilities that change creates.

In the context of AI and the digital revolution, however, civil society has faced significant limitations in capacity, understanding, and accessible pathways for participation. As a result, the ongoing transformation of power caused by emerging technologies has suffered from an evident imbalance in who benefits and who is made more vulnerable. According to a 2022 survey from Salesforce, only 12% of nonprofits across the globe classify themselves as "digitally mature" (*Nonprofit Trends Report*, 2022). These organizations are not only at a disadvantage when working to achieve their own goals but they are also missing opportunities to prepare the communities they serve to succeed in a digital future. Traditionally, the philanthropic sector would help build those pathways, but their capacity and understanding of AI remain inadequate.

Civil society and philanthropy play unique roles in the AI revolution. It is up to civil society to advocate for and protect the rights of vulnerable communities while empowering them to leverage the opportunities at hand. It is up to philanthropy to acquire the skills necessary to support civil society learning and bolster their efforts. An empowered civil society and philanthropy can help ensure that communities everywhere have access to the tools, opportunities, and skills to both participate in and design our shared digital future.

### 3.3.1 Case studies: harnessing the power of AI for impact

The benefits of AI adoption across the social sector range from optimization of internal operations to more effective delivery on strategic outcomes. Given their proximity to communities, AI fluency within civil society organizations can help enable human-centered AI solutions for long-term impact. The following examples illustrate how civil society is currently working with philanthropy to empower communities at the frontlines of vulnerability and tackle global challenges at scale.

In the climate sector, civil society organizations are leveraging philanthropic support to enable community-driven AI development. For example, in-house technologists at *Digital Democracy* collaborated with Indigenous Earth defenders to create the Earth Defenders Toolkit. The product incorporates cutting-edge AI to help communities predict and stop illegal logging and poaching on ancestral lands. Similarly, tools like Mapeo help communities map and monitor where environmental issues like pollution might have occurred and respond accordingly. This model of partnership acknowledges the critical wisdom of Indigenous communities as stewards of the earth and emphasizes that the most effective solutions center the needs and perspectives of communities at the frontlines of the fight. Emily Jacobi, the Co-Executive Director and Founder of Digital Democracy stated:

The tools we've co-developed with communities and academic researchers have had an enormous impact. In the context of pollution – for example – we've been able to apply machine learning and satellite imagery to figure out where pollution has occurred, and where it might not be safe to get drinking water or fish. So what might have seemed like a scientific or academic question at first, becomes real life or death guidance for communities on the ground – the ones who are living that reality.

(E. Jacobi, personal communication, September 14, 2023)

In global health, AI has amplified civil society and nonprofit capacity while helping to inform government interventions and support healthier communities. *Khushi Baby*, a local health organization

in India, has leveraged years of insights from community health workers to develop an integrated digital health platform (CHIP) focused on improving maternal and child health outcomes. Now used by more than 70,000 government health workers, the platform provides critical health services to 40 million people and has expanded to incorporate immunization services, treatment for noncommunicable diseases, and public health surveillance. The organization now leverages the platform to provide technical support to the Department of Health and Family Welfare in the Government of Rajasthan. It is scaling to other Indian states while contributing data to advance precision health care across communities. The CEO and Co-Founder of Khushi Baby, Dr. Ruchit Nagar, stated:

There is a lot of enthusiasm in public health around the potential applications of AI and data. At Khushi Baby, we hope to see a future where both health workers and health officials alike feel empowered to speak to their data, to ask questions of it, and to learn in collaboration with others. This level of success requires a broader public health ecosystem that is dedicated to supporting all stakeholders in the health space – from government officials, to frontline community health workers, to patients – with appropriate quality controls and oversight.

(R. Nagar, personal communication, September 11, 2023)

As AI introduces new threats to the survival of our democracy, a digitally informed and AI-empowered civil society is equally critical to help strengthen democratic infrastructure for the future. For instance, the *CyberPeace Institute* partners with philanthropic actors and other multi-stakeholder organizations to combat digital threats against humanity, focusing on vulnerable communities, supporting nonprofits to enhance their cybersecurity posture, and monitoring the enforcement of global laws and norms around responsible behavior in cyberspace. The tools they have built are uniquely human-centered. Designed with an intimate understanding of the dangers that both development and humanitarian nonprofits, as well as human rights defenders, face, they enable reporting, collection, and analysis of dangers and abuses – a critical piece of a thriving democracy – while protecting civil society from malicious actors, creating knowledge and awareness, and building capacity for long-term cyber resilience. Francesca Bosco, the Chief Strategy and Partnerships Officer at the CyberPeace Institute, stated:

Despite the opportunities, technologies like AI have the potential to create new risks and vulnerabilities in the realm of cybersecurity. This is especially true for civilians, human rights defenders, and development and humanitarian workers, caught in the crossfires of geopolitical conflicts, natural disasters, and other systemic risks. At the CyberPeace Institute, we believe these same tools — if combined with robust laws and policies that are based on human rights and responsible behavior — can help ensure that every person on the planet enjoys safety, freedom, equity, and dignity in a digital age.

(F. Bosco, personal communication, September 17, 2023)

### *3.3.2 A new social compact for digital justice*

The present discourse on ethical technology – or ethical AI – reduces a conversation about normative ethics to one of applied decision-making – that is, AI becomes a tool to achieve different goals. The private sector regards it as a tool for profit and process optimization. Governments perceive it as a tool to increase efficiency and reduce bureaucratic errors in public service. The social sector is beginning to leverage the opportunities of AI to accelerate its goals for impact and empower communities too often excluded from the benefits of technological progress. As AI becomes fully

embedded into society, it is increasingly vital that we understand humanity's role in the power transformation at hand. Humans decide whether we use AI to perpetuate and intensify the unjust status quo or build a new ethical society – defined by the foremost values of today. The latter pathway both leads to and cements a commitment to digital justice.

Scholars at the World Economic Forum have defined digital justice as a righting of “past wrongs” or harms that a particular person or group of persons have undergone in the digital space, from bias and discrimination to exploitation and reputational injury (Warren et al., 2021). Correcting such harms might involve providing access to some redress mechanism or recourse determined via due process. Alternatively, the Detroit Digital Justice Coalition defines digital justice more broadly, elevating four key principles: access, participation, joint ownership, and healthy communities. The coalition emphasizes the importance of integrating into the digital ecosystem all those who have been “traditionally excluded from and attacked by media and technology” (Aguilar, 2015). Both definitions are rooted in fundamental human rights – the right to freedom from harm, the right to communication and self-expression, and the right to human dignity, or digital dignity – the cornerstone of digital justice.

The philanthropic sector must work with civil society to advance digital justice – including AI justice – in society. However, their partnership has the potential to not merely right past wrongs or create equitable access and opportunity in the digital space, but to build community-driven platforms of power that replace a historically flawed system with a transformative system that embeds human values more deeply into our technological and social framework. It demands a new social compact in which the social, public, and private sectors unite to share tools and capabilities, create robust standards for design and decision-making, and distribute the benefits of technological progress to every person on the planet.

To realize this vision, philanthropy can work with civil society to create a shared vernacular for advocacy and pave new pathways to shape AI implementation and governance. Supported by the Patrick J. McGovern Foundation, organizations such as the Center for Artificial Intelligence and Digital Policy and The Future Society, among others, are leading the way (CAIDP, n.d., *The Future Society*, 2016). As they convene transnational policymakers, AI experts, and civil society leaders to advance conversations for human-centered governance of AI, these groups are building new mechanisms for institutional innovation, critically examining the implementation of AI policy frameworks, and producing targeted policy recommendations that promote human dignity and justice in an AI era.

At scale, establishing new pathways for impact and social norms to guide the AI revolution is a fraught and complex multi-stakeholder process – but philanthropy and civil society must play a critical role in bridging the gap in AI product and platform ownership – to ensure that communities become not just users but architects of our digital future. Digital access and literacy are just the first steps. A united, tech-forward, and visionary social sector can help communities move beyond digital literacy to digital agency – to self-advocate, to create tools that capture their human aspirations and social mission, and to participate in and shape the digital ecosystem. As they step into their rightful position as architects, people everywhere will have the power to cement community values in our AI-enabled future, bolstered by new systems and structures perpetuating those values across society.

#### **4 Practical recommendations and a roadmap for the future**

The pathway to civil society leadership in the AI era will require ongoing commitments and buy-in from philanthropy and civil society organizations. This chapter has demonstrated the benefits of

social sector fluency in AI – to empower civil society and philanthropy and ensure that emerging technologies are designed and regulated to protect human interests and dignity. New technical understanding can strengthen their capacity to work in tandem with public and private leaders for community-centered objectives. From broad-based internet connectivity to increased access to critical technologies, aligned efforts can help expand pathways to digital opportunity for all communities. At scale, a digitally empowered civil society can work with communities across the globe – to build new tools that directly address their needs and, ultimately, shape the social transformation happening because of AI.

The practical guidance below can help advance the efforts of individual organizations along their digital empowerment journey. Through broad-based collaboration, the social sector can help alter the trajectory of the AI revolution and help achieve a more equitable, just, and human-centered future for all.

#### ***4.1 Digital transformation across the social sector***

Today’s social sector has a unique responsibility to proactively acquire newfound capacity, skills, and resources to harness the possibilities of AI for human progress. Visionary, aligned, and tech-forward foundations can provide their grantees a critical ecosystem of support, championing human-centered technological creation, use, and governance across civil society. Rather than one-off, isolated interventions, it requires sustained commitments, sharing tools and best practices, and ethical standards. Across the sector, it demands courage – to learn, grow, and evolve in response to the changing digital space. Philanthropy and civil society can solve previously insurmountable challenges and embed values of human agency, dignity, and participation into our AI-enabled future.

##### ***4.1.1 Embarking on digital transformation as a philanthropy***

Building internal capacity to understand and use these tools effectively is the first step for foundations. This may require upskilling and reskilling initiatives to ensure a foundational level of digital and AI literacy across all teams. As goals, technologies, and priorities evolve – both within foundations and across the social sector – executive leadership should not only make continual learning opportunities available but also incentivize employees to update their knowledge base regularly. Similarly, foundations can work together to create shared learning platforms for the entire sector, complete with best practices and generalizable toolkits, networks of digital experts and advisors for common pain points, and mentoring opportunities for each other and their grantees.

Foundations should also invest in developing organization-wide ethical guidelines and expectations for using AI and data specific to their needs and context. Various resources can help guide these efforts, whether among boards of directors, executive leadership, program teams, or beyond. For example, Vilas Dhar’s LinkedIn course on “Ethics in the Age of Generative AI” provides a step-by-step framework for everything from responsible data management to effective communication principles around AI (Dhar, 2023). As the most widely viewed course on ethics in the age of generative AI, it provides a baseline for integrating ethical AI principles and practices within an organization. As new needs arise, the respective leaders, in-house technical experts, and DEI professionals within each organization should regularly revisit and tailor their principles for relevance and troubleshooting – guided by feedback from the teams and grantees they serve.

#### *4.1.2 Driving digital transformation among grantees*

Once internal transformation is achieved, foundations can transfer that technological know-how to civil society to help harness their passion, knowledge, and strength for good. Foundations may support nonprofit actors through tools such as “large, unrestricted grants” so that nonprofits can disburse funds as needed without preset limitations (Fleming et al., 2023). These grants can, in turn, help accommodate investments in technology or in-house technical staff. They may also enable the creation of professional development programs to train teams across the organization in responsible stewardship and management of data for impact.

Although unrestricted grants constitute a notably “rare” practice in grantmaking today, findings reveal that they can spur digital transformation within nonprofits (Fleming et al., 2023). A recent study by the Center for Effective Philanthropy and Panorama Global demonstrated the early impacts of significant, unrestricted funds from MacKenzie Scott to be “dramatically and profoundly positive,” allowing “more innovation and risk-taking” (Buteau et al., 2022). Researchers then cross-examined their findings against another group of 21 U.S.-based Ballmer group grantees who received large unrestricted grants in 2017 that were dispensed across five years. Program-specific impacts included increased investments in digital delivery innovations and scaling of programs. The report also revealed that 89% of grantees applied the funds to core operations improvement, including data systems strengthening, fundraising, and staff development (Fleming et al., 2023). In the AI era, these findings call for philanthropies to take more risks and provide flexible support that allows nonprofits to upgrade internally via incremental technology adoption or organization-wide digital transformation.

For nonprofits already advanced in their digital understanding, trust-based, unrestricted support can empower ongoing innovation and adaptation to community needs. Girl Effect is an international nonprofit that builds tech to help girls make informed choices and changes during their critical years of adolescence (Effect, 2022). Reaching millions of girls across Africa and Asia, they are developing AI-integrated chatbots with and for adolescent girls and young women that provide more personalized and tailored responses to questions about sexual and reproductive health information and services, among other topics. This includes their Big Sis chatbot in South Africa, Bol Behen chatbot in India, and WAZII chatbot in Kenya.

In a recent interview with the Patrick J. McGovern Foundation, Karina Michel, Girl Effect’s Chief Creative and Technology Officer, reiterated the importance of funders who are “willing to take a bet” on their grantees (K. Michel, personal communication, August 22, 2023). She emphasized that funders with less familiarity and comfort with technology are less likely to invest in tech-inspired ideas or recognize the monumental expense of backend infrastructure that allows nonprofits to maintain their digital systems over time. However, philanthropies that have taken that first step toward digital fluency are more readily aligned with visionary risk-taking across the social sector. Through unrestricted funds, they can help reduce the infrastructural burden and costs of those risks on grantees themselves – enabling the creation of new products, platforms, and ongoing data analysis.

In addition to providing funding, digitally savvy foundations might support grantees through more active partnership as they experiment with and apply new technologies. The philanthropic partnership innovations we previously described – including formalized capacity-building services, product development support for new ideas, or even ad hoc technical consulting – can play a critical role in empowering digital transformation across the social sector. With easy access to trusted expertise in digital and AI systems and the flexibility to take risks, grantees will have more opportunities to build foundational capacity and also innovate according to their own needs and objectives.

## **4.2 Global ethical frameworks for AI**

While establishing organizational standards for ethical practices around AI and data is essential, sector-wide collaboration and goal-setting are critical to keeping all stakeholders accountable. Philanthropy can play a key part in convening stakeholders across the social, public, and private sectors. Together, these stakeholders can set benchmarks for open sharing of data and access to compute, inform national policies and practices for AI creation and use, and build common resources for support and learning along the digital journey.

The Centre for the Fourth Industrial Revolution Centre for Trustworthy Technology models multi-stakeholder collaboration between philanthropy, business, government, civil society, and academia to advance ethical principles for trustworthy technology. Launched in June of 2023 and supported by the Patrick J. McGovern Foundation and Deloitte, the center builds on foundations of expert thought leadership around emerging technologies like AI, digital reality, blockchain, the metaverse, and quantum computing. Their mission is to develop frameworks and tools to promote the ethical development and use of technology, with principles of inclusion, diversity, equity, and environmental sustainability at its core (*World Economic Forum Launches the Centre for Trustworthy Technology*, 2023). Other initiatives, such as the AI Governance Alliance at the World Economic Forum, convene multi-stakeholder actors to support the development of responsible and transparent AI systems, foster innovation, and promote inclusive and values-driven progress across the AI landscape (*AI Governance Alliance*, n.d.). Both efforts are supported by philanthropy and demonstrate the critical importance of cross-sector partnerships for greater accountability, broader reach, and sustained impact.

Philanthropy has also helped support the creation and implementation of global frameworks and institutional standards for responsible and ethical AI. One example is the UNESCO Recommendation on the Ethics of Artificial Intelligence – the first global standard for AI ethics (United Nations Educational, Scientific, and Cultural Organization, 2022). Developed to preserve human rights and dignity, this body of work emphasizes values such as diversity, inclusion, and peace. It also names more targeted and actionable principles of fairness and nondiscrimination, transparency and explainability, right to privacy, human oversight, and beyond. UNESCO applies the values and principles to critical policy action areas – ranging from gender equity, to health and education. The Patrick J. McGovern Foundation is now working with UNESCO to operationalize their recommendations further and increase public sector capacity in the Global Majority to engage with AI in ways that support civil society and empower communities.

## **4.3 Taking action with urgency**

The AI race has already begun, and the social sector is struggling to keep pace. For the select few dominating the AI landscape in the private sector, AI is creating new pathways to increase profit, optimize operations, and meet consumer preferences. These capabilities can have profound implications for community empowerment – from improved health outcomes to more efficient public service delivery and data-driven policies for a more just and equitable society. However, civil society must first gain a seat at the table, not merely as a participant but as a leader and architect of the conversations at hand. With aligned and intentional effort, philanthropy is well-positioned to help secure its seat.

Throughout this chapter, we shared critical interventions for philanthropy to propel civil society forward into the AI and digital landscape. First, foundations must educate themselves and overcome their barriers to digital transformation and AI adoption, leveraging not only their access



to resources but also the broader network of forward-thinking foundations already building accessible inroads into this space. Next, foundations must work hand-in-hand with civil society to drive social sector transformation at scale. They can achieve this goal by integrating AI to optimize their interactions with grantees and by implementing new mechanisms of active partnership. Foundations can also help amplify civil society voices in public decision-making by working with them to create a shared language for community advocacy and a broad-based understanding of digital justice. Finally, we identified practical steps every stakeholder in the social sector can take to ensure a human-centered AI future.

AI brings profound opportunities – and responsibilities – to philanthropy in the unfolding of our digital future. The words of Dr. Martin Luther King Jr. from decades ago invite new meaning today: the world is faced with a “fierce urgency of now” – in which our present decisions and actions will shape our future. With renewed alignment between philanthropy and civil society, we can harness AI to finally achieve that hailed promise of a digital economy and, ultimately, bend the arc of our AI-enabled world toward justice for all.

## References

- 2022 *State of Philanthropy Tech* (2022). Technology Association of Grantmakers. <https://www.tagtech.org/page/philanthropytech2022>
- Abalo, T., Peterson, K., Akaakar, A., Robinson, M., Behrens, T., Robinson, P., Couturier, J., Sharp Eizinger, M., Dietz, K., Spicer, T., Layton, M., Vuyst, A., Martin, T., Williams, J., & Moody, M. (2023). *11 Trends in Philanthropy for 2023*. Dorothy A. Johnson Center for Philanthropy. <https://johnsoncenter.org/blog/11-trends-in-philanthropy-for-2023/>
- Aguilar, E. (2015, November 12). *Basics of Digital Justice*. NFCB; National Federation of Community Broadcasters. <https://nfcfb.org/basics-of-digital-justice/>
- AI Governance Alliance (n.d.). Retrieved September 6, 2023, from <https://initiatives.weforum.org/ai-governance-alliance/home>
- AkiraChix (n.d.). Retrieved November 29, 2023, from <https://akirachix.com/>
- Anderson, G. (2014, April 1). *Data Collection and Analysis in Philanthropy*. The Conference Board. <https://www.conference-board.org/publications/publicationdetail.cfm?publicationid=7333>
- Azagury, J. (2023, January 10). *How Closing the Digital Divide Can Improve the Global Economic Outlook for 2023 and Beyond*. Fortune. <https://fortune.com/2023/01/10/how-closing-the-digital-divide-can-improve-the-global-economic-outlook-for-2023-and-beyond/>
- Beasley, S. (2023, June 14). *Philanthropy Needs to Embrace AI and Fast, Experts Say*. Devex. <https://www.devex.com/news/philanthropy-needs-to-embrace-ai-and-fast-experts-say-105652>
- Boyle, A. (2021, December 9). *Survey Suggests 84% of Americans Are Illiterate about AI—So Here’s a Quiz to Test Your Own AI IQ*. GeekWire. <https://www.geekwire.com/2021/survey-suggests-84-of-americans-are-illiterate-about-ai-so-heres-a-quiz-to-test-your-own-ai-iq/>
- Buteau, E., Buchanan, P., Lopez, M., Malmgren, K., & Im, C. (2022). *Giving Big: The Impact of Large, Unrestricted Gifts on Nonprofits*. The Center for Effective Philanthropy. <https://cep.org/report-backpacks/giving-big-year-one/?section=intro>
- CAIDP (n.d.). Center for AI and Digital Policy. Retrieved November 29, 2023, from <http://caidp.org>
- Climate TRACE (n.d.). Retrieved November 29, 2023, from <https://climatetrace.org/>
- Creative Commons (2022, December 19). *Patrick J. McGovern Foundation Funds New CC Initiative to Open Large Climate Datasets*. Creative Commons. <https://creativecommons.org/2022/12/19/patrick-j-mcgovern-foundation-funds-new-cc-initiative-to-open-large-climate-datasets/>
- CSW67 Opening Statement: *Digital Rights Are Women’s Rights* (n.d.). UN Women – Headquarters. Retrieved November 29, 2023, from <https://www.unwomen.org/en/news-stories/statement/2023/03/csw67-opening-statement-digital-rights-are-womens-rights>
- Dervishi, K. (2023, July 24). Foundations Seek to Advance A.I. for Good—And Also Protect the World From Its Threats. *The Chronicle of Philanthropy*. [https://www.philanthropy.com/article/foundations-look-to-advance-ai-for-good-and-also-protect-the-world-from-its-threats?sra=true&cid=gen\\_sign\\_in](https://www.philanthropy.com/article/foundations-look-to-advance-ai-for-good-and-also-protect-the-world-from-its-threats?sra=true&cid=gen_sign_in)

- Dhar, V. (2023, May 25). *Ethics in the Age of Generative AI*. LinkedIn Learning. <https://www.linkedin.com/learning/ethics-in-the-age-of-generative-ai/generative-ai-and-ethics-the-urgency-of-now>
- Dhar, V., & Firth-Butterfield, K. (2021, March 30). Philanthropy Needs to Prepare Itself for a World Powered by Artificial Intelligence. *The Chronicle of Philanthropy*. <https://www.philanthropy.com/article/philanthropy-needs-to-prepare-itself-for-a-world-powered-by-artificial-intelligence>
- Effect, G. (2022, September 19). *Homepage – girleffect.org*. *Girleffect.org – Unlocking the Power of Girls; My WordPress*. <https://girleffect.org/>
- Ferreira, C., Natalucci, F., Singh, R., & Suntheim, F. (2021, May 13). *How Strengthening Standards for Data and Disclosure Can Make for a Greener Future*. IMF. <https://www.imf.org/en/Blogs/Articles/2021/05/13/how-strengthening-standards-for-data-and-disclosure-can-make-for-a-greener-future>
- Firth-Butterfield, K., Toplic, L., Anthony, A., & Reid, E. (2022, March 17). *Without Universal AI Literacy, AI Will Fail Us*. World Economic Forum. <https://www.weforum.org/agenda/2022/03/without-universal-ai-literacy-ai-will-fail-us/>
- Fleming, K., Abril, A. M., & Bradach, J. (2023, January 24). *The Impact of Large, Unrestricted Grants on Nonprofits: A Five-Year View*. The Center for Effective Philanthropy. <https://cep.org/the-impact-of-large-unrestricted-grants-on-nonprofits-a-five-year-view/>
- Global Inequality Rises Again* (2023, January 15). International Finance. <https://internationalfinance.com/global-inequality-rises-again/>
- International Telecommunication Union (2023, September 12). *Press Release*. ITU. <https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by-2030.aspx>
- International Women's Day: New Factsheet Highlights Gender Disparities in Innovation and Technology* (2023, March 7). <https://www.unesco.org/en/articles/international-womens-day-new-factsheet-highlights-gender-disparities-innovation-and-technology>
- Kassatly, A. (2018, May 8). How Philanthropy Infrastructure Can Promote Evidence-Based Giving. *Alliance Magazine*. <https://www.alliancemagazine.org/analysis/philanthropy-infrastructure-can-promote-evidence-based-giving/>
- Lester, A. (2014, October 8). *Why Are Foundation Leaders Slow to Adopt New Technology?* <https://www.fluxx.io/blog/why-are-philanthropy-leaders-slow-to-adopt-new-technology>
- Lynch, M. (2018, December 12). *10 Things to Know about Poverty and Technology*. The Tech Advocate. <https://www.thetechadvocate.org/10-things-to-know-about-poverty-and-technology/>
- Nielsen, R. C. (2014). *A World That Counts*. Data Revolution Group. <https://www.undatarevolution.org/report/>
- Nonprofit Trends Report* (2022). Salesforce. <https://www.salesforce.com/news/stories/nonprofit-statistics-trends-2022/>
- Open Climate Data* (2023, April 25). Creative Commons. <https://creativecommons.org/about/program-areas/open-science/open-climate-data/>
- Per Scholas Expands Tech Training to Houston* (2023, July 17). Per Scholas. <https://perscholas.org/news/per-scholas-expands-tech-training-to-houston/>
- Roadmap for Funders* (2020). NTEN; NetHope; Technology Association of Grantmakers; TechSoup. [https://www.tagtech.org/wp-content/uploads/2023/11/Digital\\_Infrastructure-Funde.pdf](https://www.tagtech.org/wp-content/uploads/2023/11/Digital_Infrastructure-Funde.pdf)
- Team4Tech* (2023, April 3). Team4Tech. <https://team4tech.org/>
- The Future Society* (2016, November 15). The Future Society. <http://thefuturesociety.org>
- United Nations Educational, Scientific, and Cultural Organization (2022). *Recommendation on the Ethics of Artificial Intelligence*. <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
- Warren, S., Cheikosman, E., & Fazelpour, S. (2021). *Pathways to Digital Justice*. World Economic Forum. <https://www.weforum.org/whitepapers/pathways-to-digital-justice/>
- World Economic Forum Launches the Centre for Trustworthy Technology* (2023, January 18). World Economic Forum. <https://www.weforum.org/press/2023/01/world-economic-forum-launches-the-centre-for-trustworthy-technology/>

# ALTRUISTIC COLLECTIVE INTELLIGENCE FOR THE BETTERMENT OF ARTIFICIAL INTELLIGENCE

*Thomas Maillart, Lucia Gomez, Mohanty Sharada,  
Dipam Chakraborty and Sneha Nanavati*

## 1 Introduction

Despite its transformative economic and social outcomes, artificial intelligence (AI) is faced with several operational, legal, and ethical challenges (Ntoutsis et al., 2020), mainly associated with algorithm robustness (Dietterich, 2019), explainability (Confalonieri et al., 2021), and biases (Osoba et al., 2017). However, as reported by Percia David et al. (2023), AI seems to be still largely in its infancy and may be decades away from becoming a mature technology. Despite concerns and room for improvement, the opportunity for a bright AI future exists. Navigating toward the betterment of AI implies engaging in a critical reflection that builds on the lessons learned and experience of computer revolutions from their start in the late 1960s (Levy, 2010). While computer systems development was mainly driven by closed source and intellectual property (IP)-protected software at the time, an open-source community developed early on. This open-source movement was initially seen as fringe and utopist, as it introduced a new form of IP, which would primarily require sharing software code and crediting code authors<sup>1</sup> in a non-exclusive non-rival economic regime, instead of extracting profit from exclusive, rival goods (Tirole & Lerner, 2002). The open-source movement led to the disruption of large parts of the software industry, from operating systems (Moody, 2009), to the Internet (Zittrain, 2009), to cryptography (Landau, 2022), to the World Wide Web (Benkler, 2011), and eventually impregnating the world of commercial software (Fitzgerald, 2006). The dominance of open-source software has culminated since Microsoft acquired GitHub, the prime open-source online development platform, in 2018. Reflecting on how past technology development has been well managed and how it has failed in its social outcomes is informative to establish a sustainable development path for AI, under economic, social, and ethical constraints.

Inspired by the open-source software (OSS) movement, we posit that two interacting mechanisms are key to the sustainable betterment of AI. First, collective intelligence (CI), embodied by peer production – composed of task self-selection, peer review, and transparency (Benkler, 2002) – plays the role of democratized control over the development of AI systems, while also boosting innovation, through decentralized, modular, and creatively destructive dynamics (Maillart et al., 2008). Second, the open-source movement has allowed a strong culture of altruism, born

from the imperative of collective action to overcome significant challenges in complex environments (Ostrom, 1990). It also grew largely with the recognition of software programming as a form of art (Bonaccorsi & Rossi, 2004), whose practice stems largely from intrinsic motivation (Ryan & Deci, 2000; von Krogh et al., 2012).

In this chapter, we critically reflect on the development of AI by drawing from the concrete example of Alcrowd,<sup>2</sup> an AI startup that has heavily bet on peer production, intrinsic motivation of a sense of community, with emphasis on fun and a strong drive for social good. Namely, Alcrowd proposes challenges offered by its paying customers or for free by not-for-profit organizations for its 68k+ community members to tackle. During challenges, contributors can submit and test for performance as many AI models as they wish. The winners receive a monetary or in-kind prize at the end of the competition period. Participants may also be offered to be co-authors of research papers to be posted to, e.g., *arXiv.org* or submitted to a computer science conference, such as NeurIPS, the famous AI scientific conference (competition track). If accepted, the lead authors are invited to attend the conference with all fees covered. Alcrowd is therefore walking a thin line combining extrinsic motivation and competition (i.e., competition for a prize and potential royalty proceedings from paying use of AI models down the road), intrinsic motivation (i.e., sense of achievement in a fun and social online environment), cooperation, and a sense of purpose. The latter sense of purpose is powerful with challenges organized free of charge for not-for-profit organizations (e.g., International Telecommunication Union – AI for Good or the United Nations). By investigating the Alcrowd community as a heterogeneous set of engaged individuals fueled by various incentives, we investigate how the selfless philanthropic side of people acting collectively, through their cooperation and competition at once, brings more ethical value to AI, enhancing cooperation, utility, robustness, and transparency.

## **2 Background**

To envision how altruistic collective intelligence could benefit from the development of AI, we consider how the open-source movement has durably shaped the development of software, which is the most important precursor technology of AI together with abundant data and the development of specialized hardware, such as graphical processing units (GPUs) chips.

### ***2.1 How altruistic collective intelligence has shaped the digital revolution***

The open-source “*hacker*” movement (Levy, 2010) and the peer production approaches (Benkler, 2002), characterized by the collaborative efforts of individuals working voluntarily toward a common goal, have significantly contributed through history to alleviating operational, organizational, and ethical challenges in software development. Operationally, peer production facilitated the rapid identification and resolution of bugs and the enhancement of software features through the diverse expertise of contributors, leading to more robust, reliable, and adaptive software (Raymond, 1999; Maillart et al., 2017). Organizationally, it democratized key development processes, breaking down barriers to entry and allowing for a more inclusive group of participants, which in turn fostered innovation and accelerated development cycles (Bosu et al., 2017). One overarching example of the influence of the open-source movement on the development of software development practices is GIT (Loeliger & McCullough, 2012), a distributed control version system invented by Linus Torvald in 2005, that has become the technical backend of widely used social coding platforms, like GitHub and GitLab.

Ethically, the open-source approach has encouraged transparency and accountability, as the open review process ensured that code was scrutinized by a broad community. This openness

helped in identifying and mitigating biases or unethical uses of software, aligning development practices more closely with societal norms and values (Coleman, 2012; von Krogh et al., 2012).

The open-source movement has indelibly shaped the software and hardware development landscape, embedding principles of transparency, collaboration, and accessibility deeply within the fabric of technological innovation. It marks a pivotal shift from the exclusivity of proprietary systems toward a more democratized approach to technology creation and dissemination. The genesis of the open-source movement can be linked to the collaborative ethos among early computer scientists and hobbyists who believed in sharing software openly as a means to foster innovation and solve problems more efficiently. In his book *Hackers: Heroes of the Computer Revolution* (2010) on the history of the hacker culture, Steven Levy highlighted the commitments of this burgeoning community to openness, information, and sharing. This period laid the groundwork for developing major open-source projects that would later revolutionize the technology sector.

The influence of open-source principles became more pronounced with the development of foundational operating systems, notably the Linux kernel, spearheaded by Linus Torvalds in 1991. Moody's exploration (2009) of open-source operating systems illustrates how this model disrupted the traditional software development paradigm, enabling a global community of developers to contribute to and improve existing codebases. This collaborative approach not only accelerated innovation but also ensured that software could be modular and adaptive (Maillart et al., 2008), as such, more secure, reliable, and adaptable to the needs of diverse users.

The expansion of the Internet and the World Wide Web further exemplified the power of open-source methodologies. Jonathan Zittrain (2009) and Barbara van Schewick (2012) have discussed how Internet's open architecture facilitated an unprecedented level of innovation and creativity, allowing individuals and small teams to create impactful technologies without the need for substantial resources. The invention of the World Wide Web by Tim Berners-Lee at CERN, who made this technology available on a royalty-free basis, epitomizes the ethos of the open-source movement, democratizing access to information and enabling the explosive growth of online content and services.

Cryptography, as explored by Susan Landau (2022), is another area where open-source principles have been instrumental. The move toward open cryptographic standards and the public sharing of encryption algorithms have significantly enhanced security and privacy in the digital age, underscoring the movement's role in building trust and safeguarding freedoms online.

The culmination of the open-source movement's integration into the commercial sector was symbolized by Microsoft's acquisition of GitHub in 2018. Once seen as antithetical to the open-source *ethos*, major corporations have now largely embraced open development practices, recognizing the value of community-driven innovation (yet without relinquishing proprietary, closed source code). The GitHub acquisition by Microsoft, as discussed by Brian Fitzgerald (2006), marks a significant acknowledgment of the impact of the open-source model on commercial software development, highlighting a shift toward more open, collaborative, and transparent practices associated with increased competitiveness for firms (Nagle, 2018) and even for nations (Nagle, 2019).

## 2.2 *Altruistic collective intelligence is a form of philanthropy*

At the root of the open-source movement is collective intelligence (CI). CI refers to the shared knowledge, expertise, and problem-solving capabilities of a group or community of individuals (Malone, 2019). It is the idea that the collective wisdom of a group and their collective decision-making can bring additional performance to that of any single member within the group: CI emerges when people collaborate, share information, and pool their insights and abilities to

tackle complex problems, make decisions, or create innovative solutions. Deeper, CI is best predicted by social interactions (Kim et al., 2012) and the capacity to sense the emotional states of others through non-verbal social cues (Woolley et al., 2010). The CI concept has become particularly relevant in the digital age, where technology and connectivity enable large and diverse groups to collaborate and generate insights and solutions collectively (Benkler, 2011).

When predominantly relying on intrinsic motivation (Ryan & Deci, 2000), CI can be seen as a form of philanthropy, albeit in a non-traditional sense. While philanthropy typically involves the individual donation of financial resources or volunteering time to support charitable causes directly through the provision of individual skills, CI offers a different kind of contribution, which is collective, integrative, and, most importantly, involves the mobilization of empathy (Woolley et al., 2010). Hence, the philanthropic contribution lies primarily in how it proceeds and delivers value collectively, by leveraging people's inner social capabilities (Dunbar, 1998). *In fine*, it enables collective action to overcome significant challenges and strive in adverse environments (Ostrom, 1990).

CI involves harnessing the wisdom, knowledge, and expertise of a diverse group of individuals to address complex problems, make informed decisions, or create innovative solutions (Hong & Page, 2004). In essence, CI is a form of giving back to society through the productive collision and integration of intellectual capital (Engel & Malone, 2018). By collaborating and pooling their collective knowledge, people can collectively benefit others by solving challenges, advancing research, or improving decision-making in areas such as science, technology, and governance (Fink, 2018). CI is even thought to provide answers to how humankind should consider tackling global catastrophic risks (Yang & Sandberg, 2023). In this sense, the act of contributing individual insights and expertise to collective efforts can be seen as a valuable and altruistic form of philanthropy, one that goes beyond financial donations and embodies the spirit of communal support for the greater good. As for contributing, collaborative giving elicits similar intrinsic rewards, and the “whole is more than the sum” financial contributions (Proulx et al., 2023), while also possibly producing the “whole is more than the sum” (Sornette et al., 2014) through peer production (Benkler, 2002).

### ***2.3 Outstanding challenges in AI and how altruistic collective intelligence can help with a dose of competition***

As we stand on the brink of a fundamental reshaping of many facets of human life by AI, the lessons from the open-source movement are more pertinent than ever. The principles of transparency, collaboration, and ethical responsibility that have driven the open-source movement can, and already largely, serve as guiding lights for the development of AI technologies. By fostering an open AI ecosystem, we can encourage a broad and diverse community of developers, ethicists, and users to contribute their perspectives and expertise, thereby ensuring that AI technologies are not only advanced but are developed in a manner that is socially responsible, inclusive, and aligned with human values.

Moreover, the OSS model can address some of the most pressing concerns in AI development, including biases, transparency, and accountability, in the same way OSS has helped alleviate similar problems for previous information technology developments. By making AI algorithms and datasets publicly available, the community can facilitate scrutiny, peer review, and iterative improvement, ensuring that AI systems are fair, reliable, and understandable. This collaborative approach to AI development has the potential to democratize AI innovation, making it accessible to a wider range of stakeholders and enabling solutions that are tailored to a variety of social,

economic, and environmental challenges. However, in AI development, benchmarks serve as computational Olympic arenas, where algorithms and pipelines compete for improvement toward a progressively optimal solution or fork through radical innovations. These standardized testing grounds for AI evaluation constitute unique tools for combining cooperation and competition, as all developers can observe and learn from each other’s solutions to create the next best one. Potentially serving as one of the major catalyzers of CI within AI, benchmarks are fundamental pieces for a future open and democratic AI landscape.

The journey from the early advocacy for open computing environments to the present-day ubiquity of open-source methodologies underscores a fundamental change in how digital technologies have been developed, distributed, and perceived. The transformative impact of open-source principles across various technological milestones offers a compelling narrative that not only charts the evolution of technology but also sets a precedent for the development of artificial intelligence (AI) in a sustainable, ethical, and inclusive manner.

### 3 Theoretical framework: altruistic collective intelligence contributing to better AI

Altruistic CI is pivotal in advancing and improving AI, as witnessed by the fast advancement of worldwide open-source communities and hubs such as *HuggingFace*,<sup>3</sup> providing new AI tools and evaluation benchmarks daily (such as their *open\_llm\_leaderboard*). It brings together individuals from diverse backgrounds, approaches, and focus areas, each of which contributes unique tools. This diversity of thought helps AI developers and researchers consider a wide range of perspectives and opens the doors for a fruitful peer-supported exchange, leading to more well-rounded and ethically sound AI systems.

With the increasing complexity of AI systems, ethical considerations have become paramount. Altruistic CI fosters discussions and debates around AI ethics, helping to establish guidelines and best practices that ensure AI technologies are developed with fairness, transparency, and accountability in mind. An exemplar case is the safety and ethics AI leaderboard hosted at *HuggingFace*, setting standards developed and continuously improved by an open community for evaluating AI models (*llm-trustworthy-leaderboard*). CI initiatives can offer valuable feedback on AI systems, helping developers identify weaknesses, vulnerabilities, and areas for improvement. This iterative process is crucial for enhancing the robustness and security of AI technologies.

Besides model development and testing, AI systems heavily rely on data. By engaging in altruistic data sharing, data annotation efforts, and collaborative-competitive development of AI algorithms, CI can help improve the quality and diversity of training data to mitigate biases, but also improve the balance in representing minorities and rare data sources, for example. This collective data sharing and pruning efforts, in turn, lead ultimately to AI systems that are more accurate, unbiased, and reliable. In sum, altruistic CI is a source of diverse and rich data, AI models, and ethics testing.

Further, as the existence of AICrowd shows, CI can also contribute to the accelerated development and the betterment of machine learning algorithms and neural networks. Nevertheless, the origins of CI performance in AI algorithm development have remained unclear.

We posit that performance is largely due to collective “*trial-and-error*” by teams. Our hypothesis is that *team size*, *number of submissions*, and incidentally *waiting time between submissions* are important, but the *diversity of submissions* by the same team matters most. The number and diversity of submissions by a team reflect the collective capacity to take risks in accepting that some submissions will not overperform the current benchmark of (i) own team submissions as well as

(ii) submissions by other teams. We contend that this capacity stems from using foraging-like *explore* and *exploit* human search algorithms (Wilson et al., 2014).

As we dive into the data of the *Alcrowd Food Recognition Challenge* (see Section 4), we focus on the structure of *cooperation* and *competition* in communities and how the teams achieve performance during the challenges.

## **4 Alcrowd: an exemplar case study on altruistic CI for AI**

### ***4.1 Alcrowd: democratizing access to AI challenges and harnessing collective intelligence***

Alcrowd is a pioneering platform at the intersection of data sharing, AI, and CI. As a business, Alcrowd operates as a hub that connects problem-solvers, data scientists, and machine learning enthusiasts with organizations seeking innovative solutions to complex AI challenges. Since it started, Alcrowd has hosted over 500 challenges and competitions in a wide range of AI sub-disciplines. Participants from around the world team up, cooperate, and submit their machine learning models to solve real-world problems.

Alcrowd challenges cover diverse domains, from computer vision and natural language processing to robotics and healthcare. The business model thrives on creating a collaborative ecosystem that allows organizations to tap into the CI of a global community of AI enthusiasts and researchers. Through these challenges, Alcrowd facilitates knowledge sharing, fosters innovation, and accelerates AI advancements while offering organizations access to cutting-edge solutions and talents. This unique business model positions Alcrowd as a catalyst for collective problem-solving in the machine learning domain.

Alcrowd faces competition primarily from platforms like Kaggle<sup>4</sup> – known for its broad range of data science and machine learning competitions, Topcoder<sup>5</sup> – which connects businesses with global developers and data scientists, DrivenData<sup>6</sup> – specializing in data science challenges with social impact, CrowdAI – focusing on computer vision competitions,<sup>7</sup> or Zindi – catering to African-specific AI challenges.<sup>8</sup> These platforms vary in their focus, community, and types of challenges they offer, creating a dynamic landscape for AI competitions and CI in the field.

Alcrowd sets itself apart from competitors by emphasizing fostering a collaborative AI community and diverse problem domains. Unlike competitors, Alcrowd encourages open knowledge sharing and collaborative learning, creating a strong sense of CI among participants. Its challenges cover a wide array of AI fields, addressing real-world problems faced by organizations and making it practical and industry-oriented. Alcrowd also actively collaborates with academic institutions and research organizations, promoting and leading cutting-edge AI research that emerges from the community breakthrough offered solutions. These aspects distinguish Alcrowd as a platform not only for innovation but also for nurturing a culture of open learning and sharing, and collective problem-solving in the AI community, making it unique in the field of AI challenge business.

### ***4.2 The Alcrowd Food Recognition Challenge***

In the landscape of modern AI-driven challenges, the *Alcrowd Food Recognition Challenge* stands out as a compelling endeavor<sup>9</sup> with implications spanning different sustainability matters such as SDGs 3 (good health and well-being) or 2 (zero hunger). This challenge revolves around the intricate task of recognizing food items from images, a capability that could empower individuals to effortlessly monitor their dietary habits by merely photographing their meals, providing



a powerful tool for personal health management and nutrition tracking (Mohanty et al., 2022). Beyond personal applications, the challenge’s focus on food recognition holds substantial medical relevance, addressing a long-standing methodological need in research studies. Traditionally, such studies relied on low-scale and imprecise food frequency questionnaires. This challenge harnesses the power of CI and AI to provide improved and scalable detection solutions.

A core element of CI in this challenge is the collaborative annotation process. The dataset used in this competition is a collection of images and a rich repository meticulously annotated with segmentation, class-belonging, and weight/volume estimation of individual food items.<sup>10</sup> This annotation effort harnesses the collective wisdom of contributors, ensuring the quality of the dataset and enhancing its utility (Mohanty et al., 2022). The Food Recognition Challenge is by no means unique in that regard. Most recent advances in AI have been possible through highly annotated datasets and evaluation benchmarks, using some form of crowdsourcing for collaborative building and benefiting from a standardized and open ground for model evaluation and optimization. Cases such as the transformative ImageNet challenge illustrate the power of CI for AI, an iconic benchmark competition that has significantly advanced image recognition datasets, research, and AI solutions widely used today, such as the AlexNet, VGG, and ResNet (Deng et al., 2009).

Beyond annotation, the Food Recognition Challenge encourages the development of new machine learning models. Participants are invited to submit their innovative algorithms and approaches, creating a forum for collective problem-solving. By pooling together diverse talents and perspectives in both the annotation and AI development parts, the Food Recognition Challenge exemplifies the principles of altruistic CI. The design of the challenge is an example of best practices in the pursuit of solutions for good and for the resolution of intricate problems.

The inner mechanisms of how performance emerges from *coopetition*, i.e., a subtle equilibrium between collaboration and competition, have remained unclear. While coopetition enjoys some popularity in the context of industrial organization (Bouncken et al., 2015), it is also applicable to collaborative communities (Gulley & Lakhani, 2010). Here, we investigate the fine-grained origins of performance in developing AI algorithms in the context of the Food Recognition Challenge.

### 4.3 Mapping collective intelligence dynamics of the AICrowd Food Recognition Challenge

#### 4.3.1 Data and method

To understand how team dynamics evolved throughout the Food Recognition challenge (2021–2022), we studied anonymized data on timestamped<sup>11</sup> solution submissions annotated by:

- 1 Round type (challenge versus benchmark)<sup>12</sup>;
- 2 Round identification (ID) number (1–4, 1–2)<sup>13</sup>;
- 3 Anonymized participant ID;
- 4 Anonymized team ID;
- 5 Solution precision score (proportion of true positive predictions among all positive predictions); and
- 6 Submission latency (in days, zero being the start of each challenge round).

To comprehensively visualize and analyze team activity across the entire challenge (challenge rounds and benchmark rounds), we structured this data in a bipartite network<sup>14</sup> (see Figure 21.1) capturing the relationships between teams (white dots) and submissions (dark dots), weighted by

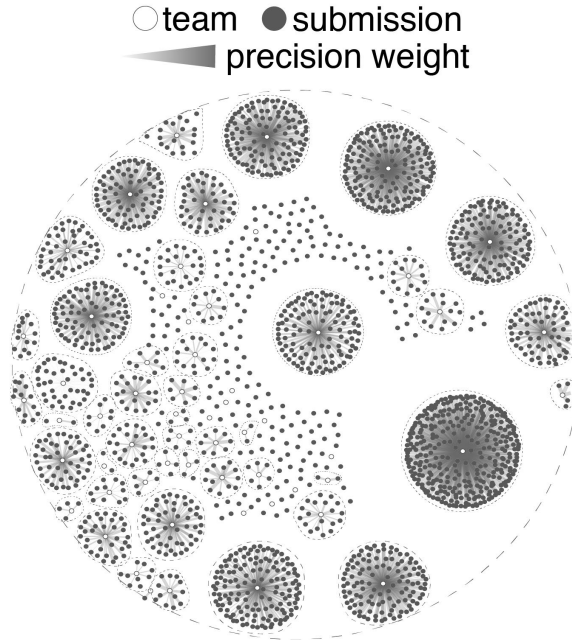


Figure 21.1 Bipartite network of relationships between teams and submissions as weighted by precision scoring.

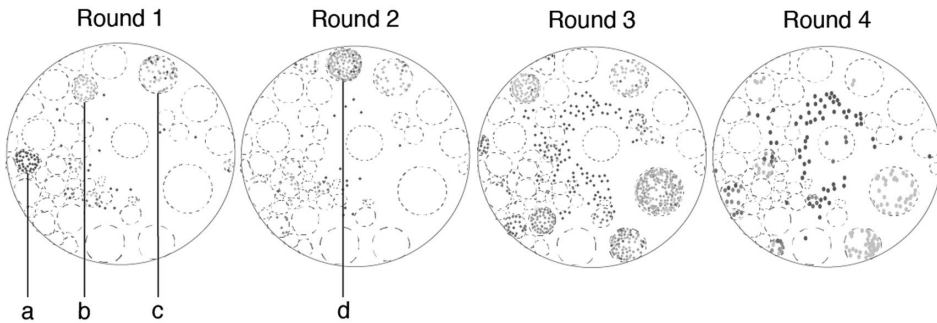
the precision of the solution (gray lines), as systematically evaluated by AICrowd. The temporal evolution of the team submission was obtained by slicing the core bipartite network per round (one to four challenge rounds, one to two benchmark rounds; see Figure 21.2).

Upon visual inspection (see Figure 21.1), we find evidence of activity heterogeneity: some teams displayed strong engagement through their submission activity (i.e., a high number of submissions – many dark dots in dense circular clusters). Conversely, other teams sparsely contributed, with a low number of submissions (sparse blue dots, scattered). Additionally, some teams display an additional layer of heterogeneity in their activity patterns: although all circular cluster teams submitted enough to create a dense cluster, some teams submitted significantly more solutions and benchmarked their performance more frequently, while others submitted less solutions (fewer dark dots; smaller clusters).

We propose that team submission density in the network could be a behavioral indicator for potential hubs of CI emergence through *coopetition* in the benchmarking of submissions open to the entire community. In other words, densely clustered teams cooperate between members and compete against themselves and other team submission. The entire network thus interacts (hub), engaging in a CI-generating behavior altogether. Real-time standardized evaluation (benchmarking) could be fundamental for indicating the status of the collective, accessing knowledge and pipelines developed by others, and, therefore, fostering CI coopetition. As all community members are informed in real time about the performance and algorithms of others, cooperation and competition fuel the emergence of innovative and ever-improving solutions.

Next, we considered the teams' submission dynamics during the challenge: Figure 21.2A shows how performance arises for each team and across the four rounds of the challenge.

## A. Challenge Phase



## B. Benchmark Phase

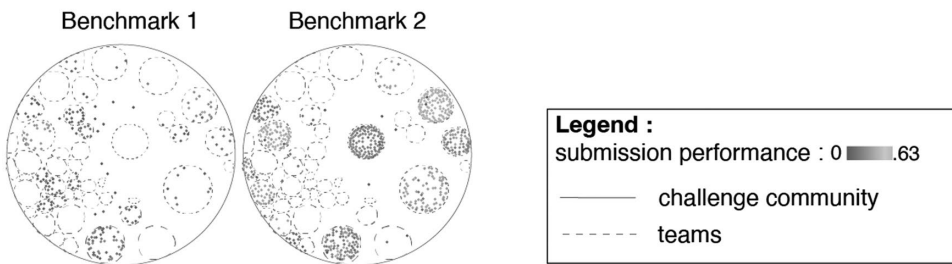


Figure 21.2 Network slice by submission rounds, in challenge and benchmark phases.

Further visual inspection suggests that teams with high and low numbers of submissions at each round can achieve both good and bad performances (high heterogeneity – visualized by color gradient). Additionally, the number of submissions per round does not directly indicate overall performance. For example, *team c* that won the Food Recognition Challenge submitted their winning solution in round 2, while they submitted most solutions during rounds 1 and 3 (61 submissions in round 1, 46 in round 2, 60 in round 3, and 2 in round 4).

Contrary to other teams, such as *a*, *b*, and *d*, participating in single rounds, *team c* showed a protracted engagement across all rounds. For teams participating in several main rounds (i.e., challenge rounds), an overall increase in both solution precision and number of submissions can be observed in consecutive phases (rounds 3–4 display more submissions compared to rounds 1–2). Additionally, during challenge rounds, the performance of submitted solutions within teams shows high variability: while some teams, such as *a* and *b*, submit solutions with a similar level of precision across trials, submissions by other teams, such as *c* and *d*, show a high degree of variability across trials.

Among these, *team c* provided the best solution to the challenge, with a precision of 0.62 (and recall of 0.88),<sup>15</sup> a high score considering the stringent food identification criteria set by AICrowd for this challenge (c.f., footnote 7 for reading all details at the AICrowd website). *Team d*, participating in rounds 1 and 2 (2 submissions in round 1, 117 in round 2) and displaying a high-performance variability, was the second-best team, with their top solution having a precision of 0.59 (and recall of 0.82). Conversely, the overall submission precision in the *benchmark* phase was visibly lower than during the *challenge* phase.

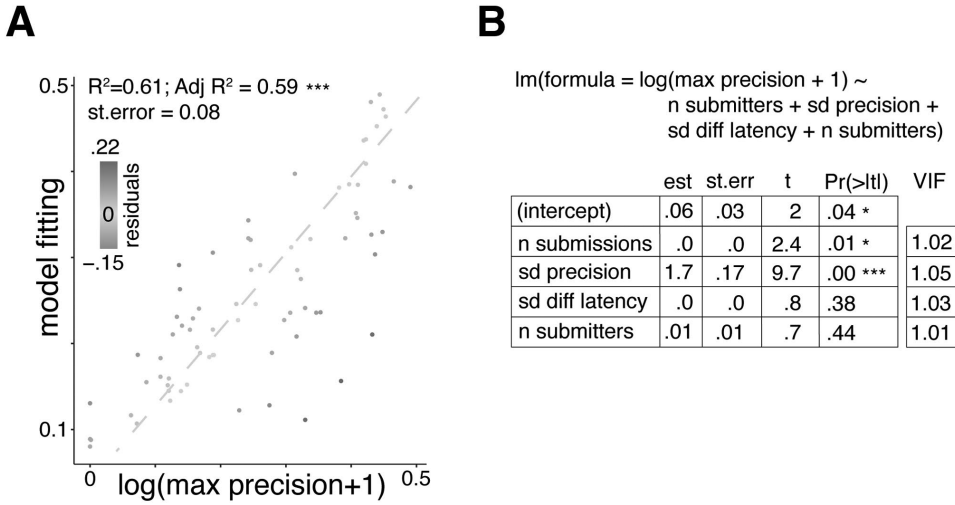


Figure 21.3 Determinants of collective intelligence in AICrowd Food Recognition Challenge.

This was due to the post-challenge re-definition of both the dataset and task requirements (among which, amount of food types to be identified) toward continuous development of the AI solution after an initial successful competition (challenge rounds 1–4). Despite the decrease in performance (caused by performance metric redefinitions from challenge to benchmark phases), an improvement in submission precision is still observable between phases 1 and 2 of the benchmark. This suggests that award expectation, which was present during the challenge phase but absent during benchmarking, is not the sole driver of continued cooperation.

To systematically understand the recipe of successful CI dynamics, we modeled *maximum precision* (dependent variable) by *team* and *round* across the competition, using a standard linear regression model (see Figure 21.3A–B). As regressors (independent variables), we considered:

- 1 The *number of submissions* by team and round (proxy for how active a team is in testing solutions);
- 2 The *number of submitters per team* (indicating how many different participants are active submitters testing solutions in parallel);
- 3 The *standard deviation of the precision* across submissions (proxy for their trial-and-error progress variation and thus appetite for risky trial and fail);
- 4 And the *standard deviation of the waiting time between two submissions* (metric of how much time it took to develop a given submission).

These independent variables were individually examined for their contributions to the maximum precision. Prior to modeling, we ensured that these predictors did not display collinearity (*VIF* score, Figure 21.3B).

#### 4.3.2 Results

We find that the independent variables of interest here explain a considerable amount of the maximum performance per team and round, with an *adjusted-R<sup>2</sup>* of 0.59 (*non-adjusted-R<sup>2</sup>* of

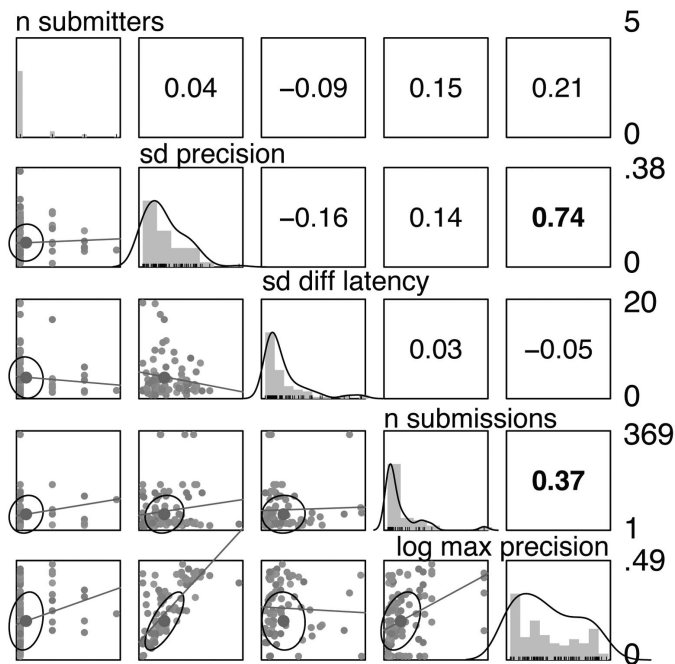


Figure 21.4 Detail on the relation between the variables used for modeling.

0.61), a residual standard error of 0.08, and statistical significance on modeling ( $p < 0.05$ ) (see Figure 21.3A).<sup>16</sup> Among predictor variables, the number and the heterogeneity of submissions were significant predictors, in particular heterogeneity (standard deviation of solution precision;  $p < 0.001$ ) (see Figure 21.3B). Importantly, as found for the best and second-best performing teams *c* and *d*, risky “*trial-and-error*” exploration of solutions with variable accuracies is an important contributor to successful CI (see Appendix 21.A2). Although this model is not able to explain all observations (high residuals associated with outlier submission: exceptionally performant and under-performant ones), it does capture the overlying trend behind CI success, fueled by a mighty predictive power of the standard deviation of submission precision (see Figures 21.3B and 21.4).

Our results indicate that the more a team tries heterogeneous solutions associated with more performance outcome risks, the more likely it is to obtain a higher maximum precision across trials. Related, however not correlated, the number of submissions per team significantly indicated overall success. Considering the volatility of submission latency, we find that the variability in the time taken for developing solutions can result in an overall successful or unsuccessful competition. Overall, indicators of *coopetition* support CI success.

## 5 Conclusion

The AICrowd Food Recognition Challenge analyses reveal non-trivial dynamics of collective intelligence (CI) in action. The varying degrees of team engagement and the performance differences illustrate the intricate balance between competition and collaboration in driving AI advancements. Notably, the significant predictors of success – *variability in submission performance* and *active*

*engagement in solution testing* – underscore the efficacy of a “*try-and-error*” sub-mechanism in the broader CI universe. These findings not only highlight the importance of seeking diversity, and hence collective intelligence, to achieve higher precision but also reflect the broader mission of Alcrowd: to democratize AI challenges and harness CI for the collective good.

The mission of Alcrowd serves as a Petri dish for the potential transformation of the AI landscape fueled by the philanthropic contribution of AI developers cooperating and competing at once. By fostering an environment where individuals from varied backgrounds contribute toward common goals, Alcrowd exemplifies how CI can lead to innovative solutions that might emerge neither from isolated nor homogeneous efforts. This approach, rooted in peer production and transparency, leverages the intrinsic motivation of participants, blending competition with cooperation for societal benefit. This, in turn, enables the strength of collective problem-solving through which AI practitioners can tap into a global pool of expertise for solving the intricate types of challenges only this technology has been able to address. The collaborative approach of CI accelerates AI advancements, helps overcome challenges more effectively and democratically, educates the public about AI and its implications, and fosters discussions of how it should be the AI for all.

The implications of such a model extend beyond Alcrowd to inform broader discussions on AI and philanthropy. The altruistic underpinnings of CI, as demonstrated through challenges organized by Alcrowd, offer a blueprint for advancing AI that prioritizes ethical considerations, inclusivity, and fairness. This collaborative innovation model, where success is derived from collective effort and diversity of thought, not only accelerates technological advancements but also ensures that these advancements are aligned with societal needs and ethical standards. As AI continues to evolve, integrating principles of CI and philanthropy into its development can help mitigate biases, enhance transparency, and ensure that AI serves as a force for good, reflecting a shared commitment to improving the human condition.

## Notes

- 1 A wide spectrum of open-source licenses have existed and co-evolved over the years. See Carver (2005) for a full review.
- 2 <https://www.aicrowd.com>
- 3 <https://huggingface.co/>
- 4 <https://www.kaggle.com/>
- 5 <https://www.topcoder.com/>
- 6 <https://www.drivendata.org/>
- 7 <https://www.crowdai.com/>
- 8 <https://zindi.africa/>
- 9 <https://www.aicrowd.com/challenges/food-recognition-challenge> and <https://www.aicrowd.com/challenges/food-recognition-benchmark-2022>
- 10 Segmentation = image annotation delimiting the perimeter of regions of interest (ROI) (food items for training); Class-belonging = Region of Interest (ROI) annotation stating the true food category; weight/volume estimation is self-explanatory.
- 11 Each submitted solution is time-stamped so that the challenge can be analyzed with temporal resolution.
- 12 Challenge submissions are registered for the competition phase toward a prized solution whereas benchmark submissions are post-challenge submissions for the permanent improvement of the algorithm outside the competition.
- 13 IDs 1–4 belong to phases 1–4 of challenge rounds; IDs 1 and 2 belong to corresponding phases of benchmark rounds.
- 14 Data structure representing two different types of samples (nodes of the network), in this case teams (white) and submissions (dark gray), connected by the performance score (precision) of the submitted solution (gray lines whose thickness, weight, is proportional to its precision).

- 15 Precision : proportion of true positive predictions among all positive predictions; Recall : Proportion of true positive predictions among all positive and negative predictions.
- 16 *adjusted-R<sup>2</sup>*: Measure of model goodness of fit indicating the proportion of the variance in the independent variable explained by the model when using selected dependent variables for regression. Ranging from 0 to 1, it is calculated as the ratio of the explained sum of squares to the total sum of squares. Adjusted means that this R<sup>2</sup> considers the number of independent variables used and penalizes the addition of unnecessary variables by adjusting for degrees of freedom. *non-adjusted-R<sup>2</sup>* does not penalize the addition of unnecessary variables by adjusting for degrees of freedom. The residual standard error is used to measure how well a regression model fits a dataset. In simple terms, it measures the standard deviation of the residuals in a regression model. The *p value* helps determine that the null hypothesis cannot be rejected, i.e., that the model used to represent the data cannot be ruled out statistically.

## References

- Benkler, Y. (2002). Coase's Penguin, or Linux and "The nature of the firm." *The Yale Law Journal*, 112(3), 369+. <https://doi.org/10.2307/1562247>
- Benkler, Y. (2011). *The Penguin and the Leviathan: How Cooperation Triumphs over Self-Interest* (1st ed.). Crown Business. <http://www.worldcat.org/isbn/0385525761>
- Bonaccorsi, A., & Rossi, C. (2004). Altruistic individuals, selfish firms? The structure of motivation in Open Source software. *First Monday*, 9(1). <https://doi.org/10.5210/fm.v0i0.1476>
- Bosu, A., Carver, J. C., Bird, C., Orbeck, J., & Chockley, C. (2017). Process aspects and social dynamics of contemporary code review: Insights from open source development and industrial practice at Microsoft. *IEEE Transactions on Software Engineering*, 43(1), 56–75. <https://doi.org/10.1109/TSE.2016.2576451>
- Bouncken, R. B., Gast, J., Kraus, S., & Bogers, M. (2015). Coopetition: A systematic review, synthesis, and future research directions. *Review of Managerial Science*, 9(3), 577–601. <https://doi.org/10.1007/s11846-015-0168-6>
- Carver, B. W. (2005). Share and share alike: Understanding and enforcing open source and free software licenses part I: Law and technology: Subpart IV: Business law: Section A: Notes. *Berkeley Technology Law Journal*, 20(1: Annual Review), 443–484.
- Coleman, E. G. (2012). Coding freedom: The ethics and aesthetics of hacking. In *Coding Freedom*. Princeton University Press. <https://doi.org/10.1515/9781400845293>
- Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1), e1391. <https://doi.org/10.1002/widm.1391>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Dietterich, T. G. (2019). Robust artificial intelligence and robust human organizations. *Frontiers of Computer Science*, 13(1), 1–3. <https://doi.org/10.1007/s11704-018-8900-4>
- Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5), 178–190. [https://doi.org/10.1002/\(SICI\)1520-6505\(1998\)6:5<178::AID-EVAN5>3.0.CO;2-8](https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8)
- Engel, D., & Malone, T. W. (2018). Integrated information as a metric for group interaction. *PLoS One*, 13(10), e0205335. <https://doi.org/10.1371/journal.pone.0205335>
- Fink, A. (2018). Bigger data, less wisdom: The need for more inclusive collective intelligence in social service provision. *AI & SOCIETY*, 33(1), 61–70. <https://doi.org/10.1007/s00146-017-0719-2>
- Fitzgerald, B. (2006). The transformation of open source software. *MIS Quarterly*, 30(3), 587–598. <https://doi.org/10.2307/25148740>
- Gulley, N., & Lakhani, K. R. (2010). *The Determinants of Individual Performance and Collective Value in Private-Collective Software Innovation* (SSRN Scholarly Paper 1550352). <https://doi.org/10.2139/ssrn.1550352>
- Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46), 16385–16389. <https://doi.org/10.1073/pnas.0403723101>
- Kim, T., McFee, E., Olguin, D. O., Waber, B., & Pentland, A. "Sandy." (2012). Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior*, 33(3), 412–427. <https://doi.org/10.1002/job.1776>

- Landau, S. (2022). The development of a crypto policy community: Diffie–Hellman’s impact on public policy. In *Democratizing Cryptography: The Work of Whitfield Diffie and Martin Hellman* (1st ed., Vol. 42, pp. 213–256). Association for Computing Machinery. <https://doi.org/10.1145/3549993.3550002>
- Levy, S. (2010). *Hackers: Heroes of the Computer Revolution* (O’Reilly, Ed.; 3rd ed.). O’Reilly. <http://www.worldcat.org/isbn/0141000511>
- Loeliger, J., & McCullough, M. (2012). *Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development*. O’Reilly Media, Inc.
- Maillart, T., Sornette, D., Spaeth, S., & von Krogh, G. (2008). Empirical tests of Zipf’s law mechanism in open source Linux distribution. *Physical Review Letters*, *101*(21), 218701. <https://doi.org/10.1103/PhysRevLett.101.218701>
- Maillart, T., Zhao, M., Grossklags, J., & Chuang, J. (2017). Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity*, *3*(2), 81–90. <https://doi.org/10.1093/cybersec/tyx008>
- Malone, T. W. (2019). *Superminds: How Hyperconnectivity Is Changing the Way We Solve Problems*. One-world Publications.
- Mohanty, S. P., Singhal, G., Scuccimarra, E. A., Kebaili, D., Héritier, H., Boulanger, V., & Salathé, M. (2022). The food recognition benchmark: Using deep learning to recognize food in images. *Frontiers in Nutrition*, *9*. <https://www.frontiersin.org/articles/10.3389/fnut.2022.875143>
- Moody, G. (2009). *Rebel Code: Linux and the Open Source Revolution*. Hachette UK.
- Nagle, F. (2018). Learning by contributing: Gaining competitive advantage through contribution to crowd-sourced public goods. *Organization Science*, *29*(4), 569–587. <https://doi.org/10.1287/orsc.2018.1202>
- Nagle, F. (2019). *Government Technology Policy, Social Value, and National Competitiveness* (SSRN Scholarly Paper 3355486). <https://doi.org/10.2139/ssrn.3355486>
- Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., & Staab, S. (2020). Bias in data-driven artificial intelligence systems – An introductory survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1356. <https://doi.org/10.1002/widm.1356>
- Osoba, O. A., IV, W. W., & Welsler, W. (2017). *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Rand Corporation.
- Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action (Political Economy of Institutions and Decisions)*. Cambridge University Press. <http://www.worldcat.org/isbn/0521405998>
- Percia David, D., Maréchal, L., Lacube, W., Gillard, S., Tsesmelis, M., Maillart, T., & Mermoud, A. (2023). Measuring security development in information technologies: A scientometric framework using arXiv e-prints. *Technological Forecasting and Social Change*, *188*, 122316. <https://doi.org/10.1016/j.techfore.2023.122316>
- Proulx, J. D. E., Akin, L. B., & Barasch, A. (2023). Let’s give together: Can collaborative giving boost generosity? *Nonprofit and Voluntary Sector Quarterly*, *52*(1), 50–74. <https://doi.org/10.1177/08997640221074699>
- Raymond, E. (1999). The cathedral and the bazaar. *Knowledge, Technology & Policy*, *12*(3), 23–49. <https://doi.org/10.1007/s12130-999-1026-0>
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. <https://doi.org/10.1006/ceps.1999.1020>
- Sornette, D., Maillart, T., & Ghezzi, G. (2014). How much is the whole really more than the sum of its parts?  $1 \boxplus 1 = 2.5$ : Superlinear productivity in collective group actions. *PLoS One*, *9*(8), e103023.
- Tirole, J., & Lerner, J. (2002). Some simple economics of open source. *Journal of Industrial Economics*, *50*(2), 197–234.
- van Schewick, B. (2012). *Internet Architecture and Innovation*. The MIT Press. <http://www.worldcat.org/isbn/026251804X>
- von Krogh, G., Haefliger, S., Spaeth, S., & Wallin, M. W. (2012). Carrots and rainbows: motivation and social practice in open source software development. *MIS Quarterly*, *36*(2), 649–676. <https://doi.org/10.2307/41703471>
- Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*, 2074–2081. <https://doi.org/10.1037/a0038199>



- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004), 686–688. <https://doi.org/10.1126/science.1193147>
- Yang, V. C., & Sandberg, A. (2023). *Collective Intelligence as Infrastructure for Reducing Broad Global Catastrophic Risks*. <https://doi.org/10.25740/mf606ht6373>
- Zittrain, J. (2009). *The Future of the Internet—And How to Stop It*. Yale University Press. <http://www.worldcat.org/isbn/0300151241>

## Appendix 21.1

### A1. Food Recognition Challenge Case Study

**Purpose and Intention:** The Food Recognition Challenge was established to address a common yet complex problem in nutritional informatics – accurately identifying food from images. It intends to harness deep learning capabilities to develop tools that assist users, ranging from individuals tracking their dietary intake to medical professionals conducting nutritional studies.

**Problem Statement:** Despite advancements in computer vision, food recognition from images remains an intricate task due to the diversity of food appearance, presentation, and context. The challenge’s problem statement revolves around creating models that can robustly identify and analyze food items in varying conditions, pushing the limits of current technology.

**Motivation:** The initiative is driven by the need for precise food-tracking mechanisms and the desire to foster community engagement in solving this problem. The challenge focuses on technological innovation and aims to improve public health outcomes and support medical research.

**Impact of the Challenge:** Over successive iterations, the challenge has made significant impacts by (i) providing an open, evolving benchmark for food recognition, encouraging ongoing participation and development, (ii) releasing annotated datasets to the public, facilitating research and application in real-world scenarios, and (iii) creating a platform for developers and researchers to collaborate and compete, spurring advancements in the field.

**Solution and Contribution:** Participants contribute solutions that utilize a novel dataset from the MyFoodRepo app, which is continuously updated with images and annotations for segmentation, classification, and weight/volume estimation. These contributions have led to improved models capable of detecting individual food items and understanding dietary patterns.

**Challenge Cycle:** The challenge operates in multiple rounds, each with specific tasks and datasets that reflect the growing dataset and technological advancements. This cyclical nature ensures the challenge remains relevant and continues to adapt to community feedback and the latest research findings.

**Common Goal:** The shared goal of the Food Recognition Challenge is to provide a high-quality dataset that serves as a foundation for developing effective food recognition algorithms. Unlike the “beautiful” but unrepresentative stock photos found online, these algorithms are expected to work with real-world images. By doing so, the challenge aims to create AI tools that can be widely adopted for personal and medical use.

**Methods:** The Food Recognition Challenge focuses on developing AI models to identify food items in images. These models should be capable of detecting and annotating individual food items with accurate segmentation, classification, and weight/volume estimation. The challenge uses a novel dataset collected through the MyFoodRepo app, contributed by volunteer Swiss users documenting their daily food intake. This dataset has been annotated to map the individual food items onto an ontology of Swiss Food items.

**Dataset:** The dataset provided by the AICrowd Food Recognition Challenge is an evolving collection of food images with annotations in MS-COCO format. It includes:

- A Training Set with 24,120 RGB images and 39,328 annotations.
- A suggested Validation Set with 1,269 RGB images and 2,053 annotations.

This is a debug Test Set for Round 3, offering the same images as the validation set. The dataset is designed to overcome the limitations of existing food databases, which often feature unrepresentative stock photography without proper annotations. The challenge dataset aims to provide real-world images with proper segmentation, classification, and volume/weight estimates.

**Results:** The Food Recognition Challenge offers substantial prizes to incentivize participants. For Round 4, significant cash prizes were awarded for scores above specific thresholds, with the top prize being 10,000 CHF for a score greater than 0.70. Additionally, the top four winners received an Oculus Quest 2, and a travel grant to AMLD 2021 was also provided. These incentives aim to encourage high-quality submissions and advancements in food recognition.

**Submissions:** Participants were required to set up a proper repository structure and create a private Git repository at GitLab with the contents of their submission. Submissions were identified using an aicrowd.json file containing specific fields, including the challenge ID and whether the submission required a GPU for evaluation. If needed, a NVIDIA-K80 GPU was made available for the submission evaluation.

**Here are some key takeaways from the Food Recognition Challenges:**

**Community Engagement Is Crucial:** The challenges have consistently emphasized community involvement, leveraging crowd-sourced data from the MyFoodRepo app and encouraging developers globally to contribute to the evolving dataset.

**Real-World Application Focus:** The practical use case of the challenge – to help track dietary intake for personal and medical purposes – highlights the importance of AI applications that can be integrated into everyday life.

**Evolving Datasets Enhance Relevance:** The datasets have grown over time, ensuring that the challenge remains relevant and that the algorithms developed are tested against a diverse and up-to-date range of food images.

**Difficulty of Image-Based Recognition:** Despite advances in deep learning, food recognition from images is still a difficult problem due to the variability in food presentation, which these challenges aim to address.

**Quality of Data Over Quantity:** The emphasis on high-quality, well-annotated datasets underscores the challenge's commitment to creating reliable and accurate AI models, moving away from unrepresentative and misleading Internet images.

**Continuous Improvement Through Iterative Rounds:** The challenge's multi-round structure fosters ongoing improvement and innovation, allowing participants to build upon previous work and adapt to new data.

**Incentives Drive Innovation:** Substantial prizes and recognition, such as co-authorships in papers and cash rewards, are significant incentives for participation and pushing the boundaries of current AI capabilities.

**Openness and Collaboration:** By establishing an open benchmark and providing resources like starter kits and discussion forums, the challenge encourages transparency and collaboration within the AI community.

**Accessibility and Ease of Entry:** The challenges have lowered the barrier to entry, allowing a broad range of participants, from those with access to powerful AI models to individuals who may just be starting.

**Results Demonstrate Feasibility and Progress:** The results and solutions generated from these challenges demonstrate the feasibility of using AI for food recognition and tracking, showcasing progress and paving the way for further advancements in the field.

## **Appendix 21.2**

### ***A2. Relation between variables used for modeling***

Exploring the relation of target variables and model prediction further indicates the overall linear nature of the phenomenon under study and indicates that the precision of exceptionally performant solutions fails to be predicted accurately given their scarcity in the data. Figure 21.4 shows the pairs' scatter plots (left quadrant), histograms (diagonal), and correlations (right quadrant) for variables used for modeling. Color-coding of scatter plots represents the different teams.

# HAND OUT OR HELP OUT

## A resource-based view of AI in philanthropy

*Joe Wheeler*

### 1 Introduction

By the end of the 1990s, “effective philanthropy” had become a common practice – its main differentiating feature from traditional giving being that it seeks to measure its impact (Katz, 2005). The primary question effective philanthropy asks is how to do “the most good” with a limited number of resources, compared to more emotionally focused giving (such as to an alma mater or passion project), where the emphasis is on the donor’s values expressed through a gift, rather than necessarily maximizing social impact (Rosqueta, 2014). This new form of philanthropy has catalyzed an array of new instruments (social impact bonds, loan guarantees, patient capital, etc.), institutions (impact investment funds, capital aggregators, internet portals, etc.), and trends (moving beyond cash grants to other forms of donations, including in-kind resources).

The latter will be the focus of this chapter, for corporations with artificial intelligence (AI) capabilities in particular (Salamon, 2014). More specifically, this chapter will refer to AI in the context of large language models (LLMs), or AI that has been trained on vast amounts of text to understand existing content or generate original content, such as OpenAI’s ChatGPT or Google’s PaLM 2 (Gartner, 2023). Given the growing prevalence and interest in AI, understanding its usefulness as a potential donated good is worthwhile.

This chapter is organized as follows. The next section will introduce a brief history of AI applications in philanthropy, what others have said in the field, why this question matters, and how the resource-based view can help answer it. Section 2 will develop the theory, arguing that under certain conditions, resources can accomplish outcomes that cash alone cannot. Section 3 will outline the parameters under which it is better to use technological resources than cash. The three contingencies are as follows: First, the donation must fulfill a unique need, i.e., not offload resources onto NGOs they do not want or need. Second, the donor must be able to manage NGO relationships and have the appropriate personnel with some knowledge and experience of the social problems the NGO is working on. Third, drawing from principal-agent theory, the staff responsible must not compromise core business objectives by advancing their interests over the shared goals of the donor and grantee. Finally, Section 4 will conclude with directions for future research and

opportunities where empirical testing would be useful. This chapter contributes to the literature by offering practical guidance – illustrating how to better meet shared goals between grantees and donors with AI technology.

### ***1.1 AI in philanthropy: literature and use cases***

AI itself is not new, but given that the proliferation of advanced predictive engines like ChatGPT has occurred just within the past year, academic research on AI in philanthropy is still relatively nascent. Farrell (2019) used AI to show a positive correlation between the growing influence of private philanthropy and the large-scale production and diffusion of misinformation (Farrell, 2019). This implies a potentially distortive impact of philanthropy on public access to credible information – further reinforcing the importance of studying philanthropy’s impact on the rest of civil society. But a key distinction here is that this study used AI in its methodology rather than choosing it as a research subject (as this chapter will do) – leaving a dearth of needed scholarly work examining relationships between AI and philanthropy.

Behl et al. (2022) studied the adoption of AI for public services, particularly for disaster relief responses. Using a structural equation model on survey data from 184 government employees, they found that resources (time, money, and skills) were key variables that determined whether AI technology was adopted (Behl et al., 2022). Additionally, they found that organizational culture and voluntariness (i.e., a willingness to use AI) were critical moderators on the effect (Behl et al., 2022). However, rather than examining relationships between donors and NGOs – as will be the focus of this chapter – Behl et al. looked at how resources influence government agencies’ use of AI specifically. There remains a gap in the literature looking at NGOs’ use of AI. Like Behl et al., this chapter uses a resource-based perspective, but it will aim to fill the research gap by focusing on donors and NGOs specifically.

There are a few factors signifying that more research should focus on AI and philanthropy, such as the growing number of AI use cases among NGOs already underway. Use cases range from inputting facial recognition and demographic information – then running through an advanced prediction engine – to help families of refugees who have separated from each other to reconnect (to date, the organization Reunite has helped over 40,000 people reconnect); scientists are tracking endangered animals by using AI and image recognition; the organization College Forward is using college data to automatically flag when students are at risk of not graduating to then recommend coaches and additional academic support, etc. (Wallace, 2018).

Recognizing the potential of AI to serve the nonprofit sector, ChangeFinder was founded to ease the process of writing grant applications (I. Winbrock, personal communication, July 13, 2023). By fine-tuning LLMs through Application Programming Interface (API) integrations, ChangeFinder analyzes grant questions and answers (and whether they were accepted), gives a numeric score based on selection criteria important to funders, and offers feedback and suggested revisions (I. Winbrock, personal communication, July 13, 2023). The goal is to ultimately save NGOs time in applying for grants and give them a higher chance of converting an application into successful funding (I. Winbrock, personal communication, July 13, 2023). Similarly, DonorSearch uses AI to prospect wealthy donors with a particular interest in healthcare funding (“DonorSearch Releases First Nonprofit Vertical Solution,” 2021). Given the wide range of applicable uses in the nonprofit sector, it is critical for corporate donors who possess AI and LLM technology to know when their products might be useful as in-kind donations.

## **1.2 Rationale for focusing on philanthropy**

Wang et al. (2020) have emphasized that corporate philanthropy remains undertheorized, leaving a gap in extant literature (Wang et al., 2020). Example areas of research they call for include exploring philanthropy as a form of reputational insurance; this is indeed important, as some results indicate that philanthropy to boost one's own reputation can cause more social harm than good, reflecting a kind of moral hazard that contradicts the fundamental purpose of corporate giving (Luo et al., 2018). They also call for more theory development on what drives corporate giving, the authenticity behind corporate actions, and how companies can give most effectively, the latter of which will be discussed in detail.

This chapter asks the question: when is it better for international corporate philanthropy to give financial grants, and when is it better to offer AI resources like machine learning and data science trainings, specialized software, and others? This question warrants scrutiny as the opportunity cost is high; companies are giving more year over year, especially in Latin America where corporate philanthropy comprises 50% of all foundations, and the people impacted by better giving models will be NGOs' clients and beneficiaries most in need of aid (Johnson, 2018). Furthermore, this topic is important to validate the original call-to-action from effective philanthropy – to make giving as impactful as possible – not just as semantic window-dressing but in real and meaningful terms.

Put another way, there is a tension between deciding whether to help others by giving cash or creating enterprises that then help others by giving goods and services. In what they call “inclusive capitalism,” Prahalad and Hart (2002) argued in favor of providing goods and services to the poorest people in the world, making the case for growing new markets and entrepreneurial opportunities as a solution to fight poverty (Prahalad & Hart, 2002). For example, several multinational banks are now providing microbanking services in lower-income countries; through shared-access models (e.g., Internet kiosks), wireless infrastructure, and focused technology development, companies are dramatically reducing the cost of WiFi connectivity around the world, and so on (Prahalad & Hart, 2002). However, for AI technology, under what conditions is that useful (particularly as global NGOs are perennially under-resourced and strapped for cash donations)?

Management literature on the firm's resource-based view (RBV) provides a sound theoretical framework to address this question (Barney, 1991). RBV offers a theory of competitive advantage when resources are valuable, rare, imperfectly imitable, and non-substitutable, also known as the VRIN framework (Barney, 1991). The same factors that sustain competitive advantage (VRIN) can also fulfill idiosyncratic NGO needs. For example, a donated AI capability that would allow NGOs to discover and complete grant applications more easily – like ChangeFinder, mentioned in the introduction – could scale exponential fundraising returns, compared to a pure cash donation. In this case, the economic value, dollar-for-dollar, of ChangeFinder's AI technology would be more valuable than cash.

This chapter proposes that offering a diverse range of AI assets is more effective than cash under the conditions that (a) the AI fulfills a specific NGO need that cannot otherwise be easily met, (b) the donor is well-equipped to manage NGO relationships, and (c) the staff responsible for facilitating donations do not compromise core business objectives. As a note, the word “NGO” could be used interchangeably with “nonprofit” in the American context, though there are many different related terms for not-for-profit organizations (civil society organizations, 501c3s, etc.) – for simplicity's sake, “NGO” will be used consistently.

### 1.3 Rationale for choosing these criteria

The above criteria were selected based on both theoretical inferences from management literature and anecdotal feedback from dozens of NGO leaders gleaned over the past decade when the author worked with corporate donors. Many of these individuals have expressed frustration when donors *do not* follow these parameters and donated goods and services that they neither want nor need are dumped onto them. NGOs are perennially strapped for time and resources, so the operating assumption for donors should be that cash is best. However, through the discussion that follows, and further supported by case studies, this chapter will show that there are some circumstances – where all three criteria are met – in which the economic value of receiving AI capabilities can exceed the value of cash if it can help the NGO accomplish more than it could do with cash alone.

## 2 Theory development

The term “resource-based view” (RBV) was first coined by Wernerfelt (1984) and later developed by Barney (1991) and others. The theory advocates not just viewing firms in terms of outputs and products but rather a broad conception of their resources (which could include brand names, in-house tech knowledge, skilled employees, trade contracts, machinery, and efficient procedures) (Wernerfelt, 1984). Part of its usefulness is in helping managers direct their attention (and aid in deciding what *not* to do); if RBV holds, then internal resources (like honing internal technological, organizational, and managerial processes) can more effectively create wealth compared to strategizing (thwarting competitors and making the market difficult for new entrants), or capital investments, or automation (Teece et al., 1997). The resource-based view suggests that organizations must develop *unique, firm-specific* core competencies that will allow them to outperform competitors by *doing things differently* (Hamel & Prahalad, 1990).

### 2.1 VRIN framework applied to philanthropy

The same factors that sustain competitive advantage (VRIN) can fulfill idiosyncratic NGO needs. RBV emphasizes that a firm’s resources cover many physical, human, and organizational categories – and the extent to which they can sustain competitive advantage depends (in part) on management and the effectiveness of internal routines and processes (to the extent that they are management-specific and not easily recreated). For example, several unique features of OpenAI distinguish it from other generative AI competitors like Google, including scale (ChatGPT has trillions of parameters), natural language understanding, and a commitment to AI research. Developing these unique features relied on a diverse set of organizational resources, including intellectual capital, management capabilities, leadership, and so on.

Let us call these diverse resources “streams of influence.” A business does not only have the products it sells, but it also has office facilities, physical goods associated with company benefits (food at the office, for some), knowledge and intellectual capital, the skills of the workers, etc.; this meting out of various types of assets form the basis of a resource-based view of philanthropy (RBVP). The steps for an AI organization are: (1) identify all the various physical, human, and organizational resources the business has; (2) determine if NGOs need one or more of those resources (that cannot otherwise easily be fulfilled); (3) ensure the business has the right staffing (such as full-time Corporate Responsibility workers, or staff experienced in NGO relationships who have part of their responsibilities doled out for social impact work); and (4) determine whether the right

monitoring mechanisms are in place to monitor performance – and use those steps to determine whether the AI company adopts RBVP or not. Rather than viewing philanthropy as an ancillary silo, why not have a social impact lens through every stream of influence?

For example, donating routines could be an option. The dynamic learning that occurs through donated and shared routines could enable tasks to be performed better and quicker (e.g., learning new AI software) and enable new types of programs (e.g., learning how to benefit from advanced LLMs that could be passed down to an NGO's clients) (Teece et al., 1997). One could also donate assistance in creating an “organizational culture” – contributing to an NGO's brand equity, helping it gain publicity, bestowing internal processes and efficiencies, design-thinking or brainstorming techniques, etc. (Barney, 1986, as cited in Barney et al., 2011; Fiol, 1991, as cited in Barney et al., 2011).

Indeed, these sorts of donations could fill a significant gap. Contemporary research suggests that – due to a lack of expertise, low financial budgets, and insufficient awareness of data analytics capabilities – many NGOs are missing out on the potential for big data and machine learning to provide valuable insights (e.g., better understanding donor behavior) (Alsolbi, 2023). AI's potential benefits to NGOs also include performing administrative tasks, optimizing resource allocation, designing programs, monitoring and evaluating, and personalizing donor engagement (Efthymiou et al., 2023a, 2023b). However, successfully using these capabilities relies upon adequate training (mainly to prevent potentially negative consequences such as the proliferation of disinformation) – and RBV suggests that donors with AI expertise can fulfill a crucial role in this regard (Efthymiou et al., 2023a, 2023b).

RBV shares some similarities with the “economies of scope” concept put forth by Panzar and Willig (1981); as an example, the company Bic was able to use a single resource for several businesses, including pens, lighters, and razors (Wernerfelt, 1984). They entered the market sequentially (one after the other) from a position of strength. This idea also applies to AI philanthropy – diversifying to a nonprofit use rather than an additional business vertical. For example, LinkedIn already has built up the largest network of individuals' professional data; they would be well-suited to deploy AI in responding to queries and/or generating content for job seekers (such as relevant job trainings or networking opportunities for underserved communities). RBV views *related* diversifications (those that build upon existing capabilities) as the only type worthy of merit, and by this logic, donating assets relevant to the core business could be more appealing than giving cash (Teece et al., 1997). RBVP is an extension of the focus and specialization that defines a company's capabilities (in the case of LinkedIn, its professional network) (Teece et al., 1997).

## **2.2 Inclusion of dynamic capabilities**

Let us apply this to a contemporary real-world example: philanthropy and pro bono during the height of the COVID-19 pandemic – where the dynamic capabilities view is especially relevant. Winter (2003) defines dynamic capabilities as those that would change the product, the production process, the scale, or the customers (markets) served – diversifying scale (including geographic expansion) and scope (Winter, 2003). Teece et al. (1997) refer to these processes as reconfiguration (transformational), with so-called “high flex” firms being those that have honed these adaptive capabilities (Teece et al., 1997). One is hard-pressed to think of a more salient modern example of how important being “high flex” is than sustaining an organization through COVID-19. Firms had to change their business models, adapt to a remote work environment, find new ways to meet customer demand, etc. This applies not only to COVID-19 but to other types of crises as well (mass shootings, natural disasters, etc.), especially if the company has staff in the locations affected.



In the context of philanthropy, there are numerous examples where RBVP is relevant. As the world's largest search engine, Google played a role in early virus detection. Their "early search trends" feature aggregated relevant searches (e.g., "What to do if I have COVID-19 symptoms?" "How long should I quarantine?" "What are the current masking guidelines?") as proxy indicators for detecting a rapid spread early (the logic being that geographic regions with high volumes of certain queries can indicate whether a major outbreak is likely). Moreover, the continued advancement of LLMs could enable even more sophisticated early detection tools in the future; of course, this information is relevant for major national and international public health organizations (like the CDC, the WHO, etc.).

Applying the steps of RBVP to this example: (1) the capacity for Google to collect this data (given that it has perhaps the most sophisticated data collection techniques of any business in the history of civilization) is a unique capability that goes beyond its primary products (using its technology to sell ads), and identifying the relevancy of early search trends for COVID-19 detection required some discernment and effort from management; (2) NGOs (especially those like the Gates Foundation who work on curbing infectious diseases) certainly need this information (Zuboff et al., 2019). It would be otherwise difficult to obtain because Google already has the behemoth infrastructure, widespread market share, and unique capabilities to capture and store this type of data. (3) Google has the right staffing, given that Google.org is a multibillion-dollar part of the company whose sole purpose is social impact through aiding NGOs and social enterprises; and (4) assuming Google.org's staff are held accountable through regular performance evaluations and feedback from NGO partners, they would have monitoring mechanisms in place. Therefore, the criteria are met to make Google's technology suited for RBVP.

The following section will detail each contingency, outlining the three propositions for when donated assets are more effective than cash.

### 3 Defining the terms: three contingencies

#### 3.1 Fulfilling a unique need

The first contingency for AI resource donations is that they must fulfill a unique NGO need. Wernerfelt's (1984) definition of resources includes any strength of a firm or any semi-permanent advantages – which could, therefore, make the resources valuable for NGOs as well, so long as they are protected from competitors (Leiblein, 2011; Lippman & Rumelt, 1982; Wernerfelt, 1984). In other words, if a company can meet an NGO need that *others cannot* (or that the NGO cannot meet itself), this makes a strong case for donating AI assets. For example, St. Jude's Research Hospital has used Google's AI tool "target cost per-acquisition bidding" to reach new donors through language and audience testing at scale (Levesque, 2023). To the extent that St. Jude's is unable to procure this service from other providers or through an open-source network, then tech training, customization to fit the needs of the NGO, etc., is ripe for Google to consider as a pro bono donation.

Dierickx and Cool (1989) developed the notion that resources are especially valuable when no effective substitutes are available (Dierickx & Cool, 1989, as cited in Barney et al., 2011). This lends itself to "helping out" rather than "handing out"; if donated AI technology is not easily substitutable, then presumably, an NGO could not simply purchase it in factor markets (or at least it would be much more costly to do so). Distinctive competence must be built; it cannot be bought (Teece et al., 1997). For example, OpenAI is launching a proprietary subscription plan, "ChatGPT Plus," that allows the user to input images or voice for content generation (Agomuo & Larson, 2023).

Broadening the scope of generative AI opens additional pathways for NGOs to generate engaging social media content, draft tailored outreach emails, etc. – which could, in turn, generate more resources for the NGO than a simple cash donation would provide (Levesque, 2023).

Donors' choices about how to give strategically should flow mainly from an internal analysis of their unique skills and capabilities (Barney, 1986). They should best exploit their internal AI resources relative to external opportunities – matching market supply to NGO demand (Barney, 1986). The alternative is dumping resources onto NGOs that they neither want nor need – reinforcing a harmful (and non-equity-based) power relationship between giver and recipient, giving with “strings attached” in a way that serves the interests of the donor more than the NGO, is another form of donating without meeting NGOs' needs. In some cases, this can create more harm than good; for example, if generative AI queries yield false or misleading results, this could promulgate disinformation. Or, if AI generates personal identifiable information (PII) that compromises an individual's privacy, the technology could work *against* an NGO's goals. Giving unique AI assets, based on self-identified demand from NGOs, is a way to safeguard against the potential of just “checking a box,” doing philanthropy for vanity's sake, or causing harm through a donation.

Non-substitutable AI gifts should complement, rather than supplement, existing NGO efforts (Barney, 1991). A donation should not displace any ongoing initiatives but elevate the work already being done. Philanthropy that is closely related to the donor's core business and complements rather than supplements NGO work will most strongly benefit society (Kaul & Luo, 2018). When this is not the case, donated assets may add little social value and may even be harmful (Kaul & Luo, 2018). As Kaul and Luo (2018) stated, donating AI organizations should ask, “What is my firm's unique advantage in serving this cause relative to alternative providers?” (Kaul & Luo, 2018)? In sum, AI donors should “help out” where they have a relative advantage to do so, and they should not step on NGOs' toes. This forms the basis of the first hypothesis:

H1: If there is a good fit between a donor's resources and an NGO's needs, then “help out” (give AI).

### **3.2 Can manage NGO relationships**

In RBV, the ability to manage inter-organizational relationships can improve performance (Kale et al., 2002). Similarly, for donated AI to add value, the donor must have the capacity and ability to manage NGO relationships; this section will explore why this is the case and provide some examples.

Schmidt and Keil (2013) highlight the role of managers both in understanding the potential of resources to create value and in building strategies to leverage resources undervalued by others (Schmidt & Keil, 2013). For managers to be effective, they must not only create but also capture value (Kim & Mahoney, 2010). “Creating value” in a philanthropy context refers to creating social value – but still, a donation's success hinges on managers' abilities to leverage resources and maintain strong inter-organizational relationships. In the context of AI and philanthropy, there are several barriers that donors and beneficiaries' managers must address to ensure that AI can aid the NGO's mission: adequate staff training, integration with other data systems, appropriate data usage, a desire among the NGO's leadership to adopt the technology, etc.

The abundant literature on public-private partnerships (PPPs) can help inform the parameters to determine whether a donor has appropriate capacity and staffing. Insofar as the government outsourcing a public service to an NGO (as in many PPPs) has some similarities with an NGO outsourcing to a donor (i.e., receiving pro bono time, skills-based volunteering, technological resources, or some combination thereof), the literature carries some relevant insights.

First, due to the complexity of any relationship that requires a contractual agreement (like in-kind grants of AI technology), Savas and Gilroy (2020) argue that strong commitment is necessary from the top. Without a clear commitment from senior leadership, mid-management is left to negotiate the minute details of a contract, which are then sent back up an ambiguous chain of command. Technology companies with corporate social responsibility (CSR) and/or charity programs – that do not have C-level buy-in – often must wrestle with unclear contracting requirements, overcoming opposition, effectively delegating employee roles, forming consistent partnerships with the recipients of donated technology, monitoring the work, evaluating the results, and determining whether to renew or terminate the contract upon completion (Savas & Gilroy, 2020).

Likewise, as senior leadership approval is a necessary first step, having clearly articulated partnership parameters follows as a next necessary step (Gerrard, 2001). The management team must know the constraints of a particular agreement – which is important for the donor and grantee. Without clearly defining the boundaries of an agreement, there is a risk of “over-tipping” (e.g., the donor has given too much, and the recipient now must absorb an exorbitant donation that is not appropriate for the size of their organization) or “under-tipping” (the donation is too small or insignificant to make a difference for the grantee). Ambiguous partnership agreements can also lead to “vendor lock-in,” where NGOs become dependent upon a donated service and must pay for it later. If proprietary AI technology that was not interoperable with other systems were donated, then the grantee would be stuck with something that may or may not be serving their mission in the long run.

Furthermore, a partnership cannot be appropriately staffed unless the commitments are explicitly stated (Akintoye et al., 2008). Staff must also be incentivized to transfer their time and skills (Akintoye et al., 2008). For example, consider a scenario where a literacy NGO would like to use a donated AI service to generate stories that children can use to practice reading. They make an appointment to be trained on using an LLM. Still, the technologist deprioritizes the appointment and does not show up because it was not tied to their compensation and/or core job responsibilities. This then hurts the relationship and future partnership opportunities between the NGO and AI company – and that problem can be avoided by a transparent staff allocation tied to performance up front.

Furthermore, a lack of enthusiasm can exist on either the donor or grantee side – NGO staff may be reticent to embrace new ideas and working methods. Indeed, from a study examining employee perceptions on the adoption of AI from 24 companies across 11 countries, they found that simply providing AI tools is not enough; effective adoption relies on communication quality and reinforcement, local and senior management support, training, ethical reviews, reporting mechanisms, sufficient technical infrastructure, and more (Kelley, 2022). This is partly dependent upon the caliber of the staff transferring time and skills; therefore, for the partnership to work, the donor must allocate competent staff to do the training. The training must be comprehensive enough that the diverse skills and expertise to get value from LLMs are adequately transferred to the NGO’s staff (Kwak et al., 2009).

This can also help ensure continuity once a partnership has ended. If an NGO’s staff gets trained on an AI platform that later breaks down or can no longer be used, the staff has no one to turn to for ongoing maintenance. In this case, the partnership could do more harm than good. Therefore, an in-kind or pro bono contract must be complete (De Bettignies & Ross, 2004). Other donated technical service provision has often faltered by relying on incomplete contracts; for example, Engineers Without Borders built 113 gravity-fed systems (networks of water pipes where communities can access clean water from taps), and only 32 of them were functional after the initial implementation (Damberger, 2011). There was no forethought about who would maintain

this system (Damberger, 2011). As this case illustrates, bestowing a new infrastructure onto an organization or group – without clearly outlining how it will be maintained – can lead to failure (Damberger, 2011). Consequently, as such, donor/grantee partnerships must have mechanisms in place to monitor the quality of the donated good or service (Flinders, 2005). That way, if donated AI proves to be superfluous or not functional, the grantee is not locked in and beholden to something not serving them.

In sum, an AI donor should have the appropriate personnel (with appropriate experience and/or some knowledge of the NGO's social problem). Yescombe (2011) calls this “institutional capacity”: Does the donor have a working knowledge of the NGO's culture, lingo, priorities, ways of working, etc.? This will help determine whether a skill and resource match exist between donor and recipient. Consider some counterexamples. In a situation where dozens of volunteers from a technology firm show up at a community shelter for houseless youth – with next to no training on how to work with them – this could not only create an administrative burden for the community shelter but also cause more harm than good for the youth clients. If they had a positive prior relationship (and the relational know-how to work with them effectively), they would have known to *ask first* and then fill in the gaps where needed.

The social processes that a donor organization possesses will inform its ability to sustain NGO relationships – in a way that is imperfectly inimitable (not the same for other donors' relationships with the NGO) (Barney, 1991). In the example above, if the appropriate processes were in place, then rather than simply showing up to volunteer, an event would be co-planned based on a salient relationship that can match skills offered with real needs. This relational element is the focus of the second hypothesis:

H2: If the donor can manage NGO relationships effectively, then “help out” (give AI).

### ***3.3 Donor's staff do not self-serve***

The third contingency is that the staff responsible for administering AI donations should not be opportunistic or self-serve in a way that compromises core business objectives. Williamson (1975) described opportunism as “self-interest seeking with guile,” which could take the concrete forms of acting to boost one's clout, compensation, position in a corporate hierarchy, etc. (Eisenhardt, 1989). In the context of donated AI goods and services to NGOs, this is problematic, as it is fundamentally at odds with the mandate of NGOs to act in the public rather than private interest.

Agency theory is helpful to draw from here, as this is a quintessential principal-agent problem: the staff responsible for managing donations want more responsibility, which serves their interests more than the shared goals between donor and grantee (Dalton et al., 2007). A central tenet of agency theory is the “potential for mischief.” In the philanthropy context, this can be seen and felt in numerous ways (e.g., the power-hungry executive who wants to appear as an altruistic socialite commits to more than 50 different NGO boards but does not have the time to serve any of them meaningfully). Incorporating agency theory into donor relations certainly fits with Eisenhardt's (1989) core recommendation to “incorporate an agency perspective in studies of the many problems having a cooperative structure.”

Like Transaction Cost Economics, agency theory deals with information asymmetry (Arrow, 1985). For example, the “shirking” issue that Alchain and Demsetz described in 1972 is reminiscent of the self-interested agent: in the absence of monitoring (in a team environment), what will stop a donation manager from doing the least work possible (and therefore failing to serve an NGO's bottom line) while reaping the maximum reputational benefit? As Alchain and Demsetz

(1972) posed, the problem transcends another level above: “Who will monitor the monitors” (a task which, in and of itself, has transaction costs and imperfect information)? Hence, an AI donation program is contingent upon some degree of making sure the managers, on the corporate side, are serving the organizations and not themselves.

For example, consider a scenario where a data science lead helps an NGO access meaningful tools like Google’s SOS Alerts – leveraging AI for disaster management (“Sounding the Alarm,” 2023). Whether or not the first-responding NGO used this tool, this data science manager can chalk up this activity as a win on their quarterly goals (sometimes referred to as Objectives and Key Results, or “OKRs” in the technology industry). To go beyond a self-serving action, this manager should work with the partner to ensure the technology was adopted meaningfully (which requires some degree of monitoring).

Prioritizing a benefit to one’s reputation over social welfare carries significant consequences. For example, in Luo et al.’s (2018) study of the US petroleum industry, they found that philanthropic donations are positively associated with subsequent oil spills (firms that give more spill more). This implies that contributions (for example, to environmental groups like the World Wildlife Fund or The Nature Conservancy) could be offset by careless behavior on the part of the donor. Or, using Google as an example, Google.org’s multimillion-dollar donations could be offset by other problems regarding surveillance or misinformation on the core search engine.

For donating organizations, one solution might be to invest more in formal information systems (including boards of directors) to reduce the cost of information-gathering (Eisenhardt, 1989). But just having a board by itself is not necessarily enough to account for the agency problem; it is contingent upon the shortcomings of the board or managers (or both) (Walsh & Seward, 1990). Therefore, organizations should work to diffuse decision control and management beyond just a few agents (Fama & Jensen, 1983). This would mean organizations like OpenAI, Google, Microsoft, etc., would stay accountable by seeking counsel from experts across sectors (academia, ethical review boards, nonprofits, etc.).

This theory applies not only to maximizing profit but maximizing social missions as well. Reducing the agency problem will ultimately help deliver demanded goods/services (i.e., a donor’s AI resources) most efficiently. This is important because reducing the likelihood of philanthropy managers acting in their self-interest will benefit both principal and agent – whether financially or protecting one’s social status in and out of work as well (Arrow, 1985; Zeckhauser, 1985). If this condition is not met, then there is no way to ensure alignment between a donor’s AI resources and an NGO’s needs (if a manager is prioritizing himself or herself over shared goals) – therefore leading to the third hypothesis:

H3: If a donor can mitigate employee self-interest, then “help out” (give AI).

In sum, if one or more of the above conditions are not met (H1, H2, or H3), cash is better than resources; eventually, it would be more economical to let the NGO decide how to spend cash based on self-identified needs. If the donor does not have the appropriate AI resources (H1: no fit between supply and demand), is not equipped to help NGOs (H2: cannot manage relationships), and the staff strays from shared objectives (H3: staff acts to self-serve), then there is no fit for RBVP.

However, if the three contingencies are met, AI resources can accomplish what cash alone cannot. For example, a software company might create a video identification feature for a human rights NGO (e.g., if common characteristics of user-uploaded videos (hate speech, doxing,

incitement of violence, etc.) are found across multiple videos in a shared geography, it could signal a human rights risk). They already possess the resources (human capital, intellectual know-how, necessary patents on technology, etc.) and have paid high sunk costs to acquire these assets. It is, therefore, more economically efficient for a company that possesses these capabilities and has already paid sunk costs to donate them over cash; the NGO could not achieve the same results with cash (assuming they do not possess the same unique set of capabilities as the donating firm, and that it would cost exponentially more for them to try to acquire the identification feature on their own). Therefore, gifting the video technology meets shared goals between donor and recipient, whether for the impact in and of itself or ancillary benefits (employee morale, strengthened external relationships, etc.).

#### **4 Discussion and conclusion**

This theory would benefit from empirical testing, and though that is not the focus of this chapter, it should be an avenue for future research. Scholars could regress a dependent binary variable, “cash or AI assets,” on independent variables defined by using H1, H2, and H3 (binaries for “yes” or “no” on each). For example, one could gather data from a randomized selection of companies that build LLMs (e.g., using a systematic sampling method, one could gather a publicly available list of companies and sample every third one) to determine coding for the dependent variable; the next step would be to determine whether the donor gives assets or cash. And finally, the independent variables could be assessed using grantee survey data. Analyses such as these will help strengthen a collective understanding of whether RBVP holds up in practice.

One limitation is that – given the vast body of literature on RBV – some theories or interpretations may have been missed. To mitigate this risk, as many review articles as possible were synthesized. Furthermore, this chapter provides a reimagining of RBV, perhaps different from the applications its architects had originally imagined; this could be a limitation, but it could be a strength in some ways. Barney et al. (2011) called for the need to innovate RBV, ensuring that it achieves revitalization and avoids decline – hence the need for this chapter and others like it. Similarly, Priem and Butler (2001) criticized RBV for lacking prescriptive implications; not only does the “hand out or help out” framework provide a prescriptive implication for AI in philanthropy, but it does so to advance theory for NGOs (arguably doing some of the most important work in the world).

In conclusion, there are many types of philanthropy; some strictly give cash, while others give combinations of resources (e.g., Salesforce). With the emergence of firms building new types of LLMs, understanding whether cash or AI is better (and under what conditions) is worthy of academic scrutiny given the substantial opportunity cost and size of the charitable sector (for example, Google alone gives more than \$100 million annually to causes like COVID-19 research, racial justice, and environmental sustainability). This chapter’s argument is that giving AI assets is preferred over cash under the conditions that (a) the AI fulfills a specific NGO need that cannot otherwise be easily met, (b) the donor is well-equipped to manage NGO relationships, and (c) the staff responsible do not succumb to an “agency problem” that compromises core business objectives. This theory contributes to the literature by applying RBV to a meaningful, materially significant, emerging part of public life. Future research should empirically examine outputs and outcomes associated with AI donations (compared to cash donations) – strengthening the potential applications of this theory in practice.

## References

- Agomuoh, F., & Larson, L. (2023). ChatGPT: the latest news, controversies, and helpful tips. *Digital Trends*. <https://www.digitaltrends.com/computing/how-to-use-openai-chatgpt-text-generation-chatbot/>
- Akintoye, A., Beck, M., & Hardcastle, C. (Eds.) (2008). *Public-private partnerships: managing risks and opportunities*. Hoboken, NJ: John Wiley & Sons.
- Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *The American Economic Review (AER)*, 62(5), 777–795.
- Alsolbi, I. N. (2023). Leveraging potentials of big data for better decision-making and value creation in non-profit organisations (Doctoral dissertation).
- Arrow, K. J. (1985). Informational structure of the firm. *The American Economic Review (AER)*, 75(2), 303–307.
- Barney, J. B. (1986). Strategic factor markets: expectations, luck, and business strategy. *Management Science*, 32(10), 1231–1241.
- Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99–120.
- Barney, J. B., Ketchen Jr, D. J., & Wright, M. (2011). The future of resource-based theory: revitalization or decline? *Journal of Management*, 37(5), 1299–1315.
- Behl, A., Chavan, M., Jain, K., Sharma, I., Pereira, V. E., & Zhang, J. Z. (2022). The role of organizational culture and voluntariness in the adoption of artificial intelligence for disaster relief operations. *International Journal of Manpower*, 43(2), 569–586.
- Dalton, D. R., Hitt, M. A., Certo, S. T., & Dalton, C. M. (2007). The fundamental agency problem and its mitigation: independence, equity, and the market for corporate control. *Academy of Management Annals*, 1(1), 1–64.
- Damberger, D. (2011, April 21). *Learning from failure—David Damberger* [Video]. [https://www.youtube.com/watch?v=HGjHU-agsGY&ab\\_channel=TEDxTalks](https://www.youtube.com/watch?v=HGjHU-agsGY&ab_channel=TEDxTalks)
- De Bettignies, J. E., & Ross, T. W. (2004). The economics of public-private partnerships. *Canadian Public Policy/Analyse de Politiques*, 30, 135–154.
- Dierickx, I., & Cool, K. (1989). Asset stock accumulation and sustainability of competitive advantage. *Management Science*, 35(12), 1504–1511.
- DonorSearch (2021, March 2). DonorSearch releases first nonprofit vertical solution for healthcare philanthropy: DonorSearch Aristotle Health aims to transform grateful patient philanthropy with Artificial Intelligence. <https://www.prnewswire.com/news-releases/donorsearch-releases-first-nonprofit-vertical-solution-for-healthcare-philanthropy-301238647.html>
- Efthymiou, I. P., Alevizos, A., & Sidiropoulos, S. (2023a). The role of artificial intelligence in revolutionizing NGOs' work. *Journal of Politics and Ethics in New Technologies and AI*, 2(1), e35137–e35137.
- Efthymiou, I. P., Egleton, T. W. E., Chatzivasileiou, S., & Emmanouil-Kalos, A. (2023b). Artificial intelligence and the future for charities. *International Journal of Non-Profit Sector Empowerment*, 2(1), e35345–e35345.
- Eisenhardt, K. M. (1989). Agency theory: an assessment and review. *Academy of Management Review*, 14(1), 57–74.
- Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *The Journal of Law and Economics*, 26(2), 301–325.
- Farrell, J. (2019). The growth of climate change misinformation in US philanthropy: evidence from natural language processing. *Environmental Research Letters*, 14(3), 034013.
- Fiol, C. M. (1991). Managing culture as a competitive resource: an identity-based view of sustainable competitive advantage. *Journal of Management*, 17(1), 191–211.
- Flinders, M. (2005). The politics of public-private partnerships. *The British Journal of Politics and International Relations*, 7(2), 215–239.
- Gartner. (2023). Information technology: Gartner glossary. <https://www.gartner.com/en/information-technology/glossary/large-language-models-llm>
- Gerrard, M. (2001). Public-private partnerships. *Finance and Development*, 38(3), 48–51.
- Hamel, G., & Prahalad, C. K. (1990). Strategic intent. *Mckinsey Quarterly*, (1), 36–61.
- Johnson, P. D. (2018). Global philanthropy report: perspectives on the global foundation sector. *Harvard Kennedy School*. <https://www.hks.harvard.edu/centers/cpl/publications/global-philanthropy-report-perspectives-global-foundation-sector>

- Kale, P., Dyer, J. H., & Singh, H. (2002). Alliance capability, stock market response, and long-term alliance success: the role of the alliance function. *Strategic Management Journal*, 23(8), 747–767.
- Katz, S. N. (2005). What does it mean to say that philanthropy is “effective”? The philanthropists’ new clothes. *Proceedings of the American Philosophical Society*, 149(2), 123–131.
- Kaul, A., & Luo, J. (2018). An economic case for CSR: the comparative efficiency of for-profit firms in meeting consumer demand for social goods. *Strategic Management Journal*, 39(6), 1650–1677.
- Kelley, S. (2022). Employee perceptions of the effective adoption of AI principles. *Journal of Business Ethics*, 178(4), 871–893.
- Kim, J., & Mahoney, J. T. (2010). A strategic theory of the firm as a nexus of incomplete contracts: a property rights approach. *Journal of Management*, 36(4), 806–826.
- Kwak, Y. H., Chih, Y., & Ibbs, C. W. (2009). Towards a comprehensive understanding of public private partnerships for infrastructure development. *California Management Review*, 51(2), 51–78.
- Leiblein, M. J. (2011). What do resource-and capability-based theories propose? *Journal of Management*, 37(4), 909–932.
- Levesque, J. (2023, July 13). How nonprofits are using AI. <https://blog.techsoup.org/posts/how-nonprofits-are-using-ai>
- Lippman, S. A., & Rumelt, R. P. (1982). Uncertain imitability: an analysis of interfirm differences in efficiency under competition. *The Bell Journal of Economics*, 13(2), 418–438.
- Luo, J., Kaul, A., & Seo, H. (2018). Winning us with trifles: adverse selection in the use of philanthropy as insurance. *Strategic Management Journal*, 39(10), 2591–2617.
- Panzar, J. C., & Willig, R. D. (1981). Economies of scope. *The American Economic Review*, 71(2), 268–272.
- Prahalad, C. K., & Hart, S. (2002). *The fortune at the bottom of the pyramid*. Booz & Company. <https://www.strategy-business.com/article/11518#:~:text=Prahalad%20and%20Stuart%20Hart%2C%20professors,they%20could%20engage%20them%20profitably>.
- Priem, R. L., & Butler, J. E. (2001). Is the resource-based “view” a useful perspective for strategic management research? *Academy of Management Review*, 26(1), 22–40.
- Rosqueta, K. (2014). Rethinking the E word. *Stanford Social Innovation Review*. [https://ssir.org/articles/entry/rethinking\\_the\\_e\\_word](https://ssir.org/articles/entry/rethinking_the_e_word)
- Salamon, L. M. (Ed.) (2014). *New frontiers of philanthropy: a guide to the new tools and actors reshaping global philanthropy and social investing*. New York: Oxford University Press.
- Savas, E. S., & Gilroy, L. (2020). Contracting: privatization and public-private partnerships. In D. Bearfield, E. Berman & M Dubnick (Eds.), *Encyclopedia of public administration and public policy-5 volume set* (pp. 1–6). New York: Routledge.
- Schmidt, J., & Keil, T. (2013). What makes a resource valuable? Identifying the drivers of firm- idiosyncratic resource value. *Academy of Management Review*, 38(2), 206–228.
- Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7), 509–533.
- Wallace, N. (2018, July). How charities are using artificial intelligence. *The Chronicle of Philanthropy*, 30(9), 14. *Gale Academic OneFile*. [link.gale.com/apps/doc/A548448655/AONE?u=euge94201&sid=bookmark-AONE&xid=64f47578](https://link.gale.com/apps/doc/A548448655/AONE?u=euge94201&sid=bookmark-AONE&xid=64f47578). Accessed 29 Aug. 2023.
- Walsh, J. P., & Seward, J. K. (1990). On the efficiency of internal and external corporate control mechanisms. *Academy of Management Review*, 15(3), 421–458.
- Wang, H., Gibson, C., & Zander, U. (2020). Editors’ comments: is research on corporate social responsibility undertheorized? *Academy of Management Review*, 45(1), 1–6.
- Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal*, 5(2), 171–180.
- Williamson, O. E. (1975). *Markets and hierarchies*. New York, NY: Free Press.
- Winter, S. G. (2003). Understanding dynamic capabilities. *Strategic Management Journal*, 24(10), 991–995.
- Yescombe, E. R. (2011). *Public-private partnerships: principles of policy and finance*. Elsevier.
- Zeckhauser, R. J. (1985). *Principals and agents: the structure of business* (p. 37). J. W. Pratt (Ed.). Boston, MA: Harvard Business School Press.
- Zuboff, S., Möllers, N., Wood, D. M., & Lyon, D. (2019). Surveillance capitalism: an interview with Shoshana Zuboff. *Surveillance & Society*, 17(1/2), 257–266.



# PHILANTHROPY’S URGENT OPPORTUNITY TO CREATE THE INTERIM INTERNATIONAL AI INSTITUTION (IIAI)

*David Evan Harris and Anamitra Deb*

## **1 Introduction: the pacing problem and the philanthropic solution**

### *1.1 Background: AI on the horizon*

The pace of artificial intelligence (AI)<sup>1</sup> innovation is dizzying. Even industry insiders struggle to keep up with the daily announcements of new technology developments and regular breakthroughs constantly on the horizon. Around the world, there is a dizzying mix of ebullience for both the societal value of AI and widespread concern about the current harms and future risks stemming from AI (De Tena et al., 2023; Orth, 2023).

The hoped-for benefits of AI remain a driving force for the industry’s continued development. Proponents of AI development cite many potential benefits, ranging from product innovations like digital assistants and self-driving cars to monumental changes in society, including solutions to climate change and a cure for cancer. That being said, AI experts broadly agree that with AI’s benefits come risks (Grace et al., 2024, p.3). These include the use of AI in surveillance, manipulation, propaganda, creation of non-consensual intimate imagery (NCII), violation of privacy, and development of dangerous weapons that could pose catastrophic risks to society (Hendrycks et al., 2023; Lakatos, 2023). Others point out that the harms of AI are already here, citing AI’s role in discrimination in criminal sentencing (Larson et al., 2016), as well as access to health care, housing, credit, and jobs (Akselrod & Venzke, 2023; Backman, 2023; McIlwain, 2020).

An unexpected twist that distinguishes AI from many other industries is that the CEOs of nearly all the major companies (Zakrzewski et al., 2023) involved in AI have themselves called for regulation of the technology itself—facing market pressures to release new products more quickly than their competitors while at the same time publicly worrying that the frenetic pace of these releases could pose risks to society if not rigorously safeguarded by regulation (McCracken, 2023). In a May 2023 congressional hearing, OpenAI CEO Sam Altman stated, “I think if this technology goes wrong, it can go quite wrong,” also saying that OpenAI would “want to work with the government to prevent that from happening” (Kang, 2023). In a September 2023 congressional committee hearing, other major technology business leaders, including Elon Musk and Mark Zuckerberg, also agreed that the government needs to play some role in regulating AI (Wong et al., 2023).

## **1.2 Governments fall short**

In the face of the aforementioned calls for regulation, elected officials from around the world are working hard to establish regulations for AI. Unfortunately, their efforts have been stymied by two major challenges. The first is the geographic flexibility of AI companies that can simply relocate their way out of regulation by ceasing to do business in jurisdictions that choose to impose regulations or, at a minimum, threaten to do so. Sam Altman demonstrated this corporate capability when he threatened to remove OpenAI's products from Europe if the EU AI Act required AI companies to disclose the use of copyrighted materials in their training data. The EU's Commissioner for the Internal Market, Thierry Breton, called Altman's bluff, and later Altman recanted (Chee, 2023; Field, 2023). Nevertheless, this dynamic puts significant pressure on regulators who do not want their work to be seen as an "innovation killer" (O'Reilly, 2023), and the only real solution is for governments to band together in regulating AI. The second major challenge is the "pacing problem," described in greater detail below, which is a known issue wherein governments struggle to make regulations at the same pace as technological development.

The U.S. Congress—the most powerful democratically elected body with the potential to meaningfully regulate the AI industry—remains an institution designed to move slowly and deliberately (Turner, 2016). While draft bills on AI have been circulated, it remains uncertain, at best, whether any of the comprehensive approaches will become law anytime soon (Covington, 2023). Without federal legislative action on AI, any other approach will be limited in its ability to stop the current "race to the bottom" on AI ethics and safety (Harris, 2023a). State- and municipal-level legislation can be at least partially circumvented through corporate relocation (as discussed above), and the force of public opinion can only have so much power. Left only to the economic pressures that incentivize the current rapid development of AI technologies, it appears likely that AI companies in the United States will continue to minimize expenditures on ethics and safety while maximizing development of new system capabilities.

In October 2023, the White House, seeking to make progress on AI regulation without needing congressional approval, issued an executive order to signal their priorities and urge several agencies and departments to take meaningful action on AI at the federal level (Harris, 2023b). Despite striking a balance between the interests of industry and the public interest, the approach has inherent limitations—executive orders can be easily reversed by subsequent administrations and lack the congressional power to allocate funds (Thrower, 2021).

For those interested in ensuring that AI brings the world maximum benefit and minimum harm, the frustrating lack of durable and meaningful progress in Washington leaves us wondering where to look. Around the world, cities, states, countries, and regional organizations are beginning to create their own AI legislation (International Association of Privacy Professionals, 2023; Johnson, 2023), creating a *weak patchwork* of laws that make up the emerging regulatory environment for AI.

The first elements of this patchwork came from China, the first nation to implement firm AI regulations, starting in 2017. Unfortunately, given the country's long history of using technology for surveillance, censorship, and control of its population, while some of its laws prohibiting deep fakes and curbing the power of AI developers seem reasonable, the overall approach does not provide a palatable example for democratic nations (Heath, 2023). There are key instances of overlap in Chinese and democratic interests, such as China's novel provisions on "Deep Synthesis Technologies," aiming to regulate each step of the deepfake process, from deepfake generation to sharing. While there may be disagreements over certain AI policies, it is clear that some of the most significant threats of AI, such as the use of deepfakes, serve as a common ground for democratic and non-democratic countries to collectively build regulation (Sheehan, 2023).

With its 2024 AI Act, the EU followed China and has now emerged as the most significant player among democratically governed nations on AI regulation (Chan, 2023; European Parliament, 2023). It places bans on numerous uses of AI, including “cognitive behavioral manipulation”—a broad term for technologies that interpret behaviors and preferences with the intention of influencing our decisions. The bans also include the “untargeted scraping of facial images from the internet or CCTV footage,” a practice already used by some companies that sell databases used for surveillance; “emotion recognition in the workplace and educational institutions,” which could be used by companies to discipline, rank, or micromanage employees; “social scoring,” a surveillance tool used in China to rate individuals on everyday activities and award (or deny) them social credit; “biometric categorization,” a practice that uses characteristics such as skin tone or facial structure to infer gender, sexual orientation, or even the likelihood of committing a crime; and “some cases of predictive policing for individuals,” which has already been shown to have racially discriminatory impacts (Harris, 2023c).

The EU AI Act also regulates “General-Purpose AI systems” (GPAI), which have great potential to be used for both good and harm. Though the law only applies directly to AI used in Europe, it has significant extraterritorial implications, in that AI developers doing business in Europe, even if not based there, will have to comply with aspects of the law in ways that are likely to shape their business practices and products around the world. This includes provisions that require AI developers to produce risk assessments of their AI systems, to take precautions to mitigate those risks, and to share information about their energy consumption.

However, even strong proposals from the world’s leading countries may not be enough to guarantee the safe development of AI technology globally. Relying on the “Brussels effect,” a term for the phenomenon whereby regulations adopted in Europe naturally become adopted as *de facto* global standards, may not be enough to ensure the safe proliferation of AI (Bradford, 2020). However, given the scale of global business, political, and individual self-interests at play, the example set by the EU AI Act may not have the widespread impact that the Brussels effect implies. One category of harm in particular that cannot be stopped by a weak patchwork of laws is the misuse of powerful AI systems by “bad actors,” ranging from vengeful individuals creating devastating NCII to scammers cloning the voices of victims’ relatives, to intelligence agencies and digital mercenaries using AI for coordinated manipulation and misinformation efforts to interference in elections to developing lethal autonomous weapons (LAWs), to the deployment of powerful cyberweapons.

Absent any meaningful regulation, companies including Meta, Stability AI, Hugging Face, Mistral AI, EleutherAI, and the Technology Innovation Institute have chosen to distribute their AI systems in ways that can be easily misused by bad actors. These companies are effectively racing against each other to release ever more powerful “open” and “open-source” AI systems, which we refer to herein as “unsecured” AI systems to signify that their model weights have been publicly released in ways that facilitate repurposing—which can in some cases be good but can also impose significant risks on society (Harris, 2023d). While these unsecured models are not the only models susceptible to misuse, online interfaces to secured AI (systems offered through hosted web or API interfaces) offer opportunities to stop bad actors from accessing and abusing high-risk AI on a large scale. While hackers have found ways to circumvent safety features of secured AI systems, the developers of these systems are able to patch vulnerabilities once discovered and also limit the rates of usage of their systems. These types of security interventions are not possible for unsecured AI systems, which can be downloaded, fine-tuned to facilitate abuse, and run in secret. No security measure is ever perfect, as hackers and spies could potentially steal model weights of secured systems, but this should not be taken to mean that security measures ought to

be abandoned completely. Regulation will only become more crucial as time passes, with higher capability models from both open and closed sourcing becoming even more risky if used by malign actors (Seger et al., 2023, p.12).

One telling illustration of these vulnerabilities can be seen in an announcement from Microsoft and OpenAI that hackers from China, Russia, Iran, and North Korea were caught using their AI systems to improve their cyber-attack techniques. Once caught, these companies are able to block these users and develop more sophisticated ways of detecting abuses based on their usage patterns. The same threat actors, however, could very easily migrate their efforts to the unsecured systems provided for download by the companies listed in the paragraph above and likely never be detected or stopped from abusing them (Satter, 2024).

Today, this particular set of companies releasing unsecured AI systems are based in the United States, UK, Europe, and the United Arab Emirates. Of these companies, the EU AI Act applies most directly to Mistral AI, which is headquartered in France. Even though the EU AI Act has yet to come into force, it appears to have impacted the company's release strategy. The company, co-founded by France's former Digital Minister, Cédric O, fiercely resisted the EU AI Act's regulations on GPAI and sought to add specific exemptions for "open-source" AI to the Act (Chatterjee & Volpicelli, 2023; Wanat, 2023). Only a few months after the EU AI Act was confirmed to apply to GPAI with only limited exemptions for less powerful "open-source" models, the company announced a partnership to distribute a new, secured model, "Mistral Large," in partnership with Microsoft (Leprince-Ringuet, 2024). However, with the weak patchwork of regulations that we have today, a company in this position could simply choose to relocate to a more permissive jurisdiction and continue to develop unsecured AI systems, albeit without commercial access to the European Union market.

What we see here is a perfect illustration of the "pacing problem," where the pace of technology development outstrips the pace of governmental regulation (Downes, 2009). This term has been applied to a wide variety of technical fields, including stem cell research, genetic testing, synthetic biology, nanotechnology, neuroscience, driverless cars, human cloning, and geoeengineering (Kuokkanen & Yamineva, 2013; Marchant, 2011), though the pace of AI development today seems to be even faster than these other examples.

### ***1.3 Enter philanthropy***

Faced with such an impasse, international networks of academics have begun working with civil society organizations such as the Center for AI and Digital Policy, Amnesty International, the Algorithmic Justice League, the Center for the Advancement of Trustworthy AI, the Centre for International Governance Innovation, Access Now, the Future of Life Institute, and many others, in an effort to advance AI governance efforts around the world (AI Ethicist, n.d.; Belfield, 2020, p.16).

Organized philanthropy, used here to refer to charitable foundations and their grantmaking initiatives, has begun to commit significant financial resources to these efforts. Based on our review of publicly available information from the Foundation Directory (Candid, 2024), we estimate that more than \$300 million has been granted by U.S. private foundations to AI programs between 2018 and 2023, with roughly a third of that total going to AI governance and policy efforts.<sup>3</sup>

Strategic philanthropists today have an outsized opportunity to bridge the gap between the runaway speed of AI technological development and the multi-layered deliberative processes that could take years—or even decades—before a stable intergovernmental AI regulatory body is established.

We argue herein for large-scale philanthropic investment to create a new Interim International AI Institution (IIAI, pronounced “aye-aye”) that could act as a stopgap measure to facilitate international collaboration on AI governance. With organized philanthropy support, funds could be allocated to establish this prototype AI governance body immediately, without waiting for the lengthy process of securing commitments from governments around the world. An institution like this one, if thoughtfully constructed, could help build the capacity for international collaboration on AI governance and support policymakers around the world looking to collaborate with one another already—in the absence of such an institution. This institutional prototype would bring together policy experts, social scientists, and AI researchers with diverse disciplinary backgrounds, facilitating the attainment of technical and policy consensus where possible and clearly identifying areas where compromise will be necessary.

Fortunately, there are historical models of international agreements and related bodies from other industries that have been studied as models for AI governance, including the International Civil Aviation Organization (Trager et al. 2023, p.19), the International Atomic Energy Agency (Nichols 2023), the Intergovernmental Panel on Climate Change (Suleyman & Schmidt, 2023), the Financial Action Task Force, and the International Telecommunications Union (Ho et al., 2023, p.9).

Some scholars describe philanthropic resources as society’s “risk capital”—the funds held by wealthy individuals or their foundations that can often be deployed quickly and with greater risk tolerance than government investments (Bosire, 2020; Buck, 2020). Philanthropic wealth is never apolitical, and it will be critical to put checks on the power of any funders contributing to the IIAI. Global civil society—of which private philanthropy is a part (McGuigan & Bass, 2022)—has an important role to play in shaping the global governance of AI and should be called upon to guide the path forward by advising philanthropists on the investments recommended here.

## **2 Historical antecedent: Ted Turner’s billion-dollar United Nations Gift**

The idea of strategic philanthropic engagement in support of international governance institutions is not new. The most salient and perhaps inspiring case study is Ted Turner’s decision to donate \$1 billion to the United Nations in 1997. Though this donation was mostly known for its historic size, it also had important structural implications for UN funding more broadly (United Nations, 2006). In 1997, the UN faced criticism of its administrative spending, which ultimately led it to reform its operations the following year (The New York Times, 1997; U.S. Department of State, 1997). Leading the charge for these critiques, the U.S. government began to withhold over \$1 billion in membership dues in protest of the UN’s inefficiencies (Crossette, 1997).

In the face of these criticisms, Ted Turner announced his \$1 billion donation, originally intended to “erase the debt” the U.S. government owed to the UN. Turner intended to allow the UN to continue operating specific programs frozen due to lack of funding (Turner, 1997). By stepping in when government funding was falling short, Turner demonstrated that intergovernmental initiatives, particularly the UN, could be funded by more than just government bodies in a way that had positive geopolitical ramifications. Moreover, the realized effects of Turner’s donation also helped to facilitate further civil society and private sector engagement with the UN in the following years.

With his donation, Turner created the United Nations Foundation (UNF), a separate organization founded with an initial objective of funneling the large donation to “UN causes” (CNN, 1997). Kofi Annan, former UN Secretary-General, credited the founding of the UNF as the moment that the UN became a “partnership organization,” an important transition for the organization that greatly increased its capabilities (United Nations, 2006). By acting as a funding body for the UN,

the UNF created a mechanism for companies and individuals to donate money to UN causes, which has since become an important part of the UN's sustainable development efforts (United Nations Foundation, 2023). The UNF has also inspired and directly channeled hundreds of millions of dollars in additional donations from individuals, corporations, governments, and NGOs to UN agencies (United Nations, 2006). In this way, the UNF serves as both a legal and financial mechanism and could be seen as a historical antecedent for the proposed IIAI, demonstrative of the potential for philanthropic contributions to meaningfully change the course of international governance institutions.

Of course, there are also reasons to be skeptical of such mechanisms—ideally, the United Nations would never have found itself in the position it did in 1997, when the United States defaulted on its debt. But at this particular historical moment, Turner's intervention positively impacted the organization and leveraged additional resources. The UN took steps, however, to make sure that the application of UNF contributions would be carefully governed by UN officials. This takes place through the United Nations Fund for International Partnerships (UNFIP), which today sits within the UN Office for Partnerships, as well as the UN-UNF Joint Coordination Committee. UNFIP's Advisory Board is chaired by the Deputy Secretary-General (United Nations Foundation, n.d.). The careful structuring of this relationship could be studied in support of the design of a similar mechanism that might tie the IIAI to existing UN bodies or facilitate a planned transition to such a relationship in the future. Depending on the direction of UN efforts on AI governance, the IIAI as an institution could, pending approval of UN leadership, potentially even be donated as a whole to the UN via the UNF and UNFIP.

The longstanding relationship between the Bill and Melinda Gates Foundation and the World Health Organization (WHO) is another example of a major philanthropic contributor dramatically increasing the capacity of an intergovernmental institution. According to Euronews, "In 2018-2019, the United States was the largest donor at \$893 million, accounting for around 15 percent of WHO's budget. The Gates Foundation came only second, with \$531 million" (Carbonaro, 2023). Based on information published on the Foundation's own website, its contribution to the WHO from 1998 to 2020 appears to be well over US 5 billion, not adjusted for inflation (Bill & Melinda Gates Foundation, n.d.).

When it comes to the governance of AI, the situation is different. There is currently no intergovernmental body that plays a role in AI governance. This presents a once-in-a-generation opportunity for one or more ambitious philanthropists to step up and use their financial resources to bridge the gap between the speed of AI governance and the speed of AI technology development.

### **3 How philanthropy is positioned to help**

Philanthropy—along with civil society, policymakers, and industry—has an urgent opportunity to push AI to fulfill its potential to deliver enormous benefits to society. This would be far from the first time that philanthropy has bridged the gap between public and private interests, with core involvement in fields ranging from climate change (Betsill et al., 2022) to access to COVID-19 vaccines (Banco et al., 2022) to nuclear nonproliferation (Rubinson, 2021). None of these efforts have transpired, however, without controversy (Morena 2023; Sklair & Gilbert 2022; Sparke & Levy, 2022), and it is critical that philanthropists interested in boldly funding AI governance efforts learn from both the successes and failures of these past efforts.

Notably, philanthropy even played a significant role in the birth of AI itself. In 1956, the Rockefeller Foundation gave a grant to support the Dartmouth Conference, a five-week-long gathering of researchers that has now become known as the birthplace of the modern notion of "artificial

intelligence.” In what might be one of the most historically significant grant proposals of all time, the proposal for the Dartmouth Conference is the site of the first documented use of the term “artificial intelligence” (Rockefeller Archive Center, 2022). Today, we come full circle, where it is time for organized philanthropy to once again play a critical role in shaping the field of AI.

Reflecting on both current and historical examples, we propose here a three-part framework for understanding how philanthropy can approach supporting the development of international AI governance in a manner that is inclusive and participatory; provides critical capital and support for innovation, speed, and risk in public policy approaches; and supports the creation and sustenance of institutional infrastructure that can increase capacity and resilience in the digital technology ecosystem.

### ***3.1 Ensure inclusive representation and participation of civil society***

One of philanthropy’s best contributions has been to build the capacity of diverse, expert, and timely coalitions in the face of major technological change, both in order to ensure that technology’s benefits are distributed and democratic and to mitigate its harms (Slaughter & Walker 2021). We are simply asking philanthropy to do this once more.

In today’s technology and media industries, a small number of actors play an outsized role in decision-making and value capture (Hutchinson, 2022; Moore & Tambini, 2022). This leads to a massive asymmetry in who benefits from technology and who bears the costs of both targeted harms (e.g., scams and fraud, biased decisions) and diffuse harms (e.g., disinformation, mental health, etc.) (Alvesalo et al., 2022; Robinson & Edwards, 2024; Thakur & Hankerson, 2021). The starting point for philanthropy should be that the power, value, and decision-making of technology cannot be concentrated in the hands of a few, be they corporations, governments, or the wealthy.

One key philanthropic strategy in this domain is to ensure that individuals with diverse lived experiences and perspectives actively shape the design, deployment, monitoring, and impact of AI. Philanthropy has played a critical role in ensuring that such individuals—and the organizations they lead or are affiliated with—are capitalized to make such contributions. In the United States, for instance, strategic philanthropists over the past several decades have tried to ensure that diverse champions, organizations, and coalitions with expertise in emerging technology and their surrounding governance and accountability ecosystems are funded and supported. One example of this type of work is the Rockefeller Foundation’s grant of \$300,000 to Black in AI, a technology research organization, to enhance the representation of Black individuals in the field of artificial intelligence. This initiative aims to cultivate a new network of Black scholars and engineers while combating bias in AI (The Rockefeller Foundation 2022). Specifically, part of the funds were allocated toward addressing issues of discrimination toward people of color in AI facial recognition technology. Indeed, what might be broadly called the responsible technology (or public interest technology) movement has seen evolving battlegrounds from net neutrality to open data and the right to information, to privacy and data governance, to competition and antitrust, and even trust and safety (Omidyar Network, 2022).

AI discourse today is colored by the ongoing sparring between what is colloquially known as “AI Ethics” and the “Effective Altruism” or “X-risk” movements; a set of (sometimes divergent) foundations and networks have undergirded each field (Arcas, 2023). We agree with calls for a recognition on the part of these divergent communities that their struggles are bound up with one another and that solving AI’s present-day ethical harms is in many ways a precondition for addressing AI’s longer-term risks (Kubzansky, 2024).

Equally important is the role that philanthropy can and has played in bringing together various interest groups—e.g., labor groups, civil rights and justice, privacy and data governance, even climate—to be represented at the negotiating table on critical issues. Too often, the rapid advance of technology, including but not limited to AI, only reveals its potential to cause serious and widespread harm. These harms are often disproportionately borne by some of the aforementioned interest groups who are not usually considered technology stakeholders. By learning from past errors and supporting these key constituencies, philanthropy can play an important role in staving off the downsides of technological advances and fostering broad-based coalitions that work together toward the collective good.

Yet another way for philanthropy to play an important role in building inclusivity is in ensuring that the responsible tech ecosystem has the capacity to speak out quickly as an early warning system for unforeseen harms and unintended consequences. This capacity was visibly evident when ChatGPT broke consumer adoption records in 2022–2023 (Hu, 2023). Since academic researchers and civil society organizations had been studying AI bias and fairness with philanthropic support for years before the arrival of ChatGPT (Partnership on AI, 2020, p.9), they were prepared to make recommendations on what to measure, how to understand progress and warnings, and where to invest in disclosure and oversight—from audits, to red-teaming, to potential licensing and accountability metrics and authorities. Philanthropy has long supported experimentation and innovation in many of these areas, with the privilege of having both a longer-term time horizon than most other institutions and the ability to take risks and move quickly if desired when needed.

In many ways, philanthropy fills in the gaps when trying to establish some parity between the big power players (corporations and governments) and those trying to raise the clarion call for better and improved stewardship (Ford Foundation, 2023). Underlying all of these goals is a critical insight into the role of philanthropy: the entire project is, in a sense, about ensuring that the digital technology ecosystem centers humanity and societal well-being. This means embedding both long-term thinking and rapid-response capabilities that can be resilient and maneuverable to match the speed and evolution of any technology. Such an ecosystem would have characteristics that allow it to respond to both structural issues (e.g., privacy and data governance, trust, and safety, etc., which continue to be governed by foundational principles) and emergent issues (e.g., confabulation and synthetic child sexual abuse material (CSAM), which may be unsavory or illegal attributes of a particular technological advancement).

### ***3.2 Providing risk capital for public policy development***

In philanthropic circles, the notion of supporting “experimental pilots” of solutions to social problems appears to be gaining popularity (Burton D. Morgan Foundation, 2024; MacArthur Foundation, 2024). Philanthropists are also increasingly looking to the term “catalytic capital” as an expression of how they can take risks to fund projects that catalyze deeper societal change rather than addressing problems at a superficial level (Schwartz, 2024). This has not always been the case. There is wide agreement (Knapp, 2023; Law, 2023) that the mixing of innovation and philanthropy was led by technology and media industry leaders such as Laurene Powell Jobs, Ted Turner, Melinda and Bill Gates, Pam and Pierre Omidyar, and the X Prize Foundation.

In the field of AI, there is an opportunity to double down on these types of “high-risk, high-reward” projects (Buck, 2020). A logical extension of these approaches is to test novel public policy approaches to rapidly advancing technologies like generative AI.

Funding projects designed to build civil society “go-to” options for model evaluation and improvement tools will go a long way to ensure that we scale the best, most effective policy



solutions possible. These initiatives have included testing AI's resistance to adversaries through "red-teaming" exercises, using algorithmic audits to evaluate bias and boost fair decision-making, or integrating public rating systems for large language models. Data and Society's Algorithmic Impact Methods Lab and Dr. Rumman Chowdhury's work on Humane Intelligence are examples of these efforts. Similar and important work can be done to ensure that we support research in emergent areas, for example: How will we ensure that technology is used to augment human capabilities and not replace them? How will our relationships evolve in the age of intelligent and communicative machines?

Another role of philanthropy is to provide funding to leverage and publish lessons from both successes and failures to advance the public conversation about what works and why. For instance, in areas such as trust and safety, philanthropy has supported associations (Integrity Institute, 2023) of fellows to ensure knowledge sharing and the development of best practices, supported journals (*Journal of Online Trust and Safety*, 2023) that supply innovation and pilot results, and created fora for sharing knowledge and create guideposts and diagnostic tools for better outcomes (Stanford Internet Observatory, 2023).

Finally, philanthropy can help to accelerate consistent global standards and inter-party negotiations on policies and regulations. Bringing harmonization and interoperability to the approach that governments around the world are taking to AI is critical, and here the philanthropic sector can speed up ongoing diplomatic work. Consider, for example, organized philanthropy's funding of the UN Secretary-General's High-level Advisory Body on AI (Advisory Body on Artificial Intelligence, 2023, p.25). This Body brings together a cross-disciplinary group of 38 global experts from all world regions to offer diverse perspectives and options on how to govern AI for humanity, including in support of the UN's Sustainable Development Goals. Philanthropy's fast-turnaround support for efforts like this allows for action to be taken without the delays involved in awaiting member state contributions. This type of funding has the potential to ensure that the public sector is not caught flat-footed on issues like AI governance that require global coordination and rapid action.

### ***3.3 Supporting shared infrastructure and new institutions***

Digital technology continues to advance faster than laws, regulations, policies, market incentives, and societal norms can keep up (Marchant, 2011). As a result, it has become an ongoing struggle to address both narrow and widespread harms, usually long after they have taken a heavy toll. As noted above, philanthropy can serve to bridge the lessons of the past with the emerging needs of today, as they have done to assist policymakers, regulators, and business leaders for other complex and multipurpose technologies, including nuclear technology and biomedicine (Toma, 2022).

This will require philanthropy to invest in the creation and sustenance of new governance infrastructure. Examples of this include funding new capacities (e.g., governmental ability to respond to threats through research, mobilization, and advocacy) and human capital pathways (so that there is a threshold of dedicated expertise that can work in the field), as well as new fields of research and inquiry and new organizational homes. Philanthropy has done this before in areas as diverse as public interest technology and AI ethics (Ford Foundation, 2023), impact investing (The Rockefeller Foundation, 2021), and drug and vaccine delivery (NBC News, 2005).

In AI, examples of such infrastructure already in development in civil society are the Center for the Advancement of Trustworthy AI, which focuses on providing governments with turnkey tools, training, consulting, and best practices for AI regulation, and the Distributed AI Research Institute, which prioritizes independent, community-based research. In the public sector, philanthropy has

supported efforts to consider the provision of AI technology in the public interest through proposals such as the National AI Research Resource and CalCompute.

The pace and scale of AI's progress demand the creation of a new ecosystem of institutions in both civil society and the public sector. These new institutions should bring about greater accountability for AI harms and diffuse the concentration of power and expertise in the hands of the tech companies and venture capitalists driving much of AI development and decision-making today. Philanthropy has done this before in the domain of technology, in areas such as open-source ecosystems and digital public infrastructure (White, 2023). A key outcome for such efforts is to ensure that new bodies and institutions have a harmonious relationship with existing authorities and serve to both assist them (e.g., with nimble research or rapid-response trials or with diverse perspectives or community engagement) and hold them accountable to the public interest.

Philanthropy can help leaders and authorities incentivize collaboration and mechanisms that lead to world-class progress. The way forward will require both government and industry to establish well-designed accountability systems with appropriate guardrails and checks and balances to prevent serious harm, establish liability, create public alternatives, and provide remedies and redress. Philanthropy's substantive engagement, done well, should accelerate a more equitable technology ecosystem.

#### **4 Racing against the race to the bottom**

The first public report of the United Nations Secretary-General's High-Level Advisory Body on AI, "Governing AI for Humanity," notes that while there are numerous options upon which a new intergovernmental body designed to govern AI could be modeled, there is no clear and obvious choice (Advisory Body on Artificial Intelligence, 2023, p.13). The Advisory Body identifies in this report seven needed "AI Governance Functions," needed in any future institutions. The Member States of the United Nations are currently negotiating a Global Digital Compact, described in a recently released "Zero Draft" to have the objective of "Govern[ing] emerging technologies, including Artificial Intelligence, for humanity," and whose adoption is expected in September 2024 (United Nations, 2024, p.1). The Zero Draft also calls for the creation of an "International Scientific Panel on AI" that appears to be at least partly modeled on the Intergovernmental Panel on Climate Change, as well as a "Global Fund for AI and Emerging Technologies for Sustainable Development."

Through these documents, it is clear that the UN sees a need for the rapid launch of institutions capable of monitoring risks and opportunities, rendering governance and technical standards interoperable, and harnessing AI for the public interest, all this with a focus on the UN Sustainable Development Goals (SDGs). In email communications with UN officials involved in this effort, we learned that the intent is for this work to be initially supported by a small team that will transition into a full-fledged UN AI Office by the end of 2025.

Under normal circumstances, this might appear to be a reasonable timeline for global action. However, for those of us concerned with the urgency of the pacing problem and the fact that each day of delay in establishing a governing body perpetuates a "race to the bottom" in rapidly deploying AI systems, it seems woefully inadequate.

A particularly striking reminder of the "race to the bottom" was a recent study that revealed that a version of Stable Diffusion, one of the most popular "open-source" AI image-generating tools, had been trained with thousands of CSAM images, presumably allowing it to be effective at generating AI-Generated Child Sexual Abuse Material (AIGCSAM) (Thiel, 2023). Another recent study found that AI "undressing" tools, which can turn an image of a clothed person (although many

of the tools only work on women) into NCII, are gaining popularity, with at least 34 such tools widely available online for public use (Lakatos, 2023). The growing availability of “open-source” AI image-generation tools has been cited as a key factor in enabling the creation of these undressing tools. At a high school in New Jersey, dozens of pornographic images of female students were created, with devastating effects on the children involved (CHILD USA, 2023).

If only a few jurisdictions worldwide take action against AI CSAM and NCII generators, it will not be effective. AI developers in unregulated locales will simply continue to build these tools unless they are held accountable (Kalia et al., 2024). If the makers of popular tools are careless in their production and those tools are trained on illegal CSAM images, they should be held accountable as well. The proliferation of AI systems that either intentionally or recklessly provide people with tools that can be wielded with devastating consequences should be a crime. However, it is not yet clear today whether the makers of these tools will, or even can, be held liable under existing law.

## **5 The big bet: The Interim International AI Institution (IIAI)**

Lawmakers are scrambling to address a multitude of urgent AI harms like those discussed above in numerous jurisdictions, but without a centralized, international body to coordinate and support these efforts, they are likely to take far too long to be effective against these already occurring harms and those on the immediate horizon (Nilsson, 2017).

The project of governing AI needs to unfold in dialogue with existing frameworks, such as UNESCO’s global agreement on AI ethics and the G7’s Hiroshima AI Process. While these frameworks are acknowledged in the UN interim report, it is necessary to clarify how any new institutional function would interact with them. This process requires the active participation of numerous intergovernmental organizations. Through close communication and collaboration, a new organization will need to work quickly to “harmonize standards, safety, and risk management frameworks” (Advisory Body on Artificial Intelligence, 2023, p.16). Thus, there is an opportunity for philanthropic big bets. Leveraging the unique power of strategic philanthropists discussed above, there is an opportunity to work alongside the creation of the emerging UN AI Office in a way that follows the principles discussed above, as well as those already enumerated in the UN AI Advisory Body’s Interim Report.

To move as quickly as possible against the race to the bottom and toward maximum common benefit, we propose that one or more philanthropists make a gift in the tens of millions of dollars to fund the first three years of operation of the Interim International AI Institution (IIAI), so that even if the creation of the UN’s AI Office is delayed significantly, this body can operate in its absence. At the end of this initial three-year period, or hopefully sooner, the IIAI and its assets could be placed under UN ownership.

The term “Interim” is important—it signals humility through the understanding that such a body must have a democratic mandate to be successful in the long run. The choice of the term “Institution” is also significant in its flexibility—it acknowledges that it could evolve in the future into an agency or body of another institution or an independent organization—again in preparation for possible unforeseen challenges within the UN system. It would bring together the highest level of technical and policy expertise in the service of rapid-response regulatory development. A key goal during this period would be to facilitate strong legal standards and enforcement mechanisms to bring AI under democratic control. Tasks that the IIAI could undertake would include three key areas.

### **5.1 *Rapid-response harmonized policy development support***

The IIAI's first priority would be to provide rapid-response support to regulatory or legislative bodies anywhere in the world working to develop AI regulation. By assisting national governments in the development of AI policy, the institution can get a head start on the process of harmonizing—to the greatest extent possible—AI regulations around the globe. This could include legal, technical, translation, communications, and even legislative strategy support—everything that regulators need to ensure that their AI laws are robust and iterative, passed as quickly as possible, and harmonized with other jurisdictions (and eventually able to integrate with an international legal regime) at the state, national, or international level. Even in a place like California, we have seen firsthand that legislators lack access to the combination of skilled technologists and policy experts who understand the legal and multidisciplinary issues at hand. We have seen this leads to confusion, overreach, or duplication of effort that unnecessarily hinders democratic oversight via the regulatory process. Some examples include:

- Prototype a global licensing and registration standard for AI systems that defines all of the details of how such a system would be implemented and then makes it accessible, as soon as possible, so that when such a system becomes legally binding in certain jurisdictions, it is already available and tested;
- Develop best-practice monitoring mechanisms to detect AI harms and risks. The use of these monitoring mechanisms can then be mandated by regulation. This could include detecting AI-powered misinformation campaigns and influence operations taking place on social media and monitoring the proliferation of unlicensed AI systems that could pose harm; monitoring of the “dark web” to find AI systems that can be used to produce NCII or CSAM; or building tools to detect content from AI systems that do not use required watermarking standards;
- Build a global version of the planned U.S. National AI Research Resource (NAIRR), both by acquiring technology capacity directly in the public sector and also requiring that AI companies reserve a certain amount of their capacity for use by vetted researchers at universities and research institutions. This parallel approach ensures that we build toward a “public option” that democratizes data, compute, and access AI resources without bankrupting public sector institutions;
- Provide technical support for the development of key requirements, standards, and best practices so that AI developers are not left uncertain about the expectations for how they can develop AI systems responsibly. This could include such examples as:
  - Guidelines for assessing the human rights impacts of AI systems, including but not limited to the fairness of and potential discrimination caused by these systems in which these systems could violate privacy;
  - Guidelines for how to transparently disclose what training data was used to produce AI systems and how to make sure that this training data was ethically and legally sourced;
  - Guidelines for how to conduct adversarial testing or red-teaming of AI systems before they are released to the public;
  - Guidelines for understanding the potential harms of AI systems across different applications.

### **5.2 *Support collaborative standards development***

We are currently seeing a proliferation of efforts to establish standards around AI, with different national and international bodies operating independently, with examples including the U.S.

National Institute for Standards and Technology (NIST), the EU's CEN-CENELEC, the International Organization for Standardization, and the Institute of Electrical and Electronics Engineers (IEEE). A technically sophisticated coordinating body is needed to facilitate collaboration between these institutions to accelerate the development of AI standards, such as a maximally indelible watermarking technique (a way for generative AI systems to mark the content they produce so that it is difficult to remove, and can be decoded and displayed wherever the content is viewed). Standards like this—or common language to set up requirements for industry best practices—could be developed in-house or through coordination between existing groups that are working on similar standards. In the case of watermarking, this could mean bridging the gap between the Adobe-led Coalition for Content Provenance and Authenticity (C2PA) and efforts already in place in China.

### ***5.3 Support coordinated public interest AI development***

In collaboration with universities and research centered around the world, the IIAII could support the coordination of world-class research and help to productively direct the allocation of AI resources to solve scientific problems in the public interest. This could include climate change, misinformation, public health, energy, agriculture, and many other fields. The UN SDGs present a strong set of guiding goals that could be used to focus these efforts.

The IIAII could also support increased access to AI education, offering public educational resources, including online courses taught by IIAII staff members working on all of the items above and designed to democratize access to AI knowledge and spread its benefits widely.

### ***5.4 Succession planning***

The final duty of the IIAII is long-term succession planning for the institution itself. This would involve regular documentation of successes and failures in its efforts at the above goals, such that it can make recommendations based on its own experience of what its successor institution(s) should look like. It should also strive to operate flexibly to prepare for that transition, using UN principles and working practices to guide its efforts and smooth a possible transition into the UN system.

While it may eventually make sense to have these types of activities in different institutional homes, there is value in having a broad set of such activities under one roof at the outset, as it has the potential to spark synergies among a multidisciplinary group of people who would otherwise have to fly across the world to connect and exchange ideas face-to-face.

It is our contention that the combined cost of these activities—including the convening, travel, and maintenance of a high-level team of 50–60 experts working fulltime in the same location for three years—should be in the tens of millions of dollars, easily funded by the philanthropic community already backing AI initiatives in this vein.

## **6 Conclusion**

Philanthropic capital is uniquely positioned to make this investment, given its ability to make fast decisions and take bold risks in the public interest. This moment represents a unique leverage point, where a vacuum has emerged and technological advances are far outpacing the pace of regulatory action.

The recent successful passage of the EU AI Act (Bertuzzi, 2024) is both a sign of good things to come and an indicator of the inadequately slow pace of government action, even in the most advanced democracy that has managed to pass meaningful AI regulation. Some of the provisions of the

EU AI Act will come into force after six months, but others will take between one and three years to take effect. Given the relative slowness of the U.S. Congress to regulate social media technology and online privacy compared to Europe, it is hard to imagine legislation that has not yet even been introduced in the U.S. Congress coming into force sooner than the EU AI Act (Pearlstein, 2023).

The longer we wait to regulate AI, the more people will be harmed by both careless and malicious use of AI tools. As discussed in the introduction, some of the often irreparable harms caused by unregulated AI systems include discrimination in lending, housing, employment, health care, and many other areas; the production of interactive misinformation, disinformation, and malinformation, election interference, and the production of non-consensual intimate imagery. For every month that we wait to regulate AI, more and more powerful unsecured AI tools proliferate, and it will be very difficult to ever get these tools out of circulation. Therefore, propelling regulation forward as quickly as possible is a critically time-sensitive effort.

Another reason that philanthropy may be the best, and perhaps only, option to fund such an endeavor is the competitive dynamics inherent to the relationship between adversarial nations. The EU AI Act notably created exemptions for national security and police use of many types of otherwise prohibited AI systems (Nolan et al., 2024). As with climate change, regulation, and nuclear nonproliferation agreements, many, if not most, countries do not want to be the first to make bold commitments to self-restraint without knowing that other nations will follow suit. This “regulation dilemma,” arguably more difficult to overcome than the prisoner’s dilemma (Han et al., 2021), is further evidence of the need for a global coordinating body to step in as soon as possible to drive forward global collaboration and progress on AI regulation.

Ted Turner’s billion-dollar donation to the United Nations is still remembered a quarter-century later. The creation of the IIAI is an opportunity to make an outsized impact at a critical inflection point for AI that will deliver broad societal benefits well into the next century.

### **Acknowledgments**

The authors would like to thank the research team from the University of California, Berkeley, that supported this work: Owen Doyle, Milad Brown, Anish Ganga, Nancy Hu, Ashley Chan, Ruiyi Chen, Maddy Cooper, Daniel Jang, Parth Shinde, and Vinaya Sivakumar. This work also benefited significantly from feedback from a senior United Nations official. For David Evan Harris, additional thanks are due to the Geneva Centre for Philanthropy at the University of Geneva, the Centre for International Governance Innovation, the International Computer Science Institute, and the John S. and James L. Knight Foundation for supporting the research that made this work possible. For Anamitra Deb, to Omidyar Network, where his position has allowed him to participate in such timely, responsible technology discussions.

### **Notes**

- 1 In this chapter, we refer to AI broadly, including the ranking and recommendation systems that power social media platforms, as well as generative AI. Specifically, we use here the OECD’s 2023 updated definition of AI, “An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment” (Russell et al., 2023).
- 2 A concerning example of this includes Chinese journalist Liu Hu, who was blocked from buying plane tickets, property, or taking loans. There was no formal government notice of the restrictions or an appeal process, with the restrictions being thought to be the result of Hu’s anti-corruption reporting and tweets against the government (Zhao, 2018).

- 3 Other categories that we assigned to grants included AI & medicine/science, AI ethics, AI & climate change, AI education, AI fairness & inclusion, and AI safety/existential risk. It is not appropriate to consider this an exhaustive review or anything more than a rough approximation of a floor on funding, due to potential significant gaps in the Foundation Directory database, delays in reporting, and the difficulty of assigning categories to grants that often blurred the lines between categories or fell into multiple categories. Corporate giving, when coming directly from a company, was also not included in these totals.

## References

- Advisory Body on Artificial Intelligence. 2023. "Interim Report: Governing AI for Humanity." *United Nations*. <https://www.un.org/ai-advisory-body>.
- AI Ethicist. n.d. "AI NGOs, Research Organizations, Ethical AI Organizations." *AI Ethicist*. February 3, 2024. <https://www.aiethicist.org/ai-organizations>.
- Akselrod, Olga, and Cody Venzke. 2023. "How Artificial Intelligence Might Prevent You From Getting Hired | ACLU." *American Civil Liberties Union* (blog). August 23, 2023. <https://www.aclu.org/news/racial-justice/how-artificial-intelligence-might-prevent-you-from-getting-hired>.
- Alvesalo, -Kuusi Anne, Hanna Maria Malik, Mika Viljanen, and Nea Lepinkainen. 2022. "Dynamics of Social Harms in an Algorithmic Context." *International Journal for Crime, Justice and Social Democracy* 11 (1): 182–195. <https://doi.org/10.5204/ijcjsd.2141>.
- Arcas, Blaise Agüeray. 2023. "Fears about AI's Existential Risk Are Overdone, Says a Group of Experts." *The Economist*. July 21, 2023. [https://www.economist.com/by-invitation/2023/07/21/fears-about-ais-existential-risk-are-overdone-says-a-group-of-experts?utm\\_medium=cpc.adword.pd&utm\\_source=google&ppccampaignID=17210591673&ppcadID=&utm\\_campaign=a.22brand\\_pmax&utm\\_content=conversion.direct-response.anonymous&gad\\_source=1&gclid=Cj0KCQiAgqGrBhDtARIsAM5s0\\_m1xOb1PttlYMwJuzoFvsQSAeG9cX-o1l-Jpf7fJpMRjBZfboG6wYkaAgdwEALw\\_wcB&gclsrc=aw.ds](https://www.economist.com/by-invitation/2023/07/21/fears-about-ais-existential-risk-are-overdone-says-a-group-of-experts?utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=17210591673&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gad_source=1&gclid=Cj0KCQiAgqGrBhDtARIsAM5s0_m1xOb1PttlYMwJuzoFvsQSAeG9cX-o1l-Jpf7fJpMRjBZfboG6wYkaAgdwEALw_wcB&gclsrc=aw.ds).
- Backman, Isabella. 2023. "Eliminating Racial Bias in Health Care AI: Expert Panel Offers Guidelines." *Yale School of Medicine*. December 21, 2023. <https://medicine.yale.edu/news-article/eliminating-racial-bias-in-health-care-ai-expert-panel-offers-guidelines/>.
- Banco, Erin, Ashleigh Furlong, and Pfahler Lennart. 2022. "How Bill Gates and Partners Used Their Clout to Control the Global Covid Response — With Little Oversight." *POLITICO*. September 14, 2022. <https://www.politico.com/news/2022/09/14/global-covid-pandemic-response-bill-gates-partners-00053969>.
- Belfield, Haydn. 2020. "Activism by the AI Community: Analysing Recent Achievements and Future Prospects." In *Proceedings of the AAI/ACM Conference on AI, Ethics, and Society*, 15–21. AIES '20. New York: Association for Computing Machinery. <https://doi.org/10.1145/3375627.3375814>.
- Bertuzzi, Luca. 2024. "EU Countries Give Crucial Nod to First-of-a-Kind Artificial Intelligence Law." *Euractiv*. February 2, 2024. <https://www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/>.
- Betsill, Michele M., Ashley Enrici, Elodie Le Cornu, and Rebecca L. Gruby. 2022. "Philanthropic Foundations as Agents of Environmental Governance: A Research Agenda." *Environmental Politics* 31 (4): 684–705. <https://doi.org/10.1080/09644016.2021.1955494>.
- Bill and Melinda Gates Foundation. n.d. "Committed Grants (Database)." February 3, 2024. <https://www.gatesfoundation.org/about/committed-grants>.
- Bosire, Lydia Kemunto. 2020. "How to Make Society's Risk Capital Riskier (SSIR)." July 29, 2020. [https://ssir.org/articles/entry/how\\_to\\_make\\_societys\\_risk\\_capital\\_riskier](https://ssir.org/articles/entry/how_to_make_societys_risk_capital_riskier).
- Bradford, Anu. 2020. *The Brussels Effect: How the European Union Rules the World*. New York: Oxford University Press. <https://scholarship.law.columbia.edu/books/232>.
- Buck, Stuart. 2020. "How Philanthropy Could Embrace More High-Risk, High-Reward Projects." *Inside Philanthropy*. September 9, 2020. <https://www.insidephilanthropy.com/home/2020/9/9/how-philanthropy-could-embrace-more-high-risk-high-reward-projects>.
- Burton D. Morgan Foundation. 2024. "Morgan Startup Grants." 2024. <https://www.bdmorganfdn.org/morgan-startup-grants>.
- Candid. 2024. "Foundation Directory." <https://fconline.foundationcenter.org/>.
- Carbonaro, Giulia. 2023. "Why Does WHO Rely so Much on Bill Gates' Money?" *Euronews*. March 2, 2023. <https://www.euronews.com/health/2023/02/03/how-is-the-world-health-organization-funded-and-why-does-it-rely-so-much-on-bill-gates>.

- Chan, Kelvin. 2023. "Europe Agreed on World-Leading AI Rules. Will They Affect People Everywhere?" *Associated Press*. December 11, 2023. <https://apnews.com/article/eu-ai-act-artificial-intelligence-regulation-0283a10a891a24703068edcae3d60deb>.
- Chatterjee, Mohar, and Gian Volpicelli. 2023. "France Bets Big on Open-Source AI." *Politico*. August 4, 2023. <https://www.politico.eu/article/open-source-artificial-intelligence-france-bets-big/>.
- Chee, Foo Yun. 2023. "Exclusive: AI Rules 'Cannot Be Bargained', EU's Breton Says after OpenAI CEO Threat | Reuters." *Reuters*. May 25, 2023. <https://www.reuters.com/technology/eus-breton-slams-openai-ceos-comments-blocs-draft-ai-rules-2023-05-25/>.
- CHILD USA. 2023. "AI-Generated CSAM: The Call for Proactive Action." *CHILD USA* (blog). November 21, 2023. <https://childusa.org/ai-generated-csam-the-call-for-proactive-action/>.
- CNN. 1997. "Ted Turner Donates \$1 Billion to 'U.N. Causes.'" October 18, 1997. <http://edition.cnn.com/US/9709/18/turner.gift/>
- Covington. 2023. "U.S. Artificial Intelligence Policy: Legislative and Regulatory Developments." October 20, 2023. <https://www.cov.com/en/news-and-insights/insights/2023/10/us-artificial-intelligence-policy-legislative-and-regulatory-developments>.
- Crossette, Barbara. 1997. "U.S. Effort to Cut Its Dues Dies at an Angry U.N." *The New York Times*. December 19, 1997, sec. World. <https://www.nytimes.com/1997/12/19/world/us-effort-to-cut-its-dues-dies-at-an-angry-un.html>.
- De Tena, Carlos Luca, Diego Rubio, Oscar Jonsson, Dario Garcia De Viedma, Carlos Lastra Anadon, Irene Menendez, Carl Benedikt Frey, and Christina J. Colclough. 2023. "Survey Reveals That 68% of Europeans Want Government Restrictions on AI." *IE University Center for the Governance of Change*. October 9, 2023. <https://www.ie.edu/cgc/news-and-events/news/european-tech-insights-2023/>.
- Downes, Larry. 2009. *The Laws of Disruption: Harnessing the New Forces That Govern Life and Business in the Digital Age*. 1st ed. New York: Basic Books.
- European Parliament. 2023. "Artificial Intelligence Act." June 14, 2023. [https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236\\_EN.pdf](https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf).
- Field, Hayden. 2023. "OpenAI's Sam Altman Reverses Threat to Cease European Operations." *CNBC*. May 26, 2023. <https://www.cbc.com/2023/05/26/openai-ceo-sam-altman-reverses-threat-to-cease-european-operations.html#:~:text=In%20just%20two%20days%2C%20OpenAI,%E2%80%9Cno%20plans%20to%20leave.%E2%80%9D>.
- Ford Foundation. 2023. "Philanthropies Launch New Initiative to Ensure AI Advances the Public Interest." November 1, 2023. <https://www.fordfoundation.org/news-and-stories/news-and-press/news/philanthropies-launch-new-initiative-to-ensure-ai-advances-the-public-interest/>.
- Grace, Katja, AI Impacts, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. 2024. "Thousands of AI Authors on the Future of AI." January. <https://arxiv.org/abs/2401.02843>.
- Han, The Anh, Francisco C. Santos, Luís Moniz Pereira, and Tom Lenaerts. 2021. "A Regulation Dilemma in Artificial Intelligence Development." In MIT Press. [https://doi.org/10.1162/isal\\_a\\_00385](https://doi.org/10.1162/isal_a_00385).
- Harris, David Evan. 2023a. "AI Is Already Causing Unintended Harm. What Happens When It Falls into the Wrong Hands?" *The Guardian*. June 16, 2023, sec. Technology. <https://www.theguardian.com/commentisfree/2023/jun/16/ai-new-laws-powerful-open-source-tools-meta>.
- Harris, David Evan. 2023b. "Biden's Executive Order on AI: Bravo, Now Back to Work." *Centre for International Governance Innovation*. November 8, 2023. <https://www.cigionline.org/articles/bidens-executive-order-on-ai-bravo-now-back-to-work/>.
- Harris, David Evan. 2023c. "Europe Has Made a Great Leap Forward in Regulating AI. Now the Rest of the World Must Step Up." *The Guardian*. December 13, 2023. <https://www.theguardian.com/commentisfree/2023/dec/13/europe-regulating-ai-artificial-intelligence-threat>.
- Harris, David Evan. 2023d. "How to Regulate Unsecured 'Open-Source' AI: No Exemptions." *TechPolicyPress*. December 4, 2023. <https://techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions>.
- Heath, Ryan. 2023. "China Races Ahead of U.S. on AI Regulation." *Axios*. May 8, 2023. <https://www.axios.com/2023/05/08/china-ai-regulation-race>.
- Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside. 2023. "An Overview of Catastrophic AI Risks." *arXiv*. <http://arxiv.org/abs/2306.12001>.
- Ho, Lewis, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, et al. 2023. "International Institutions for Advanced AI." *arXiv*. <https://doi.org/10.48550/arXiv.2307.04699>.



- Hu, Krystal. 2023. "ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note." *REUTERS*. February 2, 2023. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.
- Hutchinson, Christophe Samuel. 2022. "Potential Abuses of Dominance by Big Tech through Their Use of Big Data and AI." *Journal of Antitrust Enforcement* 10 (3): 443–468. <https://doi.org/10.1093/jaenfo/jnac004>.
- Integrity Institute. 2023. "Our Supporters." *Integrity Institute*. 2023. <https://integrityinstitute.org/our-supporters>.
- International Association of Privacy Professionals. 2023. "Global AI Legislation Tracker." September 2023. <https://iapp.org/resources/article/global-ai-legislation-tracker/>.
- Johnson, Khari. 2023. "The US Has Failed to Pass AI Regulation. New York City Is Stepping Up." *Wired*. October 19, 2023. <https://www.wired.com/story/us-failed-to-pass-ai-regulation-new-york-city-stepping-up/>.
- Journal of Online Trust and Safety*. 2023. "About the Journal." 2023. <https://tsjournal.org/index.php/jots/about>.
- Kalia, Shubham, Maria Ponnezhath, and Samrhitha A. 2024. "Musk Seeks Tesla Shareholder Vote on Moving Incorporation to Texas." February 1, 2024. <https://www.reuters.com/business/autos-transportation/tesla-take-shareholder-vote-shifting-incorporation-texas-musk-says-2024-02-01/>.
- Kang, Cecilia. 2023. "2 Senators Propose Bipartisan Framework for A.I. Laws." *The New York Times*. September 7, 2023. <https://www.nytimes.com/2023/09/07/technology/artificial-intelligence-framework-senate.html#:~:text=Framework%20for%20A.I.-,Laws,with%20the%20rapidly%20evolving%20technology>.
- Knapp, Alex. 2023. "New XPrize Will Award \$101 Million to Innovators Who Can Reverse Aging." *Forbes*. November 29, 2023. <https://www.forbes.com/sites/alexknapp/2023/11/29/new-xprize-will-award-101-million-to-innovators-who-can-reverse-aging/>.
- Kubzansky, Mike. 2024. "Addressing AI's Present and Future: A False Trichotomy? | Context." *Context*. May 7, 2024. <https://www.context.news/ai/opinion/addressing-ais-present-and-future-a-false-trichotomy>.
- Kuokkanen, Tuomas, and Yulia Yamineva. 2013. "Regulating Geoengineering in International Environmental Law." *Carbon & Climate Law Review* 7 (3): 161–167.
- Lakatos, Santiago. 2023. "A Revealing Picture – AI-Generated 'Undressing' Images Move from Niche Pornography Discussion Forums to a Scaled and Monetized Online Business." *Graphika*. <https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf>.
- Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica, November 23, 2016. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- Law, Tara. 2023. "Laurene Powell Jobs: Let's Tackle Climate Change 'Like a Speed Boat' Before The Window Closes." *Time Magazine*. April 25, 2023. <https://time.com/collection/earth-awards-2023/6274438/laurene-powell-jobs-climate-philanthropy/>.
- Leprince-Ringuet, Daphné. 2024. "Mistral's Deal with Microsoft Is Causing Controversy." *Sifted*. February 28, 2024. <https://sifted.eu/articles/mistral-microsoft-deal-controversy/>.
- MacArthur Foundation. 2024. "Grantee Profile: Lever for Change." 2024. <https://www.macfound.org/grantee/lever-for-change-10114942/>.
- Marchant, Gary E. 2011. "The Growing Gap between Emerging Technologies and the Law." In *The Growing Gap between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*, edited by Gary E. Marchant, Braden R. Allenby, and Joseph R. Herkert, 19–33. The International Library of Ethics, Law and Technology. Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-94-007-1356-7\\_2](https://doi.org/10.1007/978-94-007-1356-7_2).
- McCracken, Harry. 2023. "Microsoft's Satya Nadella Is Winning Big Tech's AI War. Here's How." *Fast Company*. August 21, 2023. <https://www.fastcompany.com/90931084/satya-nadella-microsoft-ai-frontrunner>.
- McGuigan, Elizabeth, and David Bass. 2022. "Policy Philanthropy and Its Key Role in Civil Society." *Philanthropy Roundtable*. <https://www.philanthropyroundtable.org/wp-content/uploads/2022/11/Policy-Philanthropy-and-Its-Key-Role-in-Civil-Society.pdf>.
- Mellwain, Charlton. 2020. "AI Has Exacerbated Racial Bias in Housing. Could It Help Eliminate It Instead?" *MIT Technology Review*. October 20, 2020. <https://www.technologyreview.com/2020/10/20/1009452/ai-has-exacerbated-racial-bias-in-housing-could-it-help-eliminate-it-instead/>.
- Moore, Martin, and Damian Tambini. 2022. *Regulating Big Tech: Policy Responses to Digital Dominance*. Oxford University Press.
- Morena, Edouard. 2023. "The Spirit of Climate Philanthropy." In *The Routledge International Handbook of Critical Philanthropy and Humanitarianism*, edited by Katharyne Mitchell and Polly Pallister-Wilkins. Taylor & Francis. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781003162711-21/spirit-climate-philanthropy-edouard-morena>.

- NBC News. 2005. "Gates' Foundation Donates \$750 Million." *NBC News*. January 25, 2005. <http://www.nbcnews.com/health/health-news/gates-foundation-donates-750-million-flna1C9440090>.
- Nichols, Michelle. 2023. "UN Chief Backs Idea of Global AI Watchdog Like Nuclear Agency." *Reuters*. June 12, 2023, sec. Technology. <https://www.reuters.com/technology/un-chief-backs-idea-global-ai-watchdog-like-nuclear-agency-2023-06-12/>.
- Nilsson, Adriana. 2017. "Making Norms to Tackle Global Challenges: The Role of Intergovernmental Organisations." *Research Policy* 46 (1): 171–181. <https://doi.org/10.1016/j.respol.2016.09.012>.
- Nolan, David, Hajira Maryam, and Michael Kleinman. 2024. "The Urgent But Difficult Task of Regulating Artificial Intelligence." *Amnesty International*. January 16, 2024. <https://www.amnesty.org/en/latest/campaigns/2024/01/the-urgent-but-difficult-task-of-regulating-artificial-intelligence/>.
- Omidyar Network. 2022. "Our Vision for a Responsible Tech Future." *Omidyar Network*. September 20, 2022. <https://omidyar.com/responsible-tech-future/>.
- O'Reilly, Thomas. 2023. "EU's AI Act Slammed as Innovation Killer." *The European Conservative*. December 14, 2023. <https://europeanconservative.com/articles/news/eus-ai-act-slammed-as-innovation-killer/>.
- Orth, Taylor. 2023. "Americans Are Divided on AI's Societal Impact, But Most Support Government Regulation." *YouGov*. May 25, 2023. [https://today.yougov.com/politics/articles/45747-americans-are-divided-artificial-intelligence-poll?redirect\\_from=%2Ftopics%2Fpolitics%2Farticles-reports%2F2023%2F05%2F25%2Famericans-are-divided-artificial-intelligence-poll](https://today.yougov.com/politics/articles/45747-americans-are-divided-artificial-intelligence-poll?redirect_from=%2Ftopics%2Fpolitics%2Farticles-reports%2F2023%2F05%2F25%2Famericans-are-divided-artificial-intelligence-poll).
- Partnership on AI. 2020. "2019 Annual Report: Building a Connected Community for Responsible AI." January 30, 2020. <https://partnershiponai.org/annual-report-2019/>
- Pearlstein, Steven. 2023. "Here's the Inside Story of How Congress Failed to Rein in Big Tech." *Washington Post*. July 6, 2023, sec. Opinions. <https://www.washingtonpost.com/opinions/2023/07/06/congress-facebook-google-amazon-apple-regulation-failure/>.
- Robinson, Jemima, and Matthew Edwards. 2024. "Fraudsters Target the Elderly: Behavioural Evidence from Randomised Controlled Scam-Baiting Experiments." *Security Journal*. January. <https://doi.org/10.1057/s41284-023-00410-4>.
- Rockefeller Archive Center. 2022. "'A Roomful of Brains': Early Advances in Computer Science and Artificial Intelligence." *REsource* (blog). January 6, 2022. <https://resource.rockarch.org/story/a-roomful-of-brains-early-advances-in-computer-science-and-artificial-intelligence/>.
- Rubinson, Paul. 2021. "Philanthropy, Nuclear Nonproliferation, and Threat Reduction." *Urban Institute*. [https://www.urban.org/sites/default/files/2021/02/05/philanthropy\\_nuclear\\_nonproliferation\\_and\\_threat\\_reduction.pdf](https://www.urban.org/sites/default/files/2021/02/05/philanthropy_nuclear_nonproliferation_and_threat_reduction.pdf).
- Russell Stuart, Karine Perset, and Marko Grobelnik. 2023. "Updates to the OECD's Definition of an AI System Explained." November 29, 2023. <https://oecd.ai/en/wonk/ai-system-definition-update>.
- Satter, Raphael. 2024. "Microsoft Says It Caught Hackers from China, Russia and Iran Using Its AI Tools." *Reuters*. February 14, 2024, sec. Cybersecurity. <https://www.reuters.com/technology/cybersecurity/microsoft-says-it-caught-hackers-china-russia-iran-using-its-ai-tools-2024-02-14/>.
- Schwartz, Debra. 2024. "Finding Hope in Catalytic Capital and Its Champions." *MacArthur Foundation* (blog). February 1, 2024. <https://www.macfound.org/press/perspectives/finding-hope-in-catalytic-capital-and-its-champions>.
- Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, et al. 2023. "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4596436>.
- Sheehan, Matt. 2023. "China's AI Regulations and How They Get Made." *Carnegie Endowment for International Peace*. July 10, 2023. <https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117>.
- Sklair, Jessica, and Paul Gilbert. 2022. "Giving as 'De-Risking': Philanthropy, Impact Investment and the Pandemic Response." *Public Anthropologist* 4 (1): 51–77. <https://doi.org/10.1163/25891715-bja-10033>.
- Slaughter, Anne-Marie, and Darren Walker. 2021. "Afterword." In *Power to the Public: The Promise of Public Interest Technology*, edited by Tara Dawson McGuinness and Hana Schank, 141. Princeton, NJ: Princeton University Press.
- Sparke, Matthew, and Orly Levy. 2022. "Competing Responses to Global Inequalities in Access to COVID Vaccines: Vaccine Diplomacy and Vaccine Charity Versus Vaccine Liberty." *Clinical Infectious Diseases* 75 (Supplement\_1): S86–S92. <https://doi.org/10.1093/cid/ciac361>.

- Stanford Internet Observatory. 2023. "Trust & Safety Research Conference Showcases Leading Research in Preventing Online Harms." October 25, 2023. <https://cyber.fsi.stanford.edu/io/news/trust-safety-research-conference-showcases-leading-research-preventing-online-harms>.
- Suleyman, Mustafa, and Eric Schmidt. 2023. "We Need an AI Equivalent of the IPCC." October 19, 2023. <https://www.ft.com/content/d84e91d0-ac74-4946-a21f-5f82eb4f1d2d>.
- Thakur, Dhanaraj, and DeVan L. Hankerson. 2021. "Facts and Their Discontents: A Research Agenda for Online Disinformation, Race, and Gender." *Center for Democracy & Technology*. <https://cdt.org/insights/facts-and-their-discontents-a-research-agenda-for-online-disinformation-race-and-gender/>.
- The New York Times. 1997. "A Start on U.N. Reform." *The New York Times*. March 18, 1997, sec. Opinion. <https://www.nytimes.com/1997/03/18/opinion/a-start-on-un-reform.html>.
- The Rockefeller Foundation. 2021. "Catalytic Capital Consortium Announces \$1.2 Million to Leading Impact Investing Networks." *The Rockefeller Foundation* (blog). July 28, 2021. <https://www.rockefellerfoundation.org/news/catalytic-capital-consortium-announces-1-2-million-to-leading-impact-investing-networks/>.
- The Rockefeller Foundation. 2022. "The Rockefeller Foundation Announces Nearly \$500K to Combat AI Bias and Discrimination." *The Rockefeller Foundation* (blog). March 28, 2022. <https://www.rockefellerfoundation.org/news/the-rockefeller-foundation-announces-nearly-500k-to-combat-ai-bias-and-discrimination/>.
- Thiel, David. 2023. "Identifying and Eliminating CSAM in Generative ML Training Data and Models." *Stanford Digital Repository*. <https://purl.stanford.edu/kh752sm9123>.
- Thrower, Sharece. 2021. "What Is an Executive Order, and Why Don't Presidents Use Them All the Time?" *The Conversation*. January 26, 2021. <http://theconversation.com/what-is-an-executive-order-and-why-dont-presidents-use-them-all-the-time-150896>.
- Toma, Alexandra. 2022. "The Changing Landscape of Nuclear Security Philanthropy: Risks and Opportunities in the Current Moment." *Founders Pledge Patient Philanthropy Fund*. <https://static1.squarespace.com/static/62435b2d773bcf0ad9cbb672/t/62cee05cb38a4b2487a533e0/1657725020334/The+Changing+Landscape+of+Nuclear+Security+Philanthropy.pdf>.
- Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, et al. 2023. "International Governance of Civilian AI: A Jurisdictional Certification Approach." *SSRN Scholarly Paper*. Rochester, NY. <https://doi.org/10.2139/ssrn.4579899>.
- Turner, Craig. 1997. "Ted Turner Makes \$1-Billion Pledge to U.N. Programs." *Los Angeles Times*. September 19, 1997. <https://www.latimes.com/archives/la-xpm-1997-sep-19-mn-34095-story.html>.
- Turner, Seth. 2016. "Framers Knew That Slow and Steady Wins the Race." September 16, 2016. <https://www.congressfoundation.org/news/blog/1183>.
- United Nations. 2006. "Ted Turner's United Nations Foundation Delivers \$1 Billion to UN Causes." October 11, 2006. <https://press.un.org/en/2006/dev2594.doc.htm>.
- United Nations. 2024. "Global Digital Compact Zero Draft." [https://www.un.org/techenvoy/sites/www.un.org/techenvoy/files/Global\\_Digital\\_Compact\\_Zero\\_Draft.pdf](https://www.un.org/techenvoy/sites/www.un.org/techenvoy/files/Global_Digital_Compact_Zero_Draft.pdf).
- United Nations Foundation. 2023. "Initiatives." 2023. <https://unfoundation.org/what-we-do/initiatives/>.
- United Nations Foundation. n.d. "The Relationship Agreement between the UN Foundation and the United Nations." *Unfoundation.Org*. May 7, 2024. <https://unfoundation.org/who-we-are/our-financials/un-foundation-un-relationship/>.
- U.S. Department of State. 1997. "U.S. Agenda for UN Reform." November 29, 1997. [https://1997-2001.state.gov/issues/fs\\_us\\_agend\\_un\\_reform\\_0997.html](https://1997-2001.state.gov/issues/fs_us_agend_un_reform_0997.html).
- Wanat, Zosia. 2023. "'EU's AI Act Could Kill Our Company,' Says Mistral's Cédric O." *Sifted*. October 27, 2023. <https://sifted.eu/articles/eu-ai-act-kill-mistral-cedric-o/>.
- White, Joseph. 2023. "Musk: 'AI Stresses Me Out.'" *Reuters*. March 1, 2023, sec. Autos & Transportation. <https://www.reuters.com/business/autos-transportation/musk-ai-stresses-me-out-2023-03-02/>.
- Wong, Scott, Frank Thorp V, Ryan Nobles, and Liz Brown-Kaiser. 2023. "Elon Musk Warns of 'civilizational Risk' Posed by AI in Meeting with Tech CEOs and Senators." *NBC News*. September 13, 2023. <https://www.nbcnews.com/politics/congress/big-tech-ceos-ai-meeting-senators-musk-zuckerberg-rcna104738>.
- Zakrzewski, Cat, Cristiano Lima, and David DiMolfetta. 2023. "Tech Leaders Including Musk, Zuckerberg Call for Government Action on AI." *Washington Post*. September 14, 2023. <https://www.washingtonpost.com/technology/2023/09/13/senate-ai-hearing-musk-zuckerburg-schumer/>.
- Zhao, Christina. 2018. "China: Social Credit System Will Punish the Disobedient." *Newsweek*. May 1, 2018. <https://www.newsweek.com/china-social-credit-system-906865>.

# 24

## ON FOUNDATIONS AND FOUNDATION MODELS

### What lessons can AI and philanthropy learn from one another?

*Diana Acosta-Navas*

The increasing capabilities and enhanced presence of artificial intelligence in a growing number of domains of human life have forced the public to ask what values should guide its design and development. Researchers in the field refer to this problem as the problem of alignment: how can AI be aligned with human values and interests? From this central problem, other questions arise, including “What are the values that should guide the development of AI?” and, more fundamentally, “Who gets to decide?” (Gabriel, 2020). The latter is a question of power. Given the scale at which AI is likely to impact humanity, whoever decides what values it is aligned with will have enormous power.

Decisions about the values and principles that should guide the development of AI involve a complex set of trade-offs, given the many moral challenges that this technology poses. These include long-term and existential risks to humanity caused by a hypothetical super-powerful AI, alongside more immediate concerns about the risk of AI-powered disinformation, problems related to bias and discrimination, copyright, privacy, and labor, among others. How to allocate resources to mitigate these risks is a delicate question, which requires developers to make value trade-offs that can have a significant impact.

Large philanthropic organizations fund much of the current work on AI alignment. Organizations like Open Philanthropy and others in the Effective Altruism network are channeling significant resources into this area, recognizing the crucial need to ensure that AI systems are developed with safety and ethical guidelines in mind. However, this influx of funding from major philanthropic entities has raised concerns among experts about the potential for skewing the research agenda and a resulting concentration of power in the field of AI (Ahmed et al., 2023; Lazar and Nelson, 2023). Critics argue that such concentration of financial resources in the hands of a few organizations can lead to a narrowing of perspectives and priorities, potentially overshadowing other important ethical concerns related to AI. More specifically, they have objected to the disproportionate allocation of resources to research focused on long-term existential risks and the consequent negligence of more immediate risks related to disinformation, discrimination, labor, and copyright (Geburu, 2022; Lindsay, 2023).

This chapter examines the question of how power can be concentrated at the intersection of philanthropy and AI. It draws a parallel between elite philanthropic organizations and advanced

AI developers, to analyze how each can create concentrations of power that stand in tension with fundamental democratic values. Drawing from a family of arguments in the literature, I hold that philanthropic organizations can challenge the value of political equality, by reducing the public's decision-making power in areas of common interest. Taking these arguments at face value, I argue that advanced AI developers share some of these characteristics, which render them problematic from the perspective of democratic values. I argue that the most advanced models in AI have come to occupy a space that is similar to that of large philanthropic organizations. That is a space where anti-democratic concentrations of power can occur due to their ability to bring about outsized societal impact while standing outside the scope of substantive democratic and market accountability.

The final section discusses a set of recent initiatives that aim to democratize decisions about the values with which AI models should be aligned. These initiatives seek to create spaces for stakeholder participation and deliberation over the appropriate ethical guidelines for the design and development of AI. This section argues that, on the one hand, these initiatives fall short of creating binding mechanisms of public accountability for AI developers. On the other hand, they create opportunities for more robust forms of democratic participation.

## **1 Foundations**

This section outlines a family of arguments that have been advanced, among others, by Reich (2018), Saunders-Hastings (2022), Cordelli (2020), and Illingworth (2022) for the claim that the mode of operation of large philanthropic organizations stands in tension with democratic values. These arguments have emphasized two characteristics of large foundations that are in tension with the normative ideals of a democratic society, in particular, the ideal of political equality. In a nutshell, they argue that elite foundations may have a major impact on matters of public concern while standing outside the scope of democratic oversight. For this reason, they can realize donors' conception of the social good and enhance their influence on society, with little transparency and public accountability, thus leading to large concentrations of power. Below I elaborate on each of these claims.

Beforehand, three clarifications are in order. Firstly, this argument is not intended as a moral criticism of philanthropists or their motivations. Everything I argue below is perfectly consistent with a positive view of the individual morality of donors. As a matter of fact, the argument is based on the assumption that philanthropic donations are (to some extent) meant to realize the philanthropists' conception of the social good. Rather, the subject of the argument is the larger political system in which large philanthropic organizations are situated, and the ways in which social and political relations are constituted in this context. Thus, instead of focusing on the individual ethics of philanthropic donations, this argument focuses on its political implications on a democratic society that is constituted around the idea of political equality (Reich, 2018). Secondly, this argument should be interpreted as a form of non-ideal theory. It discusses democratic values and ideals but assumes that there are significant deviations from these ideals, due to existing domestic and global inequalities. Thirdly, this argument does not focus on philanthropy as a general practice or on small philanthropic organizations. Rather, I focus on what Saunders-Hastings (2022) calls "elite philanthropy." While the contours of the category may be somewhat vague, what is relevant to the present argument is the scale of impact and the bindingness of donor intent.

Based on these three assumptions, the argument goes as follows: Political equality among citizens is a core normative ideal of democracy. Part of what it means to be a democratic citizen is to enjoy equal political standing and to share equal authority over common life. This ideal creates an imperative for societies to limit or reduce inequalities of influence, particularly over matters of

public interest. The concentration of such influence in the hands of a wealthy few is problematic, especially when outcomes of significant social importance that should be the subject of democratic decision-making fall under the disproportionate control of a few individuals. However, philanthropic donations bypass the conduits and procedures designed for democratic decision-making, leading to a redistribution of control to private entities. This shift diminishes political equality by limiting the range of public interest matters that individual citizens can effectively influence.

From the perspective of individual morality, philanthropists' actions may be praiseworthy. They may be guided by philanthropists' conception of the common good and their understanding of the needs and interests of their beneficiaries. Given that donations are made from private resources, we may be tempted to conclude that there is nothing objectionable about using them in such ways.

However, from the perspective of democratic values, this situation is problematic regardless of the intentions of philanthropists. It appears that when private donations have large-scale impact on matters of public interest, they are no longer so clearly situated in the purely private sphere. Given their oversized impact on matters that profoundly affect the lives of large numbers of people, we need a critical perspective on their political implications (Cordelli, 2020; Reich, 2018). While it seems perfectly acceptable for philanthropists to use their private resources to pursue their conception of the good, we may be more cautious in examining the receiving end of such transactions. How funds are allocated and how they affect society at large may require a different kind of justification altogether than the donors' private conception of the good. That is, it may require that such impacts be deemed acceptable to beneficiaries and other stakeholders who hold a plurality of values and interests that may, or may not, coincide with those of the philanthropist.

However, decisions about the use of philanthropic funds are not mandated to undergo public deliberation or any other democratic procedure. These decisions are treated as being the prerogative of donors, given their private ownership of resources. While this situation may not constitute grounds for moral objections to donors, it is a cause for concern in terms of its broader political implications.

For one thing, by isolating decisions over matters of public interest from the sphere of democratic procedures, philanthropic activity can lead to the erosion of political equality and the concentration of excessive power in the hands of a wealthy few. This power is exercised through agenda-setting decisions, decisions over the allocation of resources to some causes over others, and choices over the downstream trade-offs created by such decisions (Wenar, 2011). In a democratic context, decisions that affect the public should ideally be made through processes that allow for citizen input and oversight, to ensure that they reflect the collective will and public interest. However, when philanthropists make these decisions unilaterally, they bypass democratic procedures. The problem is exacerbated by the absence of robust and binding mechanisms for those impacted to challenge and contest the way that philanthropic funds are used.

For another thing, philanthropic organizations are exempt from the normal mechanisms of market accountability, such as competition and consumer choice. Philanthropic decisions are not subject to market forces. As a result, their decisions may not be responsive to the actual needs and preferences of those they aim to serve. Shielded from both types of public accountability, philanthropic organizations have no incentive to seek alignment with the preferences and interests of recipients and other stakeholders (Saunders-Hastings, 2022).

In the absence of both democratic and market accountability, Saunders-Hastings adds, elite philanthropic organizations can exert unbounded control over the use of funds. The legal framework governing charitable gift trusts prioritizes the protection of donor intent, barring any balancing of the donor's plans against the needs or preferences of recipients, or broader considerations of social utility. Thus, while elite philanthropists often view their donations as an expression of

altruism and moral commitment, they also wield significant sway over recipient organizations. They can exert prolonged control over the expenditure of funds through formalized mechanisms of influence, such as conditions or restrictions attached to their donations. Because foundations can operate indefinitely, this approach can result in the preservation and transmission of donors' influence over extended periods. Recipient organizations, on the other hand, have a particular incentive to prioritize the satisfaction of major donors, who possess the capacity to make substantial contributions. Under conditions of stark economic inequality at both domestic and global scales, organizations and beneficiaries may not be in a position to decline generous philanthropic donations.

In sum, elite philanthropy has the ability to substantially advance donors' conception of the good. Because decisions over the use of philanthropic funds occur outside the scope of democratic processes, oversight, and control, they remove decisions over matters of common interest from the public domain. Hence, the societal impact of philanthropy falls outside the scope of public accountability (including both democratic and market accountability). It is not treated as something that ought to be justified to stakeholders with a plurality of conceptions of the good. Such isolation from public accountability may lead to large concentrations of power on the part of elite philanthropic organizations and stands in tension with the ideal of political equality.

The following section argues that frontier AI models – so-called foundation models – share some of the characteristics that place elite philanthropy in tension with democratic values. Like large philanthropic organizations, the developers of foundation models have the ability to realize private conceptions of the good by making unilateral decisions about the values that inform these models. In the absence of accountability mechanisms, this could lead to a large concentration of power over matters of public interest.

## 2 Foundation models

The term “Foundation Model” was coined in a 2021 report, written at Stanford University to describe a specific type of technology that would quickly gain dominance in the field of AI. Foundation Models are described in their paper as the result of a confluence between technical and sociological inflections in AI research. The technical inflection point that led to the development and eventual adoption of these models is traced back to the use of “transfer learning,” namely, the technique by which an AI model can transfer “knowledge” acquired from one task and apply it to the performance of another task. Transfer learning allowed AI researchers to train models on surrogate tasks and then adapt them to perform other downstream tasks of interest via *fine-tuning* (i.e., providing human feedback to improve the models' performance on a given task).

Together with improvements in hardware, model architecture, and the availability of larger datasets, these technical advances allowed for a small number of natural language models to become the basis on which a wide range of tasks were performed. These models are used not only to process and predict patterns in natural language but also in images, molecular structures, and protein sequences. Moreover, they are not only trained on textual data but are increasingly multimodal.

This kind of model is characterized by two features: (1) *Emergence*, or the development of behaviors and abilities that were neither explicitly trained nor anticipated by the developers; (2) *Homogenization*, or the consolidation of methodologies for building AI systems based on foundation models.

Due to homogenization, characteristics of the foundation model are inherited by all models that are built as adaptations of it. To clarify, a foundation model is trained on large corpora of multimodal data and serves as the basis for the training of more specialized domain or task-specific models.

Foundation models may be adapted for tasks ranging from sentiment analysis to image captioning, to object recognition, to conversation (in popular chatbots, like ChatGPT). These adaptations can then be employed for a variety of purposes. Any biases or values encoded in the foundation model are reproduced by the adapted models that are built on them. When applications are deployed, such biases find their way into society, by influencing the decisions and actions of end users. These downstream effects are likely to occur on a large scale, given the variety of possible applications built upon them and the scale at which any given application is likely to impact society.

In what follows, I argue that foundation models can lead to concentrations of power that are in tension with democratic values, first because of their outsized influence on matters of common interest and second because these models are shielded from existing mechanisms of public accountability.

### ***2.1 Large-scale impact on matters of public interest***

Before the advent of foundation models, scholars had already begun to discuss the massive scale at which AI models impact society (Weinstein et al., 2021). An AI algorithm deployed on social media for content moderation may impact users worldwide, affecting the quality and nature of public debate. A credit scoring algorithm can impact large numbers of loan applicants. The same is true of AI algorithms used to screen job applicants, assess recidivism risk in criminal justice, and direct law enforcement resources through crime prediction and recommendation algorithms.

The role of technology as a value amplifier is particularly clear in the case of AI systems, given the rapid scalability of their deployment. To the extent that developers' values are reflected in design decisions, these values are likely to determine the ways in which end users are impacted by AI systems (Flanagan & Nissenbaum, 2016). Because of their scalability, these impacts are likely to be significant. Consider, for instance, the role that content moderation algorithms on social media may play during a contested election.

The (expansive and not always evident) influence of developers' values may be the result of deliberate choices on their part, or simply the result of unacknowledged biases. Developers may, unbeknownst to themselves, prioritize a group of users while neglecting others. This has been the case when content moderation algorithms have been deployed for low-resource languages, leading to poorer performance. On the other hand, developers may engage in fairly sophisticated deliberation over value trade-offs and compromises. This was the case during the COVID-19 pandemic when social media platforms were forced to balance the value of free speech against the value of public health, in their content moderation algorithms. Whether or not developers rank and prioritize these values in a way that is responsive to stakeholders, their choices are likely to influence high-stakes outcomes for large groups of people.

Foundation models amplify this phenomenon. Using foundation models as a basis for a wide range of applications makes developers' choices far more impactful than other AI algorithms. Early on, Bender et al. (2021) warned of some risks associated with these models, including their potential to privilege hegemonic voices and perspectives due to their outsized representation in the models' training data (see also Weidinger et al., 2021). Another risk, they argued, is for training data to contain unaccounted biases due to designers' lack of understanding of salient social categories. The authors cite the example of #MeToo as an important contribution to the public discussion of gender discrimination that could be filtered out by a model's guardrails on what constitutes inappropriate sexual communication (Bender et al., 2021). Likewise, Welbl et al. (2021) discuss the disparate impact of toxicity mitigation tools on non-hegemonic English dialects and topics related to women and minorities.



Note that the biases discussed by these authors are the result of design decisions, aimed at mitigating potential risks stemming from these models. In this way, they are the expression of value-driven decisions – some based on deliberate value trade-offs and some based on implicit assumptions. When these choices are encoded in foundation models, they are likely to trickle down to downstream adaptations, from content moderation to recommendation systems, to digital healthcare services. How developers rank, prioritize, and balance values in conflict will then trickle down to applications and have significant impact on end users and decision subjects.

The implications of these models should be analyzed at two morally relevant levels: individual and societal. At the individual level, researchers have drawn attention to the possibility of *systemic failure* resulting from AI-based decision-making. Systemic failure occurs when undesirable outcomes consistently accrue to the same individuals or groups (Bommasani et al., 2022). Research on outcome homogenization suggests that decision-making using AI systems often leads to the accumulation of negative outcomes for individuals. Furthermore, foundation models can lead to outcome homogenization across a wide variety of tasks and domains:

employing many adaptations of the same foundation model for multiple automated decision-making tasks means that decision subjects may face a more homogeneous set of judgments rooted in the underlying foundation model. This algorithmic monoculture could lead to consistent and arbitrary rejection, misclassification, or ill-treatment of individual decision subjects.

(Bommasani et al., 2021)<sup>1</sup>

In addition to the impact on individuals, it is important to consider the impact of foundation models on society more broadly. On a larger scale, Kleinberg and Raghavan (2021) argue that the use of shared algorithms for high-stakes decision-making leads to algorithmic monoculture, that is, to the homogenization and convergence of the decisions of different actors. Given that foundation models are likely to be employed by a large number of actors in a wide variety of domains and that applications built on top of them can inherit their biases, it would be reasonable to expect that any impact from these models would occur at a broad systemic level. Thus, “the application of foundation models across domains has the potential to act as an epistemically and culturally homogenizing force, spreading an implicit perspective, often a socially dominant one, across multiple domains of application” (Bommasani et al., 2021).

Consider again the case of toxicity evaluation and mitigation tools for foundation models. These may be used as the basis for recommendation systems, as well as the deletion, promotion, or demotion of content on social media platforms, captioning systems for audiovisual media, translation, social chatbots, personal assistants, and educational materials. Whatever value choices and trade-offs are coded into the model will be reflected in these applications. Developers may define toxicity in a way that prioritizes the interests of the digital platforms rather than users. They may prioritize child safety by removing sexually explicit content, disproportionately impacting content posted by LGBTQ influencers, or feminists denouncing sexual harassment. They may prioritize readers’ sense of comfort and willingness to engage with content and thus disproportionately impact politically sensitive topics and contributions in non-dominant dialects (Welbl, 2021; see also: Liang et al., 2023). These decisions can affect the kind of content we can contribute to public discussions and, just as importantly, what kind of content we have access to. The value trade-offs coded into these tools will be reflected in what perspectives, forms of expression, and types of content are part of the informational environment of large numbers of people.

Mitigation tools for toxic content are just one area of public interest where foundation models are likely to have an outsized societal impact. Values, trade-offs, priorities, and biases in these models are likely to be inherited by applications in other domains, from hiring to credit screening, risk assessment, and medical diagnoses. More generally, algorithmically mediated decision-making occurs in areas that determine individual access to, and distribution of, basic social goods. To this extent, they have a large-scale impact on matters of public interest.

## **2.2 Limited stakeholder control**

One of the central ethical concerns raised by the use of AI in high-stakes decisions is that of the transparency and explainability of its algorithms. The opacity of algorithmic decisions, especially those that are high stakes, is concerning for two reasons. First, AI operates by detecting patterns in large swaths of data that are unintelligible to human beings. As such, their outcomes are difficult to understand and evaluate for the users of the systems and the individuals impacted by them. If we cannot understand how an algorithm reaches a given decision, we cannot judge whether it is based on relevant (and fair) criteria. This opacity, in turn, is problematic because it makes it substantially difficult for decision subjects to challenge unfavorable outcomes.

The field of explainable AI has taken steps toward determining what would make algorithmic decisions explainable in different fields of application, and how explainable AI systems may be built (see: Longo et al., 2023). However, foundation models raise a particular set of challenges. Due to the sheer volume of their training data and the number of parameters in the model, there are difficulties associated with maintaining adequate documentation and understanding the models' behavior (Bender et al., 2021). These properties detract from stakeholders' ability to contest negative outcomes that affect them.<sup>2</sup>

They also make it difficult for researchers, third parties, and society at large to understand and address existing weaknesses in the models. There are no mechanisms by which stakeholders could contest or challenge the risks created by these models. More generally, it may be the case that broader society does not recognize or accept the value systems that are encoded in the models. If this were the case, there would be no mechanism for society to express disapproval or to pressure developers to change the models.

Concerns about accountability to stakeholders could potentially be mitigated if the design of the systems took stakeholder input into account. Participatory design has been upheld as a means of making technology responsive to different stakeholder groups and, particularly, those that have been historically marginalized and excluded (Constanza-Chock, 2020). Bondi et al. (2021) argue that some form of participatory design must be a constitutive aspect of AI systems designed to promote the social good. By setting up consultation processes with representatives of stakeholder groups and keeping communication channels open throughout the life cycle of the system, developers make their products responsive to the interests and preferences of the people impacted by their models.

However, participatory approaches to AI systems have been strongly criticized. Sloane et al. (2022) argue that participatory design is sensitive to the particular contexts in which technologies are deployed. Hence, the accountability-granting aspect of participatory processes is constrained to the contexts in which stakeholder consultations are conducted. Because AI systems scale so easily, their impact will rapidly extend beyond such contextual boundaries. Thus, participatory design makes AI systems responsive only to a limited set of stakeholders and in a limited context, but the impact of these systems far exceeds these limits. Hence, participatory design can only make AI accountable to the public in a very minimal sense.

Foundation models only exacerbate these concerns. Due to their adaptability to a wide variety of tasks and applications, each of which is likely to have a significant impact, the creation of participatory processes that take into account all relevant stakeholder groups requires the creation of new participatory methodologies, which enable the scalability of meaningful stakeholder engagement. One such methodology is discussed in Section 3.

As things stand, however, the developers of foundation models are not subject to any meaningful form of stakeholder control or accountability, despite their outsized impact on matters of public interest. The reason is twofold. Firstly, because of their opacity and the difficulty of understanding their operation, their results are not easily challenged, nor are their flaws easily identified. Secondly, given the variety of applications for which they can be used, and their rapid scalability, existing participatory mechanisms are not an effective means of incorporating stakeholder input.

### ***2.3 Concentration of power***

Section 1 of this chapter discusses the idea that the activities of elite philanthropic organizations can lead to concentrations of power that are in tension with democratic ideals. This is due to their ability to bring about large-scale societal impact in ways that are not constrained by the preferences or interests of those whose lives are impacted. This section argues that foundation models share some of the characteristics of elite foundations. Because of their large-scale societal impact and the absence of meaningful stakeholder control or accountability, foundation models allow for power to disproportionately accrue in the hand of their developers.

As these models encode developers' value systems, their impact is likely to reflect their implicit values, biases, and assumptions as well as the explicit value trade-offs that constitute an inescapable aspect of technological design. Hence, foundation models are poised to realize developers' conceptions of the good. In doing so, they will affect large groups of stakeholders in ways that are driven by developers' values but are not responsive to stakeholders' preferences and interests. Like elite philanthropic organizations, these outsized concentrations of power have the potential to strain democratic ideals. Not least, they undermine the basic democratic assumption that matters of public interest should be decided by actors and procedures subject to public oversight, accountability, and control. This assumption is derived from the fundamental value of political equality, which is undermined when power is disproportionately concentrated in a few individuals and organizations.

This concentration of power is exacerbated by the fact that only a handful of organizations have access to the resources needed to train and deploy foundation models. Training these models requires access to data and computational power that only a few firms have. Therefore, "the organizations most capable of producing competitive foundation models will be the most well-resourced: venture-funded start-ups, already-dominant tech giants, and state governments" (Bommasani et al. 2021). This implies that the unchecked and unrestrained social influence wielded through these models is held by a select number of entities. While this does not entirely isolate developers from market accountability (in the ways that philanthropic organizations are isolated), it does significantly reduce the size of the existing market to a homogeneous group of developers. Furthermore, the opacity of models and the difficulty of interpreting them make it difficult for customers to make informed judgments regarding what model best promotes or represents their preferences.

Importantly, the concentration of power that is created by foundation models compounds some of the issues raised in the introduction of this chapter regarding the disproportionate power of large philanthropic organizations in setting the agenda for AI safety and its alignment with human values. Ahmed et al. (2023) describe how sizable philanthropic donations, in addition to policy

advocacy and other forms of influence, have advanced an agenda for AI alignment that is endorsed by a specific epistemic community of researchers, developers, and philanthropists, which they refer to as the “AI safety epistemic community.” This is “a community with clearly-defined shared values and methods of knowledge production,” the impact of which has extended beyond the community’s bounds and has “translated their shared moral and normative claims into technical solutions and recommendations for AI policy that may have lasting, global implications” (Ahmed et al., 2023, 1-2). This community, which endorses a homogeneous set of values, is pursuing a particular agenda in AI alignment, leading to the prioritization of some issues over others, and to the assumption of important trade-offs. To the extent that these choices influence the development of AI models, they can be crystallized in the design of these models, leading to large-scale impact, which stands outside the scope of existing forms of public accountability. As is the case with the activity of philanthropic organizations, questions of the individual morality of these decision-makers are independent from questions about the political implications of their actions. Whether or not their agenda for value alignment is morally appropriate, the power concentration at this intersection stands in tension with democratic values.

### **3 Balancing acts: can foundation models be democratized?**

There is a set of strategies that may mitigate the concentration of power that results from the design and deployment of foundation models. These range from some forms of regulation to the creation of open-source models that expand the scope of actors who can benefit from foundation models. This section focuses on “deliberative alignment” as an improvement and enhancement of existing participatory design methods. I argue that through the deliberative alignment of foundation models, developers have the opportunity to shape their products in ways that are responsive to the preferences and interests of a broader group of stakeholders. This section explains the notion of deliberative alignment and how recent initiatives on the part of AI developers have used it to obtain stakeholder input on what values should guide the development of AI systems.

“Deliberative Alignment” refers to a set of initiatives to create deliberative spaces where stakeholders are invited to deliberate with one another with the aim of generating a set of recommendations for the developers of AI models. The end goal of these recommendations is to inform developers on how to align models and applications with stakeholders’ preferences and interests. Section 2 argued that participatory design is limited as a tool for artificial intelligence, especially foundation models. The main challenge for participatory approaches to AI is to develop methodologies that work at the required scale, while retaining the accountability-granting properties of traditional participatory methods. To this end, participatory methods should be scalable and inclusive, ensuring that the full range and diversity of stakeholders are adequately represented. At the same time, they should be open, thus allowing stakeholders to set the agenda in accordance with their interests and preferences.

These desiderata are seen as being in tension with each other. Allowing stakeholders to set the agenda and deliberate to reach a set of agreements is thought to require smaller groups of participants. These qualitative participatory spaces occur in focus groups or citizen assemblies, where the limited number of participants allows for deliberation to be held and for a tractable set of outputs to be generated, which can, in turn, produce an actionable list of recommendations. Quantitative participation, on the other hand, allows for a greater scale and is a more efficient method of consultation that more easily ensures the inclusion of all relevant stakeholders. However, quantitative methods fall short of providing a space for open participation. The reason is threefold: (1) Quantitative methods do not grant stakeholders agenda-setting power. (2) Quantitative methods do not

allow for deliberation or the search for common ground among competing interests. And (3) by quantitatively aggregating results, they may favor majority views while setting aside the interests and perspectives of (potentially marginalized and vulnerable) minority groups. This tension between the desiderata of scale, inclusion, diversity, openness, and tractability is one of the main reasons why participatory processes have been regarded as ill-suited to inform the development of AI models (Sloane et al., 2022).

Recent attempts at deliberative alignment have sought to resolve this tension by deploying technologies that allow for deliberative participation at scale. They use data analysis of participants' contributions to find areas of common ground among them, with the goal of integrating a wide diversity of stakeholders' perspectives and interests. This approach to stakeholder consultation is based on the recognition of the scale at which AI will impact society. Consequently, stakeholder participation is designed to ensure that the outcome can be responsive to as broad and diverse an array of stakeholders as is necessary.

A group of technologies enable these participatory spaces, which I will refer to as *deliberative technologies*, following Konya et al. (2023a, 2023b). Deliberative technologies underpin platforms where large numbers of stakeholders can express their perspectives and interests in their own words and can vote on the contributions of others. They use machine learning to process this information and create visualizations that make the information tractable to both participants and decision-makers. Importantly, these technologies analyze stakeholder input in a way that allows them to understand voting patterns among different stakeholder groups.

One example of such technology is Polis (n.d.): an online participation platform that generates real-time visualizations of the opinion landscape, allowing participants, moderators, and decision-makers to understand the distribution of opinion groups and to identify areas of consensus among opinion groups who otherwise disagree.

Polis's innovation in the space of online deliberation platforms lies in its nuanced revealing of the overall opinion landscape in a way that preserves opinion groups and respects minority dissent, as well as assuming no relationships between various comments, other than that they could be compared.

(Small et al., 2021, 4)

Similar tools, such as Remesh, are used to host collective dialogues with the aim of identifying areas of common ground among participants.

AI labs and governing bodies can use this process to develop concise sets of common ground policy guidelines that bridge demographic divides and reflect what a given population wants. It is ideal for those that have to make policy decisions that impact large populations and want a democratic process to align those decisions with informed public will.

(Konya, et al., 2023a, 2023b, 2)

These tools employ “bridging-based ranking” to identify and promote contributions that generate agreement across political or demographic divides (Ovadya, 2022).<sup>3</sup>

These and similar tools have been employed in the context of policymaking (Horton, 2018), as well as during peace negotiations (Konya et al., 2023a, 2023b), and for enabling fact-checking (Miller, 2022). More recently, AI companies have begun working with technology-enabled participatory processes, as a consultation mechanism for model alignment (Coy, 2023). In 2023,

Anthropic piloted the use of Polis to enable collective deliberation over the values and principles that should serve as a “constitution” for its chatbot, Claude (Ganguli et al., 2023; Roose, 2023). They describe the process as enhancing transparency in the design and development of AI models that can be better tailored to the needs of specific communities.

Open AI piloted a similar project, creating participatory spaces for deliberation on the appropriate ethical guidelines for Chat GPT. The program, called “Democratic Inputs for AI,” was the company’s first attempt to democratize decisions over the values encoded in its product. With it, they sought to address the question of whose values Chat GPT should be aligned with and, more importantly, the question of who should decide (Eloundou, 2024; Perrigo, 2024; Zaremba et al., 2023). To some extent, it constitutes an attempt to empower stakeholders to make impactful decisions about the behavior of the system. Such deliberative spaces revolved around questions like: “What principles should guide AI when dealing with issues that involve both human rights and local cultural or legal differences, such as LGBTQ rights and women’s rights?” “Should AI’s responses change based on the location or culture in which it is being used?” or “What categories of content, if any, do you believe creators of AI models should focus on restricting or denying and what criteria should be used to determine those restrictions?”

The use of deliberative technologies to enable these participatory processes is intended to enhance the scale of deliberation, while maintaining tractability. It is also intended to make the outputs of the process responsive to a variety of stakeholder groups. Thus, rather than producing recommendations that would favor majority groups, the processes are specifically designed to surface values and interests that are common across stakeholder groups and to generate as broad a consensus as possible. This set of methods and tools can thus ease the tension between the desiderata of scalability, inclusion, diversity, tractability, and openness.

The question remains, however, whether these forms of participatory design could reduce the concentration of power that is created by foundation models and mitigate their tension with democratic values.

In some respects, this form of participatory design falls short of creating the kind of democratic accountability that would mitigate concentrations of power. For one thing, these deliberative spaces are only open to stakeholders with access to technology and a certain level of technological literacy. This excludes a large number of stakeholders and may potentially bias the outcomes of the process.<sup>4</sup> For another thing, it remains an open question whether such deliberative processes should be employed to determine design features at the level of downstream applications (chatbots, in the two cases described above), at the level of models, or at an intermediate level. Holding participatory spaces for the underlying models would be more challenging, especially as stakeholders may find it difficult to conceptualize design decisions at this level and understand how they may impact them. However, given the large number of potential applications for foundation models, more exhaustive efforts at deliberative alignment could prove more resource-intensive than they appear at this early stage.

More importantly, these participatory processes fall short of creating public accountability, given developers’ tenuous commitment to abide by the resulting recommendations. Such participatory processes allow for the recognition of the plurality of stakeholders’ values and interests and provide developers with information about areas of overlapping consensus. In that respect they constitute an important step in broadening and diversifying inputs for value alignment. However, in the absence of a strong commitment to abide by the resulting recommendations, developers do not hold themselves accountable to stakeholders in any meaningful sense. Given the absence of other mechanisms of public accountability, developers could disregard the public’s

recommendations and continue to exercise unilateral power. In these circumstances, the specific role and value of deliberative alignment in distributing power remain unclear.

As it stands, however, deliberative alignment goes further than existing forms of public accountability in promoting democratic ideals. In short, technologically enabled participatory spaces allow for a more precise and nuanced representation of the public's interests and preferences. Existing forms of democratic decision-making and accountability tend to limit the types and the richness of signals through which the public can communicate its interests, values, and will: "Political reality on the ground is vastly more complex than our political symbols and categories allow us to express [...] Elections and referendums severely constrain the information populations are able to send to governing bodies" (Megill, 2019). Qualitative signals with greater nuance and resolution are restricted to small spaces (such as citizen assemblies) and are more difficult to integrate into a coherent set of guidelines for decision-makers.

Unlike existing democratic processes, technology-enabled participatory spaces do not force voters to distill their entire set of values and beliefs into a single vote (Perrigo, 2024). Unlike market signals, they allow stakeholders to convey complex sets of preferences and interests on their own terms. They allow the whole range of values and interests to surface, generating a landscape of public opinion at a higher resolution and in more dimensions than any existing channels. They also allow the public to jointly determine how a rich set of preferences and interests should be prioritized by more effectively managing and integrating both qualitative and quantitative signals (Megill, 2019).

In this way, they constitute a step in the direction of greater political equality, by allowing participants to inform decision-makers of the full range of their preferences and interests. Efficient and regular large-scale consultations would thus allow decisions to better reflect public values. This can indeed increase public control and accountability over matters of public interest. These technologies can also advance the liberal ideal of public reason by revealing with greater precision the contours of overlapping consensus among conflicting conceptions of the good, and by portraying with greater clarity the kinds of reasons that people with different comprehensive doctrines can reasonably endorse.

For these reasons, these nascent initiatives to create democratic accountability for foundation models have the potential to mitigate anti-democratic concentrations of power. The technologies that enable them have the capacity to create more robust forms of public accountability than existing democratic or market mechanisms. Whether they do so will largely depend on the extent to which AI developers make binding commitments to abide by the will of their stakeholders. More specifically, developers will need to make decisions about the conditions under which the outcomes of participatory processes become binding. These decisions (along with other design decisions for the technological platforms) will themselves involve a series of value trade-offs (e.g., along which axes are stakeholder groups identified? How to debias datasets so that they don't prioritize some perspectives while marginalizing others? How to mitigate the risk of hallucination in the generation of recommendations? etc.). Furthermore, these frontier deliberative technologies are likely to have an impact on fundamental political concepts, like consensus, representation, and democracy. How power is distributed and exercised in making these judgments is itself a question worthy of further study.

#### **4 Conclusion**

This chapter analyzes the existing relation between philanthropic organizations and the alignment of AI models with human values. It draws a parallel between elite philanthropic foundations and foundation models in AI and discusses how foundation models can compound the power

concentration of large philanthropic organizations that has concerned democratic theorists for decades. It holds that foundation models have come to occupy a political space that has been previously occupied by elite philanthropic organizations, that is, a space that allows for large concentrations of power, where a small group of individuals can have substantial impact on matters of public interest without substantive public accountability.

Section 1 presents the argument that, in the absence of mechanisms for public accountability, the operation of philanthropic organizations can lead to concentrations of power that strain democratic values, in particular, political equality and respect for reasonable value pluralism. Section 2 argues that foundation models share the characteristics of elite philanthropy: they are poised to have substantial societal impact and contribute to the realization of individual conceptions of the good; they stand outside the scope of public oversight and accountability; and they can, therefore, lead to problematic concentrations of power. Section 3 analyzes a particular response to this problem that industry leaders have begun to explore, namely, the creation of technology-enabled deliberative spaces that allow for robust public participation at scale and are intended to inform the value choices of model developers.

Whether these participatory processes will mitigate the anti-democratic tendencies of foundation models remains to be seen. What is clear is that these technologies have the capacity to reframe democratic participation in ways that more closely approximate the ideals of public accountability and political equality. For this reason, other types of organizations seeking to enhance their democratic accountability and the legitimacy of their societal impact could potentially turn to technologically enhanced consultation processes. With these tools, philanthropic organizations could take a step toward being more accountable to those whose lives they impact. This step could begin to alleviate existing concerns about concentrations of power and their potential strain on democratic values. Any such attempt, however, should be informed by a critical interrogation of the value judgments that inform the design of deliberative technologies and how they are incorporated into decision-making processes.

### Notes

- 1 Bommasani et al. (2022) suggest that certain fine-tuning methods used to adapt foundation models to specific applications can help reduce outcome homogenization.
- 2 Stanford researchers recently published a transparency index for ten leading foundation models, including open-source models such as Meta's Llama. None of the frontier models has an average score above 60%. Across the ten models, the average transparency score for their societal impact is 11%. The average transparency score for the risks associated with the models is 24% (Bommasani et al., 2023).
- 3 More recent initiatives have explored the use of more advanced natural language processing to enhance these participatory processes. Bakker et al. (2022), for instance, report on the use of large language models (LLMs) to assist groups of people in collaboratively creating written content that achieves high levels of agreement among its users. In this study, the model was trained to generate statements that generated the greatest consensus, based on the opinions contributed by the group. Small et al. (2023) report on the use of Large Language Models to enhance Polis, by automatizing some resource-intensive processes, like topic modeling, summarization, and moderation, among others.
- 4 Polis has adopted a general policy that it can only host discussions where all stakeholders are digitally enabled. In the context of the deliberative alignment for Foundation Models, it may be especially challenging to ensure that all stakeholders are able to participate (see: <https://pol.is/home>).

### References

Ahmed, S., Jazwińska, K., Ahlawat, A., Winecoff, A., & Wang, M. (2023, November 22). Building the Epistemic Community of AI Safety. *SSRN*. <https://doi.org/10.2139/ssrn.4641526>



- Bakker, M. A., et al. (2022). Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences. *arXiv*. <https://arxiv.org/abs/2211.15006>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. <https://doi.org/10.1145/3442188.3445922>
- Bommasani, R., et al. (2021). On the Opportunities and Risks of Foundation Models. Retrieved from <https://crfm.stanford.edu/assets/report.pdf>
- Bommasani, R., et al. (2022). Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *arXiv*. <https://arxiv.org/abs/2211.13972>
- Bommasani, R., et al. (2023). The Foundation Model Transparency Index. *arXiv*. <https://arxiv.org/abs/2310.12941>
- Bondi, E., Xu, L., Acosta-Navas, D., & Killian, J. A. (2021). Envisioning Communities: A Participatory Approach towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '21)*. <http://dx.doi.org/10.1145/3461702.3462612>
- Constanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press.
- Cordelli, C. (2020). *The Privatized State*. Princeton University Press.
- Coy, P. (2023, April 5). Can A.I. and Democracy Fix Each Other? *The New York Times*. Retrieved February 6, 2024, from <https://www.nytimes.com/2023/04/05/opinion/artificial-intelligence-democracy-chatgpt.html>
- Eloundou, T. (2024, January 16). Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans. Retrieved February 6, 2024, from <https://openai.com/blog/democratic-inputs-to-ai-grant-program-update>
- Flanagan, M., & Nissenbaum, H. (2016). *Values at Play in Digital Games*. The MIT Press.
- Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3), 411–437. <http://dx.doi.org/10.1007/s11023-020-09539-2>
- Ganguli, D., et al. (2023, October 17). Collective Constitutional AI: Aligning a Language Model with Public Input. Retrieved February 6, 2024, from <https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input>
- Gebru, T. (2022, November 30). Effective Altruism Is Pushing a Dangerous Brand of ‘AI Safety’. *Wired*. Retrieved February 5, 2024, from <https://www.wired.com/story/effective-altruism-artificial-intelligence-sam-bankman-fried/>
- Horton, C. (2018, August 21). The Simple but Ingenious System Taiwan Uses to Crowdfund Its Laws. *MIT Technology Review*. Retrieved February 6, 2024, from <https://www.technologyreview.com/2018/08/21/240284/the-simple-but-ingenious-system-taiwan-uses-to-crowdfund-its-laws/>
- Illingworth, P. (2022). *Giving Now: Accelerating Human Rights for All*. Oxford University Press.
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic Monoculture and Social Welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118. <https://doi.org/10.1073/pnas.2018340118>
- Konya, A., et al. (2023a). Deliberative Technology for Alignment. *arXiv*. <https://arxiv.org/abs/2312.03893>
- Konya, A., et al. (2023b). Democratic Policy Development Using Collective Dialogues and AI. *arXiv*. <https://arxiv.org/abs/2311.02242>
- Lazar, S., & Nelson, A. (2023). AI Safety on Whose Terms? A Machine-Intelligent World. *Science*, 381(6654), 138
- Liang, P., et al. (2023). Holistic Evaluation of Language Models. *arXiv*. <https://arxiv.org/abs/2211.09110>
- Lindsay, D. (2023, July 31). 33 Leaders Standing Up to Big Tech in the Age of A.I. *Chronicle of Philanthropy*. Retrieved February 5, 2024, from <https://www.philanthropy.com/article/33-leaders-standing-up-to-big-tech-in-the-age-of-a-i>
- Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *Information Fusion* 106 (June): 102301. <https://doi.org/10.1016/j.inffus.2024.102301>
- Megill, C. (2019, June 24). Beyond Flatland: Machine Learning and the End of the Two-Party Binary. *Civicist*. Retrieved February 6, 2024, from <https://web.archive.org/web/20190629035125/https://civichall.org/civicist/beyond-flatland-machine-learning-end-two-party-binary/>

- Miller, C. (2022, November 20). Elon Musk Embraces Twitter’s Radical Fact-Checking Experiment. *Wired*. Retrieved February 6, 2024, from <https://www.wired.co.uk/article/elon-musk-embraces-twitters-radical-crowdsourcing-experiment>
- Ovadya, A. (2022). *Bridging-Based Ranking*. *Belfer Center for Science and International Affairs*. Retrieved from [https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Aviv\\_BridgingBasedRanking\\_FINAL\\_220518\\_0.pdf](https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Aviv_BridgingBasedRanking_FINAL_220518_0.pdf)
- Perrigo, B. (2024, February 5). Inside OpenAI’s Plan to Make AI More ‘Democratic’. *Time Magazine*. Retrieved February 6, 2024, from <https://time.com/6684266/openai-democracy-artificial-intelligence/>
- Polis (n.d.). Polis: Input Crowd, Output Meaning. Retrieved February 6, 2024, from <https://pol.is/home>
- Reich, R. (2018) *Just Giving: Why Philanthropy Is Failing Democracy and How It Can Do Better*. Princeton University Press.
- Roose, K. (2023, October 17). What If We Could All Control A.I.? *The New York Times*. Retrieved February 6, 2024, from <https://www.nytimes.com/2023/10/17/technology/ai-chatbot-control.html>
- Saunders-Hastings, E. (2022). *Private Virtues, Public Vices: Philanthropy and Democratic Equity*. Chicago University Press.
- Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation Is Not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. <https://doi.org/10.1145/3551624.3555285>
- Small, C., Bjorkegren, M., Erkkilä, T., Shaw, L., & Megill, C. (2021). Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *Recerca. Revista de Pensament i Anàlisi*, 26(2), 1–26.
- Small, C. T., et al. (2023). Opportunities and Risks of LLMs for Scalable Deliberation with Polis. *arXiv*. <https://arxiv.org/abs/2306.11932>
- Weidinger, L., et al. (2021). Ethical and Social Risks of Harm from Language Models. *arXiv*. <https://arxiv.org/abs/2112.04359>
- Weinstein, J., Reich, R., & Sahami, M. (2021). *System Error: Where Big Tech Went Wrong and How We Can Reboot*. Hodder & Stoughton.
- Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., & Huang, P.-S. (2021). Challenges in Detoxifying Language Models. *Findings of EMNLP*. <https://doi.org/10.18653/v1/2021.findings-emnlp.210>
- Wenar, L. (2011). Poverty Is No Pond: Challenges for the Affluent. In P. Illingworth, T. Pogge, & L. Wenar (Eds.), *Giving Well: The Ethics of Philanthropy* (pp. 104–132). OUP USA.
- Zaremba, W., et al. (2023, May 25). Democratic Inputs to AI. Retrieved February 6, 2024, from <https://openai.com/blog/democratic-inputs-to-ai>

# THE AI EXTREME RISK MITIGATION PHILANTHROPIC SECTOR

## A philanthropic ecosystem at the forefront of AI

*Siméon Campos and Daniel S. Schiff*

Artificial intelligence (AI), like previous advanced information technologies, offers new opportunities and challenges for philanthropic efforts. As a tool, it offers a variety of new methods for designing new interventions, evaluating organizational effectiveness, or improving workflows and organizational infrastructure. As a substantive field, AI represents a complex and dynamic set of social challenges that need to be addressed whatever the objective of a non-profit organization, foundation, or public sector actor.

In this chapter, we explore the implications of AI for actors and organizations focused on philanthropy, in the broadest sense. In Section 1, we introduce two notions of AI in the context of philanthropy: the implications of *AI-as-a-tool* for philanthropic actors and organizations, and AI as a substantive *field* and *target* of philanthropic action. We then focus on the latter topic by providing a history, overview, and analysis of the AI philanthropic sector. In particular, we discuss a loose set of communities focused on assessing and mitigating extreme risks related to AI, which are often associated with attention to AI governance and AI safety. While the appropriate definitions and boundaries between subcommunities are themselves a focus here, for simplicity, we refer to the movement as the “AI extreme risk mitigation philanthropic sector” (AIERMPS), defined as the set of actors and organizations that seek to mitigate extreme AI-related risks through and with private philanthropy.

In Section 2, we review the founding of the AIERMPS and provide an overview of the core elements of its culture, strategy, and organizational landscape. We discuss highly unique aspects of its community, such as its strong intellectual entanglement with transhumanism, rationalism, and utilitarianism. These aspects of the AIERMPS, combined with its close relationship with major philanthropic donors, help to explain its rise in leading private sector organizations during the new AI spring of the 2010s. In turn, the engagement of these actors in leading AI research labs and policy conversations foregrounded its rise from a niche to a mainstream movement in the 2020s, following the public popularization of generative AI.

In Section 3, we turn to familiar and unique challenges faced by the movement. For example, while many philanthropic movements face persistent disagreements about structure or strategy, the AI extreme risk mitigation movement faces a unique challenge in that some of its concerns may only bear out in the future and thus remain inaccessible to study, limiting the opportunity for

useful feedback on effective strategies. In addition, philanthropic efforts are unusually focused on technical, rather than social issues, such as addressing the inexplicability of black box AI models and promoting the alignment of these systems with (unsettled) human values. We examine related structural challenges associated with the community, such as the tensions between trying to advance highly capable AI through leading research labs and technology companies and doing so as part of advancing safety efforts.

Based on these analyses, Section 4 highlights what the AI philanthropic community could learn from related fields and social movements. We consider the significance of centering AI itself as a technical issue with relatively low public salience compared to climate change or animal welfare, which were instead framed early on as social problems. Indeed, this difference between mass public movements and an elite-led project has implications for advocacy, fundraising, influence efficacy, and policymaking efforts, along with aspects of the community like its demographic profile. We also consider an issue common to philanthropic communities: internal division, particularly in the AI domain, between actors focused on so-called short-term issues (i.e., AI ethics) and long-term ones (i.e., AI safety). We discuss how other social movements have worked to manage conflicts and build coalitions. While the AI philanthropic movement may be most similar to other elite-driven efforts, such as the open-source Internet community or the nuclear safety community, we conclude that it has much to learn from other enduring social movements.

## **1 An overview of the AI philanthropic landscape**

### *1.1 AI-as-a-tool*

To a significant degree, AI continues historical conversations in the philanthropic community related to the use of technology and analytics. Associated concepts include data, ICT systems, advanced analytics, big data, automated decision systems, and e-government. Organizations in the non-profit and public sectors, and even wings of the private sector focused on corporate responsibility have sought to leverage these tools to advance their efforts.

For example, the philanthropic movement has considered the use of advanced analytics to predict the profile or giving behavior of donors (Eiland et al., 2021; Mittal & Srivastava, 2021; Sulaeman, 2018), to enhance fundraising efforts (Key, 2001; Vequist, 2014), and to guide potential donors or government organizations in the selection of effective charities themselves (Ramirez & Saraoglu, 2009; Singer, 2019; Stern, 2013). This emphasis on effectiveness and efficiency became increasingly popularized through several movements, including the turn to evaluation planning and logic modeling (Kaplan & Garrett, 2005), the early effective altruism movement, and the evidence-based philanthropy and policy movements of the 2000s and 2010s (Braverman et al., 2004; Fiennes, 2017; Johnson, 2018; Pawson, 2002).

Government agencies and private foundations alike increasingly require the use of logic models, evaluation plans, cost-benefit analysis, cost-effectiveness analysis, randomized controlled trials, or other high-quality quasi-experimental studies in areas ranging from health care, education, and public welfare to innovation policy and global development (de Souza Leão & Eyal, 2019; Pawson, 2002; White, 2019). A shared logic across these efforts is the desire to ensure that philanthropic or public money is effectively achieving its goals, while a secondary logic is to do more with less, especially in the face of reduced federal funding or stressed fundraising (Gore, 1993; Osborne, 1993).

For these reasons, the use of data, analytics, and now, AI, is an ongoing but largely unfinished ambition of the philanthropic sector, pursued in the hope that it can improve program efficiency,

streamline operations, enable more effective intervention design, and enhance evaluation (Henriksen & Blond, 2023; Madeo, 2022; Shapiro & Cody, 2015; Voda, 2014). Moreover, while some of these advances are particularly salient to the philanthropic sector, still other operational advances due to informatics and AI are increasingly being adopted across the public and private sectors in general, such as the use of AI to improve recruitment hiring, performance management, legal and compliance functions, procurement, financial management, worker training, and more (Noordt & Tangi, 2023; Wirtz et al., 2019; Zuiderwijk et al., 2021). In short, much of AI's promise to the philanthropic sector is similar to its potential for productivity gains more broadly, simply applied to the set of organizations focused on philanthropy. However, the philanthropic sector has historically been resource-constrained, with below-market wages, precarious work, and scrutiny over operations that often limit its ability to retain technical talent and experiment with new innovations. It is no surprise, then, that sectors like financial services and telecommunications are better positioned to innovate and have become more robust adaptors of AI, while the philanthropic sector lags behind.

A characteristic of AI compared to other analytical tools is that there are some unique capabilities associated with AI that enable relatively novel ways of adopting AI-as-a-tool. These efforts leverage the powerful predictive capacity of AI, emanating from its ability to find subtle patterns in large data sets. Most commonly referred to as “AI for good” or “AI for social good,” the actors and organizations aligned with this movement focus on creative applications of AI to solve large-scale social problems. For example, in the environmental space, AI has been used to promote forest ecosystem restoration, track wildlife diversity, and identify appropriate locations for renewable energy installations (Guo et al., 2023; Isabelle & Westerlund, 2022; Schiff et al., 2021; Schwartz et al., 2021). In health care, AI has been lauded for its application to medical imaging, medical diagnosis, robotic surgery, and even the creation of new drugs and vaccines (Isbanner et al., 2022; Morley et al., 2020; PricewaterhouseCoopers, 2017).

In public services and social welfare, AI has been used to triage access to housing services, provide mental health care, translate government documents for non-native language speakers, or answer questions via chatbots, among many other “for good” applications in finance, education, and other sectors (Chui et al., 2018; Cowsls et al., 2019, 2021; Herzog et al., 2021). This movement is also associated with efforts such as applying AI to achieve the sustainable development goals (AI4SDGs) or goals related to equity (AI4Equity), development, human rights, well-being, and more (Cath et al., 2020; Mazzi et al., 2023; Schiff et al., 2020; Stahl et al., 2023; Wakunuma et al., 2022). Despite criticisms related to ethical washing, corporate capture, or narrow technological solutionism (Holzmeyer, 2021), “AI for good,” which represents a family of philanthropic uses of AI-as-a-tool, may be the most prominent manifestation of AI in the philanthropic space.

## ***1.2 AI-as-a-domain***

However, AI's implications for philanthropy extend beyond its use as a tool. There is now a prolific body of work and research considering the impact of AI on essentially every social and economic sector. Across civil society, private sector organizations, academia, and government, actors have been thinking through the challenges and risks that AI poses for discrimination, accountability, manipulation, surveillance, labor displacement, arms races, power imbalances, inequality, and much more (Attard-Frost et al., 2022; Coeckelbergh, 2020; Prunkl & Whittlestone, 2020; Schiff et al., 2022). These issues cut across individuals, communities, and populations, economic and social sectors, professional roles and disciplines, and time frames.

As a consequence, AI ethics has developed into a subfield in its own right, with branches in philosophy, policy, sociology, science and technology studies, economics, information science, communications, history, and other disciplines. The movement draws on predecessors in engineering and computing ethics, robot ethics, machine ethics, and numerous other traditions, and was driven in the 2000s–2020 by concerns about autonomous weapons, labor displacement, and algorithmic bias, among other issues. Sometimes associated with the acronym FEAT or FATE (fairness, accountability, transparency, and ethics), the ethics movement has especially advanced a focus on bias and discrimination, transparency of AI systems and governance, power and inequality, and more (Floridi & Cowls, 2019; Howard et al., 2019).

The AI ethics movement has had unusual success for an emerging technology policy domain in receiving policy attention; it is overwhelmingly common for government policy documents, as well as private sector documents focused on AI strategy, to include large sections discussing AI's ethical implications (Schiff, 2023). Academic and industry conferences, professionals in newly created job roles in the public and private sectors, and a suite of new non-profit and for-profit organizations focused on trustworthy, ethical, and responsible AI have advanced a range of policy problems and solutions to address challenges in this field (Benjamins, 2020; Maas, 2023; Perry & Uuk, 2019), with proposals as narrow as the ethical development of future computer scientists to as broad as the creation of global institutions or major cultural and economic reforms.

While this review simplifies a vast movement of actors concerned with the philanthropic implications of AI, it is helpful to distinguish this community from a related but distinct movement, the AIERMPS. The AIERMPS is more commonly associated with concepts such as AI safety, AI alignment, or extreme AI risk. This field is chiefly concerned with the study, practice, and governance of AI systems to ensure that they remain aligned with human goals and do not pose an unacceptable (or even existential) threat to human flourishing. Like the AI ethics community, the AI safety or extreme risk community is composed of researchers, organizations, and donors, as well as a growing cohort of policy actors. However, the AIERMPS differs somewhat from the AI ethics community. Indeed, it focuses substantially on technical aspects of AI safety, including the robustness, interpretability, and security of AI systems, and emphasizes severe to catastrophic risks, which are sometimes considered speculative by critics of the community. The movement has also encouraged its members to focus explicitly on AI policymaking, fundraising, and technical research to advance adequate regulation, accountability and transparency measures, independent auditing, export controls, and other regulatory regimes that could mitigate AI-induced risks.

To some extent, the AI ethics and AI extreme risk mitigation communities share similar goals and strategies (Baum, 2018; Stix & Maas, 2021). Actors in both subcommunities may be concerned about privacy, manipulation, deception, autonomy, misinformation, trustworthiness, independent auditing, negative environmental impacts, labor displacement, human rights, and so on. However, the AI extreme risk mitigation community is particularly notable in its focus on catastrophic, extinction-level, or “existential” risks that could arise from AI-enabled threats to physical and financial infrastructure, manipulation of humans, biological weapons, cyber warfare or terrorism, and rogue AI systems (Hendrycks et al., 2023). Key concepts associated with this community include transformational AI, AI safety, long-termism, transhumanism, rationalism, effective altruism, artificial general intelligence (AGI), artificial super intelligence (ASI), and more. Importantly though, these broad strokes descriptions of these two communities inevitably oversimplify and mischaracterize many individuals and organizations active in these philanthropic spaces. Nevertheless, they are useful for providing a rough understanding of the unique (and non-unique) aspects of the AI extreme risk mitigation community.

To briefly summarize, we have so far presented a discussion of the relationship between AI and philanthropy, including the many uses of AI-as-a-tool that can be employed by philanthropic actors for operational improvement or for “social good” use cases, followed by a discussion of the philanthropic movements focused on AI-as-a-domain or cause area. In the next section, we turn to a particular subcommunity, the AIERMPS, given its important role in the overall AI landscape and its relative lack of scholarly and analytical attention.

## **2 The AI extreme risk mitigation philanthropic sector: culture, strategy, and organizational landscape**

### ***2.1 The founding of the sector: a strong cultural identity and shared principles***

The AI extreme risk mitigation philanthropic sector (AIERMPS), which we again define as the set of actors and organizations that seek to mitigate AI-related extreme risks through and with private philanthropy, has coevolved since its inception and to date with a very specific culture that strongly shapes its identity (Lazar & Nelson, 2023). The first major component of this culture is transhumanism, a movement defined by its “taking a decidedly positive view of the prospect of a ‘post-human’ future” (Birnbacher, 2009). Two of the first research organizations dedicated to studying AI risks have close ties to this component.

Founded in 2000 by Eliezer Yudkowsky, originally as the “Singularity Institute for AI,” the Machine Intelligence Research Institute (MIRI) was one of the first organizations to work on AI extinction risks, even before AI gained public attention in the 2010s. While its initial focus was on *accelerating* the development of artificial general intelligence (AGI), that is, an AI more capable than humans at any cognitive task, MIRI began working in 2005 on countering the risks that such an AI might pose to humanity. This turn came after staff realized that a superintelligence could potentially cause significant harm to humans, up to and including extinction (Nast, 2015). This formative insight, highlighting the potentially extreme harms of advanced AI systems, would become the core mission of the field of AI safety and the AIERMPS.

Starting modestly with philanthropy from a small cohort of private donors, the AIERMPS has grown steadily over time, especially as AI has attracted more attention thanks to new players such as the research-focused organizations (now industry-affiliated organizations) DeepMind and OpenAI, and thanks to major AI discoveries in the 2010s. At this time, the first large academic center dedicated to mitigating potential risks from AGI, The Future of Humanity Institute (FHI), was founded in Oxford in 2003 by Nick Bostrom, with the mission to study existential risks (including but not limited to those posed by AI) (Ó hÉigeartaigh, 2017). After more than a decade of conversations with members of MIRI on mailing lists covering a wide range of topics, including transhumanism (Taillandier, 2021), Bostrom published the book *Superintelligence* in 2013, drawing largely on a range of ideas that had emerged from members of MIRI and FHI (Bostrom, 2014). This substantially contributed to making AI risks more well-known.

Pioneers in the field, FHI and MIRI have developed idiosyncratic views that focus substantially on the technical challenge of aligning AGI with human values. First, they consider the robust alignment of AGI with core human values to be an extremely difficult technical problem (Bostrom, 2014; Yudkowsky, 2016). Second, they consider it likely that at some point, AI progress will accelerate sharply (Yudkowsky, 2013), manifesting in an exponential increase in AI’s capabilities, an event called the “singularity.” This rapid increase in capabilities can be analogized to the same way that DeepMind’s AlphaZero (one of the first superhuman-level AI Go players) went from subpar to substantially better than world champions at Go after only a few dozen hours of training (Silver

et al., 2017). Some core MIRI staff believe that the same rapid increase in AI capabilities could happen with respect to all tasks currently performed by humans, once the field of AI develops sufficiently powerful and general AI systems. Finally, a criticism of this movement has been that while bringing a fresh perspective to the field, only a minority of FHI and MIRI members have backgrounds in technical AI research, which carries the risk of imparting views and concepts that may lack relevance when applied to concrete AI systems.

However, by writing most of the early influential articles and shaping core conceptualizations of AI risks, MIRI and FHI have had a long-lasting impact on the AIERMPS, on the field of AI safety, and on AI more broadly. Even prominent industry actors such as DeepMind, a startup created in 2010 and now an industry leader and subsidiary of Google (Alphabet), have been closely linked with MIRI since its inception. Indeed, Eliezer Yudkowsky introduced DeepMind's founders, Demis Hassabis and Shane Legg, to MIRI's lead donor, Peter Thiel. Thiel subsequently became the first investor in DeepMind (Metz, 2022), the first AI company with the stated mission to develop AGI. Beyond these personal and financial connections, MIRI and FHI's intellectual conceptualizations of AI risks have also had a lasting influence. For instance, while new technological advances, especially related large language models, have led to the evolution of some concepts and frameworks applied to current AI systems (Ngo et al., 2023), many of the concepts that MIRI and FHI emphasized, such as "agents," "corrigibility," and "alignment," are still widely discussed and actively used in the AI safety research field (Byrnes, 2021) as well as in policy, evidenced by the creation of international AI safety workshops and the U.S. NIST AI Safety Institute.

During the late 2000s and the early 2010s, large segments of the transhumanist component of the culture evolved into a new subculture, significantly shaped by MIRI and its leader Eliezer Yudkowsky, now commonly referred to as "rationalists" (Matthews, 2023). Built around a set of concepts, ways of thinking, and idiosyncratic preferences around discourse, all dedicated to improving rationality, MIRI and the surrounding community have fostered a remarkable subculture. For example, a unique component of this subculture is that much of it has developed, and continues to be expressed and practiced, through a shared blog called *LessWrong*, based on a core set of writings by Yudkowsky known as "The Sequences."

This culture values norms that differ from those more commonly used in everyday society, such as its openness to inconvenient truths, high standards for what constitutes acceptable or productive discourse, and a preference for systematic and quantitative calculation to determine true or ethically sound positions. Because of its influence on the AIERMPS, it is partly responsible for the criticism that AI safety is characterized by a "near-monoculture" (Lazar & Nelson, 2023). Beyond the AI risk mitigation philanthropic sector, this culture has also affected AI industry leaders (Matthews, 2023), in large part because the culture has become very prominent in Silicon Valley and technology circles, where most of the top AI companies working to develop AGI have been founded. Thus, intellectual, social, historical, and even geographical factors have played a role in shaping the AIERMPS.

Finally, another critical dimension of this community with a special relationship to philanthropy is the community focused on evidence-based philanthropy. In the early 2010s, and increasingly over time, the rationalist culture has been increasingly influenced by the rapidly growing effective altruism (EA) movement. Coined in 2011 by a group of Oxford academics, the EA movement emerged from the convergence of a focus on rationalism (how to think better), altruism (how to organize philanthropy and charitable giving more efficiently, including making career decisions), and futurist concerns (how to ensure that human civilization thrives) (Chivers, 2019).

Rooted in the prominent framework of normative moral philosophy, utilitarianism, and centered around a small but powerful core of principles—rationality and altruism—the EA movement



has itself become increasingly influential and even dominant in the AI risk philanthropy landscape. Among the reasons for its influence was a shift away from the original views of MIRI and FHI, which were more closely linked to the academic field of AI, and instead toward influence through industry and policymaking, including cultivating major financial backing to enable these goals. This ambition has been reinforced by the arrival of other major players in the field, to whom we turn to next.

To summarize, the early years of the AIERMPS are characterized by the emergence and convergence of several communities steeped in academic and social philosophies and subcultures surrounding rationalism, transhumanism, futurism, and later, evidence-based philanthropy. While the associated ideologies and cultures have evolved over time, many of the core principles have remained highly operative and even determinative in the AIERMPS.

## ***2.2 A contesting power: the rise of AI industry in AI safety***

The year 2015 marked a significant turning point in AI extreme risk mitigation philanthropy, with the entry of new philanthropic players less tied to the initial cultural epicenter of the AIERMPS, MIRI. One of the major drivers of these new resources was the increased interest of tech billionaire Elon Musk, which brought new funding and visibility to the ecosystem. Musk began his philanthropy in the AIERMPS by providing a \$10 million gift to the newly founded Future of Life Institute (FLI), a non-profit organization co-founded by MIT professor Max Tegmark that is dedicated to mitigating risks arising from advanced AI systems (Kosoff, 2015). Additionally, Musk played a leading role in the founding of the (then entirely non-profit) organization OpenAI, donating \$1 billion to the organization in conjunction with donations from Peter Thiel, Sam Altman, and other high-profile donors (Markoff, 2015). This influx of resources empirically dwarfed existing philanthropic efforts related to AI and AI safety.

The stated goal of OpenAI (now a subsidiary of Microsoft), like its already existing rival DeepMind, was to build AGI, which it defines as “highly autonomous systems that outperform humans at the most economically valuable work” and to ensure that it “benefits all of humanity” (OpenAI, 2018). Initially true to its name, OpenAI initially placed openness at the center of its philosophy, intending to open source its technology and even collaborate with competitors to foster beneficial rather than harmful AGI. However, it later changed its mind on its approach to openness, at least in part to growing concerns about risks arising from putting such a powerful technology in the hands of anyone, including malevolent actors.

Consequently, 2015 and 2016 marked the beginning of a power shift within the AIERMPS away from the MIRI-centered academic ecosystem toward a set of new actors from the effective altruism movement who held different views on how best to address AI safety and mitigate AI risks (Christiano, 2022; Karnofsky, 2012). Some prominent members of the former MIRI-centered ecosystem joined and led the core organizations of this group. Notably, this new landscape of actors operated under comparatively less pessimistic assumptions than did the original MIRI cluster. For instance, individuals in the community and their organizations were less pessimistic about the difficulty of making AI safe, believing that the fundamental technical problems were not as intractable. Accompanying this stance was their view that AI capabilities would evolve more gradually rather than suddenly, making incremental research and learning about effective governance more feasible. This set of positions would become the lynchpin of many subsequent disagreements in the field over the next decade about the viability of different strategies, organizations, and philanthropic priorities.

Embodied by three key individuals, research scientists Dario Amodei and Paul Christiano and philanthropist Holden Karnofsky, this new worldview increasingly shaped the trajectory of the

AIERMPS and the AI industry beginning in 2015 (Amodei et al., 2016; Karnofsky, 2012). The trajectory of these leading individuals also strengthened the AIERMPS's important connections with AI industry leaders (Coldewey, 2021; Piper, 2023). This became particularly important because the vast majority of AI-related research and development occurs in the private sector rather than the public sector and even within a handful of leading companies, making them the epicenter of AI's trajectory.

OpenAI was influenced by that worldview, at least for the first few years of its existence. This led them to prioritize and produce empirical research, which became one of the main drivers in bringing AI safety into the mainstream academia, with seminal research such as *Concrete Problems in AI Safety*, by Amodei, Christiano, and other leading researchers. With a focus on technical research, working closely with the AI industry to steer the trajectory of AI toward its definition of safety, and access to the best existing AI models and industry-leading resources, Christiano and Amodei have played an influential role in the nascent AGI industry. Both were present at OpenAI's founding dinner (Brockman, 2016), and each led the AI safety team at OpenAI at different points in time.

After 2020, they both left to start new institutions: Amodei left OpenAI to create a competing AI industry player, Anthropic (responsible for the Claude series of models), and Christiano left to create the Alignment Research Center, a research non-profit focused on research at the intersection of AI safety theory and deep learning AI systems. This center also later incubated an auditing organization, which aims to evaluate systems created by AI industry players. Both institutions became increasingly influential in the field of AI, starting in 2021, reinforcing the ties between parts of the AIERMPS and leading AI industry actors.

Importantly, the research and industry arms of the movement have been significantly fueled by its growing philanthropic arm. As CEO of Open Philanthropy, a major philanthropic foundation that gives tens to hundreds of millions of dollars annually to a wide range of causes, Holden Karnofsky developed an AI risk philanthropy program beginning in 2016. Under his leadership, Open Philanthropy quickly became the main grantmaker of the AIERMPS, cumulatively administering \$330 million in grants for the field by 2023 (Open Philanthropy, 2023b).

This new funding enabled the field to grow substantially and in many directions: training young researchers, organizing conferences, providing fellowships and internships, supporting academic research, fostering independent researchers, sustaining organizations such as MIRI and FHI, and enabling the creation of a variety of new organizations and initiatives tied to the ecosystem. Indeed, it is difficult to overstate the importance of this large and sustained influx of funding, which has allowed Open Philanthropy to express its strategic preferences and ideology through numerous channels. Open Philanthropy's strategic views and influence on the AIERMPS also served to further connect this philanthropic community to the industry, allowing it to have an ongoing influence on the development of AI. Ultimately, these efforts, fostered over only a little more than a decade, positioned the AIERMPS to be at the forefront of AI when it exploded into the public eye in the 2020s.

The 2010s then represented an important period in the evolution of the AIERMPS. With important leaders in research, industry, and philanthropy reaching new levels of relevance, and armed with increased funding and the support from prominent academic experts and a growing pool of philanthropists, the AIERMPS was ready for the mainstream.

### **2.3 From niche to mainstream**

After gaining an influential position in the laboratories and boardrooms of leading AI industry players, the AIERMPS became even more central thanks to a major shift in public and political

awareness of AI. Namely, the release of AI systems such as DALL-E 2, ChatGPT, Microsoft Bing, and Google Bard represented an important shift that first came to prominence in November 2022 with the public release of ChatGPT by OpenAI. Whereas AI systems had largely been discussed in specialized circles, the general public and policymakers have now become aware of the implications and risks of AI (Miyazaki et al., 2023).

For instance, the uptake of generative AI tools for academic misconduct and embedded in popular search platforms increased the salience of AI for students, parents, and workers alike. The public and decision-makers were increasingly exposed to high-profile AI-related failures involving data leaks, offensive content, misinformation, and bias, leading to organizational bans on large language models and increased regulatory attention. As an example of such an early focus event documented in leading publications, a New York Times article detailed a chatbot's attempt to persuade a journalist to leave his wife (Roose, 2023b). These and other incidents heightened awareness of AI-related concerns, though matched by the contemporaneous eagerness with which numerous organizations sought to adopt AI systems into their workflows.

This growing concern among experts and the general public was amplified by three events, all of which led to major news cycles and intensified debates over the course of several months:

- An open letter for an “AI pause” was signed by major actors, such as Elon Musk and Turing Prize winner Yoshua Bengio (Future of Life Institute, 2023);
- Another prominent AI expert and Turing Prize winner, Geoffrey Hinton, known as the “Godfather of AI,” resigned from his position at Google, citing escalating concerns about AI risks as the core reason (Metz, 2023);
- A one-sentence statement from the Center for AI Safety, stating only that “mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war,” was signed by many of the world’s leading AI figures, including the CEOs of the top AI companies, hundreds of professors, and top AI researchers (Roose, 2023a).

This major shift in the landscape heralded and accompanied significant changes for the AIERMPS as a whole. First, the movement shifted a large portion of its focus to AI policy and governance, compared to its predominant focus on technical AI safety research. For example, major funders such as Open Philanthropy rapidly ramped up their grantmaking in this area (Open Philanthropy, 2023a). The sea change also led to a large influx of interested talent, stakeholder discussions, and public and policymaker attention, causing a surge of both opportunities and demands for the small number of experts in the field. Finally, a massive increase in investment in AI by leading technology companies contributed to a power shift from the non-profit and academic AI sector to the industrial sector. Due to the massive inequality of resources which allowed only the largest companies access to the data and infrastructure needed to develop AI models as well as the capacity to offer extremely high wages, industry players were far better positioned to attract top talent, develop leading AI models, and have access to information that enabled cutting-edge research (Mickle, 2023).

In response to this newly acquired publicity, a growing segment of individuals associated with the AIERMPS have endorsed strategies more reminiscent of traditional public-facing activism (Meaker, 2023). Their objective is to solicit public and policy attention to counterbalance the expanding influence of the AI industry, which continues to advance AI, despite arguably heightened risks. An “AI pause” is one of the core themes proposed and supported by these segments of the movement, with calls for a moratorium on AI development until developers of AI systems

can guarantee that their next AI systems will be developed and deployed safely. The contest between “accelerationist” and “decelerationist” philosophies is ongoing, and reflects the role of the AIERMPS at a moment in time when AI advances, governance, and public attention make these issues perhaps historically urgent and reflective of a contingency point.

In summary, this section discussed the history of the AIERMPS and its evolution since the early 2000s. From humble beginnings with a passionate and unique subculture based on shared principles and focused on theoretical work, it evolved into a sector with significant resources, important industry ties, and a more applied research and philanthropic approach, before rapidly reinventing itself in light of a massive increase in interest in AI risk mitigation. Familiarity with the ideological and cultural evolution of the AIERMPS is critical to understanding the ongoing challenges facing the movement and its possible trajectory, which we detail in the next section.

### **3 Challenges faced by the AI extreme risk mitigation philanthropic sector**

#### ***3.1 A sector bound to disagree despite a strong willingness to agree***

The AIERMPS faces several major difficulties in its efforts to mitigate AI risks. One such difficulty is that many catastrophic risks—and extinction risks in particular—are associated with events that are rare. Indeed, an extinction event is, by definition, unique and irreversible. As such, there may be few or no instances from which to learn lessons about mitigation successes or failures, limiting the possibility of feedback loops that could improve the movement’s prospects for learning and improving its philanthropic strategies.

Even in the *absence* of such a catastrophic event, it is difficult to attribute this potential success to specific interventions or to even determine whether the interventions are causally responsible for mitigating these risks. Hence, philanthropists in the AIERMPS, who focus on evidence-based grantmaking, must rely on indirect indicators to measure whether a given funding initiative or strategy is effective. Such indicators often derive from and rely on strategic views about which trajectories are likely to increase or decrease the likelihood of catastrophic risks. Not unlike other social movements, then, differences in underlying philosophy can lead to wide disagreements among grantmakers, despite a shared culture centered on rationality.

Lightcone Infrastructure is an example of an organization with a central role in the AIERMPS that decided to shut down one of its core programs because of serious concerns about whether it had had positive or negative impacts, given the opportunities it had to help accelerate AI progress (Habryka, 2023). The AIERMPS frequently refers to this concern as “capabilities externalities,” implying that advancing AI progress also entails the creation of externalities such as increased AI risks. Thus, whether to advance (accelerate) or alternatively limit (decelerate) AI progress is itself a recurring question, debated as grantmakers and strategic leaders assess the value of various research and funding directions.

A related complicating factor is that many of the strategies, often deemed necessary to ensure AI’s safe development, rely critically on predictions of how actors will behave when these risks are elevated. This additional degree of contingency built into speculative scenarios leads many actors in the AIERMPS to disagree about whether leading AI industry actors such as OpenAI or Anthropic are beneficial or negative overall. One part of the community argues that these organizations are much more concerned about risk than other existing industry actors like Meta, owing to their strong tradition in AI safety. The other side argues that these organizations have ironically been the main culprits responsible for accelerating AI progress and AI risks, and are additionally skeptical that their efforts in technical safety offset this effect (Matthews, 2023).

This state of affairs constrains, for example, the ability of actors in the movement to robustly judge whether an industry actor is beneficial or harmful. Strategic planners may need to rely on information that may be absent, inherently available only in the future, or otherwise difficult to measure, and thus ripe for disagreement, such as:

- The likely technical trajectories of AI development, including what types of advances are needed for AI to become increasingly or decreasingly risky;
- The quality of the company's contributions to AI safety research;
- A judgment of the company's culture, including the extent to which employees care about and focus on mitigating AI risks;
- The true intentions of the company's CEO and leadership; and
- Whether the company or regulation is likely to be captured by industry motives.

This situation implies that many disagreements about grantmaking strategy are difficult to resolve, due to the lack of empirical evidence and clear metrics for success or failure. Contrast the challenges here with the environmental movement, for example, which has many metrics for measuring outcomes such as wildlife diversity, water quality, or carbon dioxide emissions.

In addition to these disagreements, another core aspect of the culture of the AIERMPS, which is also grounded in its focus on evidence-based philanthropy, presents a further challenge. Because of its origins in academic philosophy and its preference for economic methods that favor causal inference, the AIERMPS places significant emphasis on understanding counterfactual impact, including the opportunity costs of philanthropy. The concept is simple: if Alice receives a grant to pursue a line of research, the impact of the grant is not merely Alice's research output, but rather the difference in the advancement of that line of research *compared* to the situation in which Alice had not pursued this line of research. In such an alternative scenario, a grant could have gone to Bob instead, or Alice might have pursued another more or less promising research direction. As an example, if Alice is pursuing a research direction that ten other research teams are also pursuing, then the true impact of her research might be far more minor, given that if she hadn't pursued her research, other teams might have found her results anyway.

Core to the culture of effective altruism (Gabriel, 2017), this reasoning adds yet another layer of uncertainty to the evaluation of the impact of philanthropic interventions. An example widely discussed in the AIERMPS is that of Open Philanthropy, one of the core actors in the sector described previously, which granted \$30 million to OpenAI. Those in the community who argue that this grant was harmful tend to say that it contributed substantially to the success of OpenAI overall, and thus helped OpenAI to accelerate AI progress writ large. Thus, by contributing to the success of its grantee, Open Philanthropy has accelerated AI progress, which, in the eyes of this community, unacceptably increases AI risks. In response to this criticism of Open Philanthropy, others presented a counterfactual: if Open Philanthropy had not given \$30 million, OpenAI would not have failed, but would simply have raised funds slightly earlier or from other actors. In effect, this would have made little difference to the overall progress of AI, while Open Philanthropy's strategic giving might have increased its "say" in OpenAI's approach including an increased focus on AI safety (Moskovitz, 2023). This discussion illustrates how considering counterfactual impact, in addition to other measures of impact, can make it especially hard for actors in the AIERMPS to reach agreement on even core decisions.

Other considerations such as technical and geopolitical perspectives constitute areas where disagreement can affect grantmaking strategies pursued by various actors of the AIERMPS. One such persistent disagreement, another defining feature of the strategic decisions considered by the

AIERMPS, is the assessment of how difficult or easy it is to make powerful AI systems “safe.” One perspective, with increasing following, is that AI safety research should *leverage* the development of large language models until they are sufficiently advanced to automate AI safety research itself (Karnofsky, 2022). As a consequence, this cluster of the community is comparably more enthusiastic about the safety strategies favored by AI industry players such as OpenAI or Anthropic. This contrasts with other segments of the AIERMPS who believe that AI safety is a much harder problem, that the use of large language models will not succeed and may backfire, and that the most desirable policy to pursue is to enforce a pause in advanced AI development until a technical solution to AI risks is in sight (Yudkowsky, 2023).

A first set of challenges faced by this community, then, relates to the numerous compounding difficulties in assessing the efficacy of their strategies. Determining an effective philanthropic strategy for AIERMPS’s causes is no easy task. This is because there is little actual empirical evidence available, as the events in question are definitionally rare and because strategic decisions rely on forward-looking predictions which are contingent on uncertain technical advances and human behavior. Even such core questions as whether advancing or halting the progress of AI are more effective in risk mitigation remain controversial. While many social movements face disagreements, some of the factors faced by the AIERMPS are indeed unique.

### ***3.2 A philanthropic sector that needs strong adaptation abilities***

Another difficulty the AIERMPS had to face was that its originating community preceded most of the other key institutional actors that would later address AI risks, including those from academia. In practice, the leading actors had to determine what strategic directions were most likely to advance an otherwise non-existing field, and were thus limited in their ability to draw on surrounding infrastructure and perspective. This required, for example, making key direction-setting decisions such as funding individuals or teams in the absence of established signals of credibility, due to the absence of specialized curriculum, established university programs and credentials, or proven success in the field.

This led the AIERMPS to pursue a wide range of approaches, including creating a large infrastructure to enable a new research, policy-engaged, and philanthropic field. Activities have included funding online courses on AI risks (sometimes taken by students without background in the typical prerequisites), establishing forums and platforms dedicated to sharing AI safety research (using formats without the traditional checks and rigors of academic research, such as formal peer review), or funding non-profit organizations led by individuals from non-traditional academic paths to train cohorts of young researchers and build the community. One example is SERI MATS, a non-profit organization providing six-month training programs which pair candidates with experienced mentors to teach them AI safety research skills.

A well-known defining characteristic of AI since the 2010s is the pace at which the technology and the surrounding landscape evolve. This subsequently urges the AIERMPS to frequently change its grantmaking strategy and be continually forward-looking. For instance, while the technology underlying ChatGPT called *transformers* (Vaswani et al., 2023) was brand new and still a proof of concept in 2017, it is now widely deployed and considered one of the most powerful and standard technologies in AI. Similarly, the AI policy landscape has radically changed after the release of ChatGPT in November 2022.

For the AIERMPS, the only way to make sure that their giving can effectively address the dynamic challenges of AI development, governance, and risk mitigation at any given time is to remain up to date on the latest developments of the technology, of the fast-evolving landscape,

and to employ grantmakers with relevant technical and policy backgrounds. In practice and due to associated uncertainties, this can limit the capacity of grantmaking organizations to robustly evaluate a wide number of possible grant projects. Such a problem is particularly exacerbated by how difficult it is to hire and retain grantmakers with unusually high technical skills (e.g., advanced machine learning skills), another unique feature of the AIERMPS compared to other philanthropic movements. This dynamic may also make it more difficult for donors to engage in long-term giving commitments, as any specific research direction, policy solution, or funding initiative could be made obsolete by a new breakthrough.

In addition to the pace of AI, the field of AI safety research is notable for being at an extremely early stage, leading some to call the field “pre-paradigmatic” in that there is no strong agreement on the most crucial areas to work on (Hernandez-Orallo et al., 2020). On the one hand, grantmakers want their efforts to help mitigate risks relevant to the current state of the technology. This has driven much grantmaking on large language models, such as research on the explainability of the most advanced AI systems. On the other hand, the lack of clear plans for solving key technical safety problems also suggests a need to provide grants to help explore radically new approaches that may be largely speculative and even ineffective. Both are risky philanthropic approaches. Improper calibration toward “safer” or “riskier” grantmaking could undermine the core goals of the AIERMPS; striking an appropriate balance is similarly difficult. Informed grantmaking in this context hence requires grantmakers with a large range of technical (as well as social, political, and geopolitical) knowledge, in order to be able to evaluate numerous scientific areas. It also leads grantmakers to rely on external advisors and technical experts to properly evaluate proposals in highly narrow fields, introducing further complexity.

There are still remaining issues for philanthropic strategy when a subfield is very new. Some areas that are considered promising are pursued by no more than a few dozen individuals, which could severely undermine the movement’s effectiveness if the donor strategy turns out to be less than optimal. This risk applies, for example, to a highly specialized branch of AI safety research called *Infrabayesianism* (Kosoy, 2020) or Open Agency Architecture (Dalrymple, 2024). Compounding this challenge, limited early investment or the inability to even initiate work in potentially promising areas can preclude the ability of funders to understand whether these trajectories are promising.

In combination, the unique cultural and ideological characteristics of the AIERMPS, its novelty, the high uncertainty surrounding key factors that affect the impact of grants, the pace of development of AI, and several other key factors mean the AIERMPS has had to face a significant number of challenges in its philanthropic strategy. This has led it to foster significant adaptability and flexibility to react to dynamic developments as they arise. Beyond its idiosyncratic challenges, the AIERMPS has also had quite unusual relations with other actors in the AI field, in contrast to how other social movements have evolved. We discuss these unique relationships and the associated opportunities and limitations they presented in the next section.

## **4 What the AI extreme risk mitigation philanthropic sector could learn from other fields and social movements**

### ***4.1 Commonalities and differences with other social movements***

Early on, the AIERMPS framed AI risk mitigation primarily in terms of technical problems that researchers needed to solve, despite its understanding of AI risk as, at its core, a societal issue that

could affect everyone. As Ó hÉigartaigh summarized in his 2017 overview of the field, much technical AI research

has focused on translating some of the more foundational questions raised by early work at FHI and MIRI and elsewhere into crisp technical research problems that can be worked on today. This includes approaches involving fundamental mathematical frameworks for agent decision-making and behavior, as well as research programs exploring how some of the behaviors that would be of concern in long-term systems may manifest in the near-term systems we are building currently.

This technical view of the issue might also be partially driven in part by the “tendency to valorize corporate-driven tech solutions” that Broad (2018) notes regarding the effective altruism movement, as well as the movement’s general origins in elite and intellectual research communities rather than activist or advocacy circles.

This focus contrasts with other social movements such as climate change or animal welfare, which were instead framed early on as predominantly social problems (McCright & Dunlap, 2000; Singer, 1975), affecting the set of interventions and approaches that are considered viable to solve these problems. This focus also impacts which kinds of actors are deemed relevant (or irrelevant) in shaping the movement’s aims. If progress requires advanced computer science research rather than, say, cutting down on one’s consumption of meat or recycling in the workplace or public protest, the space for public engagement is de facto diminished. Moreover, the types of concerns raised were not historically promoted to the mass public in the way that other social movements were, meaning that AI risk remained an issue of low public salience until it exploded into public attention.

Thus, despite the continued rise in interest since the late 2010s of the AIERMPS in the social aspects of AI risks, such as how to govern AI (Ó hÉigartaigh, 2017), the movement has not yet seen the development of a social movement comparable in size or magnitude to the climate change or animal welfare movements. The largest existing AI safety social movement to date is still very small, with overwhelming participation from experts rather than the general public (Meaker, 2023), and little evidence of penetration into the general public’s consciousness.

This technical coloration of the movement prevents the AIERMPS from using strategies that other social movements have used to achieve comparable goals. While a comprehensive review of relevant strategies, commonalities, and differences is beyond the scope of this chapter, some examples help to illustrate the point. For instance, like the climate change movement and the animal welfare movement, the AIERMPS faces a problem in the 2020s that essentially surrounds the behavior of a few corporations. Hence, the AIERMPS could likely learn from how other movements have approached these dynamics.

In *Ethics Into Action*, written in 1998, Singer shows how Henry Spira, one of the pioneers of the animal welfare movement, achieved significant social change among corporations despite having few resources by using both (external) adversarial and (internal) cooperative strategies. He used these strategies to respectively acquire bargaining power and to use it through interactions with the corporations to achieve concrete outcomes. By mixing concrete threats to the public relations of organizations that mistreated animals, such as *McDonald’s* or *Revlon*, with frequent interactions with employees from these companies, Spira was able to increase reputational pressure on these firms through targeted ads against the companies and demonstrations until they made the concrete changes Spira demanded, ultimately improving animal welfare.



Another example of how the lack of a strong social movement in AI safety limits the field's ability to achieve social change is its inability to tap one of the core mechanisms of the animal welfare and climate change movements known as the radical flank effect (Evans, 2023; Lange, 1990; Simpson et al., 2022; Singer, 1998). The radical flank effect describes the effects that radical activists have in increasing the likelihood of counterparties to negotiate alternatively with more moderate activists, or by changing the public issue framing around which strategies are considered moderate or radical. Critically, this effect has been found to have both negative and positive implications and is still much debated in the academic community. However, it arguably reflects another element missing from the AERMPS movement.

Still, other ways in which the AIERMPS's unique composition and history limit its effectiveness may revolve around voting, lobbying, fundraising, and overall resilience to threats. Limiting fundraising to a small number of wealthy, dedicated donors renders movements dependent on those donors and the risks they pose, as evidenced by the Sam Bankman-Fried scandal (Kim, 2022), while alternatively cultivating a broader base of funding support might increase resilience. Similarly, without broad-based public support, individuals are unlikely to engage in protests, walkouts, create civic clubs or student groups, call their political representatives, and so on.

A shorthand way of making this critique is that the movement remains centered on elite and esoteric perspectives aimed at rationally identifying optimal strategies. However, the animal welfare, environmental, civil rights, and gay rights movements have arguably not achieved such levels of success through rational efficacy alone. Additional exploration and serious incorporation of multidisciplinary perspectives could help the AIERMPS learn from these insights.

#### ***4.2 Contingent coalitions with natural enemies and contingent tensions with natural allies***

On top of those distinctive aspects, the AIERMPS has pursued unusual coalition-building strategies to advance its goals. One of these unusual characteristics is that the social networks of those who are developing (potentially harmful) AI systems and those who are trying to *prevent* those harms are closely tied together (Lazar & Nelson, 2023). By analogy, imagine dedicated environmental activists and oil executives working as close friends (Alexander, 2022). As discussed previously, many of the most successful AI companies are closely tied in multiple ways to prominent members of the philanthropic sector and, more broadly, to the rationalist community. While there may be benefits to such ties, for instance in terms of policy learning, they may also contribute to limiting the ability of core organizations of the AIERMPS to fund interventions that would be seen as too hostile to AI industry leaders. The potential for conflicts of interest is real.

Nevertheless, this link has allowed the AIERMPS to form highly atypical coalitions that have arguably had a major impact on public discourse. Today, several of the leading AI organizations are managed by figures in contact with the movement, and many have prominent mission statements that explicitly call for the development of responsible or safe AI, with dedicated teams focused on AI safety. As an example, the prominent AI extinction risk statement produced by the Center for AI Safety generated a major news cycle in large parts thanks to the signatures from essentially all of the CEOs of leading AI firms (Roose, 2023a). This suggests that the AIERMPS, despite its nature as an elite and technical community, has fostered a growing ability to build unusual but powerful coalitions capable of shaping policy and public discourse toward their perspective on AI risks.

This can be better understood in light of Van Dyke and Amos (2017) who explain which factors are crucial to coalition-building. Among these factors, the AIERMPS has cultivated:

- Strong social ties with prominent “bridge builders,” such as the Future of Life Institute, which has organized major conferences with a broad range of actors (Ó hÉigeartaigh, 2017);
- A shared ideology and culture across many organizations (Chivers, 2019);
- Increased political opportunities due to the rise in interest in the issue, which, according to political opportunity theory, raises the chances of successful coalitions; and
- A significant amount of philanthropic resources available to be deployed, allowing organizations in the movement to dedicate some resources to coalition-type activities.

Paradoxically, while the AIERMPS has maintained unusually strong ties with industry actors, it has largely failed to form coalitions with the AI ethics communities, whose interests and goals arguably make them much more natural allies than industry players. Some cultural and ideological differences have turned into conflicts that may have calcified actors and made coalition-building harder.

The original seeds of the disagreement, as Prunkl and Whittlestone (2020) explain, arose from disagreements over which issues were more important when attempting to minimize the harms (or risks) from AI. To simplify, those closer to the “AI ethics” side argue that AI safety concerns are overblown, even pseudoscientific, and a distraction from what they view as more pressing problems, such as bias or privacy concerns arising from already existing AI systems. In contrast, some closer to the “AI safety” perspective suggest that the problems they focus on are catastrophic or existential in magnitude, and dwarf so-called “short-term” AI ethics concerns in importance.

This disagreement became cemented by episodes where lack of mutual support and contestation over public attention and funding have created resentment. Two recent examples illustrate this conflict:

- Some prominent voices in the AI ethics community blamed Hinton for leaving Google in 2023, due to his concerns about AI safety, when he had not similarly reacted two years earlier when prominent AI ethics advocate Timnit Gebru was pushed to resign from Google after having been asked to not publish one of her papers related to the problems of large language models (Chan, 2023).
- Despite having founded the Distributed Artificial Intelligence Research Institute (DAIR), an organization focused on preventing harms of AI, and aiming to launch a “Slow AI” movement in 2022 (Strickland, 2022), she along with other prominent AI ethics researchers severely criticized statements on the AI pause proposed by an organization from the AI safety movement, due to their emphasis on AI extinction risks (Sætra & Danaher, 2023).

To resolve these disagreements and conflicts and create room for greater impact overall, various proposals have been offered. Stix and Maas (2021) emphasize many avenues for positive collaborations, such as the study of available policy levers that would help achieve changes that both camps find amenable, or jointly pushing for mechanisms to maintain the integrity of public discourse in the face of AI systems. Prunkl and Whittlestone (2020) emphasize how the division between “long term” (AI safety) and “short term” (AI ethics) risks overemphasizing these differences, including their associated time scales. Instead, they propose four dimensions that could better identify disagreements in prioritization and thus foster the potential for collaboration.

More systematic discourse and shared analysis could help the communities identify common ground on which AI capabilities to focus on, when to focus on current or future impacts, whether to focus on more or less uncertain issues, and whether to focus on extreme risks or risks at all scales. As Sætra and Danaher (2023) identify, the movements could endeavor to build bridges to achieve common goals and avoid a situation where “neither short- nor long-term risks are managed and mitigated,” which would represent a failure of both communities. For the AIERMPS to achieve its goals effectively then, it may need to continually revisit both its unusual alliances with industry actors and its disagreements with actors in adjacent communities.

## 5 Conclusion

This chapter began with a review of the status of AI and philanthropy, articulating a distinction between AI when used as a tool to advance numerous aspects of philanthropic practice, and AI when considered as a domain or cause area. Here, we focus on the latter, presenting a history and evaluation of the prominent and increasingly important community focused on extreme AI risks.

We reviewed its unusual intellectual and philosophical origins in rationalism, effective altruism, and technical safety research before discussing its transition to public relevance. Some of the movement’s unique features may have played a role in its recent successes, such as close alliances between leading AI industry actors and AI safety researchers. However, the movement’s largely elite nature and difference from typical broad-based social movements also pose limitations and threats to its viability, including distancing it from potentially natural allies.

This chapter only begins to articulate some important characteristics of the movement. Substantial research, including historical analysis, interviews, studies in management and political science, and so on, is needed to unpack many related issues, understand possible trajectories, and provide analysis to evaluate, achieve, and perhaps modify the movement’s aims and efficacy. As a starting point, we suggest greater research is needed to understand the social and intellectual history, political coalitions, and trade-offs involved with the movement, as well as the movement’s positionality in broader philanthropic and AI circles.

For individuals who are members or observers of the movement, this chapter echoes calls for multidisciplinary engagement, learning from outside perspectives, and learning from the successes and failures of other social movements. We suggest here, echoing numerous commentators, that direct engagement with other social movements, critical scrutiny of current alliances, and efforts to build bridges across coalitions could be prudent, along with deeper engagement with the general public. While the impact of the AI extreme risk philanthropic sector is yet to be fully understood, it is likely to be monumental.

## References

- Alexander, S. (2022, August 8). Why not slow AI progress? *Astral Codex Ten*. <https://www.astralcodexten.com/p/why-not-slow-ai-progress>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI Safety. *arXiv: 1606.06565 [Cs]*. <http://arxiv.org/abs/1606.06565>
- Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2022). The ethics of AI business practices: A review of 47 AI ethics guidelines. *AI and Ethics*. <https://doi.org/10.1007/s43681-022-00156-6>
- Baum, S. D. (2018). Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & SOCIETY*, 33(4), 565–572. <https://doi.org/10.1007/s00146-017-0734-3>
- Benjamins, R. (2020, May 22). A new organizational role for Artificial Intelligence: The responsible AI champion. *Think Big*. <https://business.blogthinkbig.com/a-new-organizational-role-for-artificial-intelligence-the-responsible-ai-champion/>

- Birnbacher, D. (2009). Posthumanity, transhumanism and human nature. In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Posthumanity* (pp. 95–106). Springer Netherlands. [https://doi.org/10.1007/978-1-4020-8852-0\\_7](https://doi.org/10.1007/978-1-4020-8852-0_7)
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies* (First edition). Oxford University Press.
- Braverman, M. T., Constantine, N. A., & Slater, J. K. (2004). *Foundations and Evaluation: Contexts and Practices for Effective Philanthropy*. John Wiley & Sons.
- Broad, G. M. (2018). Effective animal advocacy: Effective altruism, the social economy, and the animal protection movement. *Agriculture and Human Values*, 35(4), 777–789. <https://doi.org/10.1007/s10460-018-9873-5>
- Brockman, G. (2016). *My Path to OpenAI*. <https://blog.gregbrockman.com/my-path-to-openai>
- Byrnes, S. (2021, December 14). Consequentialism & corrigibility. *AI Alignment Forum*. <https://www.alignmentforum.org/posts/KDMLJEXTWtkZWheXt/consequentialism-and-corrigibility>
- Cath, C., Latonero, M., Marda, V., & Pakzad, R. (2020). Leap of FATE: Human rights as a complementary framework for AI policy and practice. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 702–702. <https://doi.org/10.1145/3351095.3375665>
- Chan, W. (2023, May 5). *'I Didn't See Him Show Up': Ex-Googlers Blast 'AI Godfather' Geoffrey Hinton's Silence on Fired AI Experts*. Fast Company.
- Chivers, T. (2019). *The AI Does Not Hate You: The Rationalists and Their Quest to Save the World*. Weidenfeld & Nicolson.
- Christiano, P. (2022, June 19). Where I agree and disagree with Eliezer. *LessWrong*. <https://www.lesswrong.com/posts/CoZhXrhpQxpy9xw9y/where-i-agree-and-disagree-with-eliezer>
- Chui, M., Harrysson, M., Manyika, J., Roberts, R., Chung, R., Nel, P., & Heteren, A. van. (2018). *Applying AI for Social Good*. McKinsey Global Institute. <https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good>
- Coeckelbergh, M. (2020). *AI Ethics*. MIT Press. ISBN: 9780262357074
- Coldewey, D. (2021). Anthropic is the new research outfit from OpenAI's Dario Amodei. *Yahoo News*. <https://www.yahoo.com/now/anthropic-ai-research-outfit-openais-175923024.html>
- Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). *Designing AI for Social Good: Seven Essential Factors* (SSRN Scholarly Paper ID 3388669). Social Science Research Network. <https://doi.org/10.2139/ssrn.3388669>
- Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, 3(2), Article 2. <https://doi.org/10.1038/s42256-021-00296-0>
- Dalrymple, D. (2024). Safeguarded AI: Constructing safety by design. *Aria*. <https://www.aria.org.uk/wp-content/uploads/2024/01/ARIA-Safeguarded-AI-Programme-Thesis-V1.pdf>
- de Souza Leão, L., & Eyal, G. (2019). The rise of randomized controlled trials (RCTs) in international development in historical perspective. *Theory and Society*, 48(3), 383–418. <https://doi.org/10.1007/s11186-019-09352-6>
- Eiland, J., Hammonds, C. M., Ponos, S. M., Weigand, S. M., & Scherer, W. T. (2021). Developing models to predict giving behavior of nonprofit donors. *2021 Systems and Information Engineering Design Symposium (SIEDS)*, 1–6. <https://doi.org/10.1109/SIEDS52267.2021.9483771>
- Evans, E. M. (2023). Animal advocacy and the “good cop-bad cop” radical flanking of laboratory research. *Sociological Inquiry*, 93(3), 662–686. <https://doi.org/10.1111/soin.12521>
- Fiennes, C. (2017). We need a science of philanthropy. *Nature*, 546(7657), Article 7657. <https://doi.org/10.1038/546187a>
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.8cd550d1>
- Future of Life Institute (2023). Pause giant AI experiments: An open letter. *Future of Life Institute*. <https://futureoflife.org/open-letter/pause-giant-ai-experiments>
- Gabriel, I. (2017). Effective altruism and its critics. *Journal of Applied Philosophy*, 34(4), 457–473. <https://doi.org/10.1111/japp.12176>
- Gore, A. (1993). *Creating a Government That Works Better & Costs Less: The Report of the National Performance Review*. The Review.
- Guo, Y., Dong, Y., Wei, X., & Dong, Y. (2023). Effects of continuous adoption of artificial intelligence technology on the behavior of holders' farmland quality protection: The role of social norms and green cognition. *Sustainability*, 15(14), Article 14. <https://doi.org/10.3390/su151410760>

- Habryka, O. (2023). Shutting down the Lightcone offices. *LessWrong*. <https://www.lesswrong.com/posts/psYNRb3JCncQBjd4v/shutting-down-the-lightcone-offices>
- Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks* (arXiv: 2306.12001). arXiv. <https://doi.org/10.48550/arXiv.2306.12001>
- Henriksen, A., & Blond, L. (2023). Executive-centered AI? Designing predictive systems for the public sector. *Social Studies of Science*, 03063127231163756. <https://doi.org/10.1177/03063127231163756>
- Hernandez-Orallo, J., Martinez-Plumed, F., Avin, S., & Whittlestone, J. (2020). AI paradigms and AI safety: Mapping artefacts and techniques to safety issues. In *European Conference on Artificial Intelligence* (2020). Santiago de Compostela, Spain. [https://ecai2020.eu/papers/1364\\_paper.pdf](https://ecai2020.eu/papers/1364_paper.pdf)
- Herzog, P. S., Naik, H. R., & Khan, H. A. (2021). *AIMS Philanthropy Project: Studying AI, Machine Learning & Data Science Technology for Good*. Indiana University Lilly Family School of Philanthropy and Indiana University School of Informatics and Computing, IUPUI, Indianapolis. <https://hdl.handle.net/1805/25177>
- Holzmeier, C. (2021). Beyond 'AI for Social Good' (AI4SG): Social transformations-not tech-fixes-for health equity. In *Interdisciplinary Science Reviews* (Vol. 46, Issues 1–2, SI, pp. 94–125). Routledge Journals, Taylor & Francis. <https://doi.org/10.1080/03080188.2020.1840221>
- Howard, A., Borenstein, J., & Gosha, K. (2019). *NSF-Funded Fairness, Ethics, Accountability, and Transparency (FEAT) Workshop Report* (10139705). Georgia Institute of Technology. <https://par.nsf.gov/biblio/10139705>
- Isabelle, D. A., & Westerlund, M. (2022). A review and categorization of artificial intelligence-based opportunities in wildlife, ocean and land conservation. *Sustainability*, 14(4), Article 4. <https://doi.org/10.3390/su14041979>
- Isbanner, S., O'Shaughnessy, P., Steel, D., Wilcock, S., & Carter, S. (2022). The adoption of artificial intelligence in health care and social services in Australia: Findings from a methodologically innovative national survey of values and attitudes (the AVA-AI study). In *Journal of Medical Internet Research* (Vol. 24, Issue 8). JMIR Publications, Inc. <https://doi.org/10.2196/37611>
- Johnson, P. D. (2018). *Global Philanthropy Report: Perspectives on the Global Foundation Sector*. <https://policycommons.net/artifacts/1847356/global-philanthropy-report/2593720/>
- Kaplan, S. A., & Garrett, K. E. (2005). The use of logic models by community-based initiatives. *Evaluation and Program Planning*, 28(2), 167–172. <https://doi.org/10.1016/j.evalprogplan.2004.09.002>
- Karnofsky, H. (2012). Thoughts on the Singularity Institute (SI). *LessWrong*. <https://www.lesswrong.com/posts/6SGqkCgHuNr7d4yJm/thoughts-on-the-singularity-institute-si>
- Karnofsky, H. (2022). How might we align transformative AI if it's developed very soon? *AI Alignment Forum*. <https://www.alignmentforum.org/posts/rCJQAkPTEypGjSJ8X/how-might-we-align-transformative-ai-if-it-s-developed-very>
- Key, J. (2001). Enhancing fundraising success with custom data modelling. *International Journal of Non-profit and Voluntary Sector Marketing*, 6(4), 335–346. <https://doi.org/10.1002/nvsm.159>
- Kim, W. (2022, November 15). Sam Bankman-Fried's arrest is the culmination of an epic flameout. *Vox*. <https://www.vox.com/the-goods/23458837/sam-bankman-fried-ftx-sbf-downfall-explained>
- Kosoff, M. (2015). Elon Musk donates \$10 million to the future of life institute. *Business Insider*. <https://www.businessinsider.com/elon-musk-donates-10-million-to-the-future-of-life-institute-2015-1>
- Kosoy, V. (2020). Infra-Bayesianism. *AI Alignment Forum*. <https://www.alignmentforum.org/s/CmrW8fCmSLK7E25sa>
- Lange, J. I. (1990). Refusal to compromise: The case of earth first! *Western Journal of Speech Communication*, 54(4), 473–494. <https://doi.org/10.1080/10570319009374356>
- Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science*, 381(6654), 138–138. <https://doi.org/10.1126/science.adi8982>
- Maas, M. (2023). International AI institutions: A literature review of models, examples, and proposals. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4579773>
- Madeo, E. (2022). Fundraising in the higher education context: A topical and theoretical literature review. *Education Society and Human Studies*, 3, 1–22. <https://doi.org/10.22158/eshs.v3n2p1>
- Markoff, J. (2015). Artificial-intelligence research center is founded by Silicon Valley investors. *The New York Times*. <https://www.nytimes.com/2015/12/12/science/artificial-intelligence-research-center-is-founded-by-silicon-valley-investors.html>
- Mathews, D. (2023, July 17). The \$1 billion gamble to ensure AI doesn't destroy humanity. *Vox*. <https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2>

- Mazzi, F., Taddeo, M., & Floridi, L. (2023). AI in support of the SDGs: Six recurring challenges and related opportunities identified through use cases. In F. Mazzi & L. Floridi (Eds.), *The Ethics of Artificial Intelligence for the Sustainable Development Goals* (pp. 9–33). Springer International Publishing. [https://doi.org/10.1007/978-3-031-21147-8\\_2](https://doi.org/10.1007/978-3-031-21147-8_2)
- McCright, A. M., & Dunlap, R. E. (2000). Challenging global warming as a social problem: An analysis of the conservative movement's counter-claims. *Social Problems*, 47(4), 499–522. <https://doi.org/10.2307/3097132>
- Meaker, M. (2023). Meet pause AI, the protest group campaigning against human extinction. *Wired UK*. <https://www.wired.co.uk/article/pause-ai-existential-risk>
- Metz, C. (2022). *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. Penguin Publishing Group.
- Metz, C. (2023, May 1). 'The godfather of A.I.' leaves Google and warns of danger ahead. *The New York Times*. <https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>
- Mickle, T. (2023). Big Tech rebounds and preps for transformative A.I. investments. *The New York Times*. <https://www.nytimes.com/2023/08/05/technology/tech-nvidia-chips.html>
- Mittal, P., & Srivastava, V. K. (2021). A review of supervised machine learning algorithms to classify donors for charity. *International Journal of Advanced Research in Computer Science*. <https://ijarcs.info/index.php/Ijarcs/article/view/6685/5388>
- Miyazaki, K., Murayama, T., Uchiba, T., An, J., & Kwak, H. (2023). Public perception of generative AI on Twitter: An empirical study based on occupation and usage (arXiv: 2305.09537). *arXiv*. <http://arxiv.org/abs/2305.09537>
- Morley, J., Machado, C. C. V., Burr, C., Cows, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, 260, 113172. <https://doi.org/10.1016/j.socscimed.2020.113172>
- Moskovitz, D. (2023). What happened to the OpenPhil OpenAI board seat? *EA Forum*. <https://forum.effectivealtruism.org/posts/CmZhcEpz7zBTGhksf/what-happened-to-the-openphil-openai-board-seat>
- Nast, C. (2015, November 16). The doomsday invention. *The New Yorker*. <https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom>
- Ngo, R., Chan, L., & Mindermann, S. (2023). The alignment problem from a deep learning perspective. <https://doi.org/10.2307/3381012>
- Noordt, C., & Tangi, L. (2023). The dynamics of AI capability and its influence on public value creation of AI within public administration. *Government Information Quarterly*, 101860. <https://doi.org/10.1016/j.giq.2023.101860>
- Ó hEigeartaigh, S. (2017). *The State of Research in Existential Risk* (SSRN Scholarly Paper 3446663). <https://papers.ssrn.com/abstract=3446663>
- OpenAI. (2018, April 9). OpenAI charter. *OpenAI*. <https://openai.com/charter>
- Open Philanthropy (2023a). New roles on our global catastrophic risks team. *Open Philanthropy*. <https://www.openphilanthropy.org/research/new-roles-on-our-gcr-team/#4-ai-governance-and-policy-aigp>
- Open Philanthropy. (2023b). Potential risks from advanced artificial intelligence. *Open Philanthropy*. <https://www.openphilanthropy.org/focus/potential-risks-advanced-ai/>
- Osborne, D. (1993). Reinventing government. *Public Productivity & Management Review*, 16(4), 349–356. <https://doi.org/10.2307/3381012>
- Pawson, R. (2002). Evidence-based policy: In search of a method. *Evaluation*, 8(2), 157–181. <https://doi.org/10.1177/1358902002008002512>
- Perry, B., & Uuk, R. (2019). AI governance and the policymaking process: Key considerations for reducing AI risk. *Big Data and Cognitive Computing*, 3(2), Article 2. <https://doi.org/10.3390/bdcc3020026>
- Piper, K. (2023). Can society adjust at the speed of artificial intelligence? *Vox*. <https://www.vox.com/future-perfect/2023/3/18/23645013/openai-gpt4-holden-karnofsky-artificial-intelligence-ai-safety-existential-risk>
- PricewaterhouseCoopers (2017). What doctor? Why AI and robotics will define New Health. *PricewaterhouseCoopers*. <https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/ai-robotics-new-health.pdf>
- Prunkl, C., & Whittlestone, J. (2020). Beyond near- and long-term: Towards a clearer account of research priorities in AI ethics and society. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 138–143. <https://doi.org/10.1145/3375627.3375803>

- Ramirez, A., & Saraoglu, H. (2009). *An Analytic Approach to Selecting a Nonprofit* (SSRN Scholarly Paper 1488870). <https://doi.org/10.2139/ssrn.1488870>
- Roose, K. (2023a). AI poses 'risk of extinction,' industry leaders warn. *The New York Times*. <https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html>
- Roose, K. (2023b). Why a conversation with Bing's Chatbot left me deeply unsettled. *The New York Times*. <https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html>
- Sætra, H. S., & Danaher, J. (2023). Resolving the battle of short- vs. long-term AI risks. *AI and Ethics*. <https://doi.org/10.1007/s43681-023-00336-y>
- Schiff, D., Ayesh, A., Musikanski, L., & Havens, J. C. (2020). IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2746–2753. <https://doi.org/10.1109/SMC42975.2020.9283454>
- Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the principles to practices gap in AI. *IEEE Technology and Society Magazine*, 40(2), 81–94. <https://doi.org/10.1109/MTS.2021.3056286>
- Schiff, D. S. (2023). Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy. *Review of Policy Research*, 40(5), 729–756. <https://doi.org/10.1111/ropr.12535>
- Schiff, D. S., Laas, K., Biddle, J. B., & Borenstein, J. (2022). Global AI ethics documents: What they reveal about motivations, practices, and policies. In K. Laas, M. Davis, & E. Hildt (Eds.), *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields* (pp. 121–143). Springer International Publishing. [https://doi.org/10.1007/978-3-030-86201-5\\_7](https://doi.org/10.1007/978-3-030-86201-5_7)
- Schwartz, D., Selman, J. M. G., Wrege, P., & Paepcke, A. (2021). Deployment of embedded edge-AI for wildlife monitoring in remote regions. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1035–1042. <https://doi.org/10.1109/ICMLA52953.2021.00170>
- Shapiro, D., & Cody, S. (2015). Data quality to further philanthropy's mission. *Mathematica Policy Research Reports*. <https://www.mathematica.org/publications/data-quality-to-further-philanthropys-mission>
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm (arXiv: 1712.01815). *arXiv*. <https://doi.org/10.48550/arXiv.1712.01815>
- Simpson, B., Willer, R., & Feinberg, M. (2022). Radical flanks of social movements can increase support for moderate factions. *PNAS Nexus*, 1(3), 110. <https://doi.org/10.1093/pnasnexus/pgac110>
- Singer, P. (1975). *Animal Liberation: A New Ethics for Our Treatment of Animals*. Eweb: 10461. <https://repository.library.georgetown.edu/handle/10822/769929>
- Singer, P. (1998). *Ethics into Action: Henry Spira and the Animal Rights Movement*. Rowman & Littlefield.
- Singer, P. (2019). *The Life You Can Save: How to Do Your Part to End World Poverty* (10th Anniversary ed. edition). [www.thelifeyoucansave.org](http://www.thelifeyoucansave.org).
- Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). AI for good and the SDGs. In B. C. Stahl, D. Schroeder, & R. Rodrigues (Eds.), *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges* (pp. 95–106). Springer International Publishing. [https://doi.org/10.1007/978-3-031-17040-9\\_8](https://doi.org/10.1007/978-3-031-17040-9_8)
- Stern, K. (2013). *With Charity For All: Why Charities Are Failing and a Better Way to Give*. Knopf Doubleday Publishing Group.
- Stix, C., & Maas, M. M. (2021). *Bridging the Gap: The Case for an 'Incompletely Theorized Agreement' on AI Policy by Charlotte Stix, Matthijs M. Maas*: SSRN. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3756437](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3756437)
- Strickland, E. (2022). Timnit Gebru is building a slow AI movement. *IEEE Spectrum*. <https://spectrum.ieee.org/timnit-gebru-dair-ai-ethics>
- Sulaeman, D. (2018). Smart charities? Analyses of IT-enabled charitable fundraising. *PACIS 2018 Proceedings*, 340.
- Taillandier, A. (2021). “Staring into the singularity” and other posthuman tales: Transhumanist stories of future change. *History and Theory*, 60(2), 215–233. <https://doi.org/10.1111/hith.12203>
- Van Dyke, N., & Amos, B. (2017). Social movement coalitions: Formation, longevity, and success. *Sociology Compass*, 11(7), e12489. <https://doi.org/10.1111/soc4.12489>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. <https://doi.org/10.48550/arXiv.1706.03762>
- Vequist, D. (2014). *Nonprofit Fundraising Transformation Through Analytics* (pp. 116–125). <https://doi.org/10.4018/978-1-4666-7272-7.ch009>

- Voida, A. (2014). A case for philanthropic informatics. In S. Saeed (Ed.), *User-Centric Technology Design for Nonprofit and Civic Engagements* (pp. 3–13). Springer International Publishing. [https://doi.org/10.1007/978-3-319-05963-1\\_1](https://doi.org/10.1007/978-3-319-05963-1_1)
- Wakunuma, K., Ogoh, G., Eke, D., & Akintoye, S. (2022, January 1). Responsible AI, SDGs, and AI Governance in Africa. *2022 IST-Africa Conference (IST-Africa)*, 1–13. <https://doi.org/10.23919/IST-Africa56635.2022.9845598>.
- White, H. (2019). The twenty-first century experimenting society: The four waves of the evidence revolution. *Palgrave Communications*, 5(1), Article 1. <https://doi.org/10.1057/s41599-019-0253-6>
- Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration*, 42(7), 596–615. <https://doi.org/10.1080/01900692.2018.1498103>
- Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics*. Machine Intelligence Research Institute.
- Yudkowsky, E. (2016). *The AI Alignment Problem: Why It's Hard, and Where to Start*. Machine Intelligence Research Institute.
- Yudkowsky, E. (2023). The only way to deal with the threat from AI? Shut it down. *Time*. <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>
- Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3), 101577. <https://doi.org/10.1016/j.giq.2021.101577>





# Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

## **PART IV**

# Ethics, AI, and philanthropy



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# HOW CAN PHILANTHROPY PROMOTE ETHICAL, INCLUSIVE, AND RESPONSIBLE AI DEVELOPMENT?

Lessons from impactIA Foundation

*Laura Tocmacov*

## **1 Artificial intelligence, from tool to non-human entity**

The emergence of AI represents an unprecedented turning point in the history of technology and humanity. This revolution goes far beyond mere technology: it confronts us with a new era in which the boundaries between species and those of intelligence are redefined. As the catalyst of this transformation, AI stands out for its uniqueness and far-reaching implications, marking the beginning of a unique interspecies collaboration, no longer as a mere tool, as we often like to say, but as a non-human entity with which we develop relationships. This is what “AI assistants” bring us: a new kind of relationship, where, no matter what we do, we forget that it is “just an AI,” despite all our efforts to de-anthropomorphize it. This partnership, bringing together human, collective, and artificial intelligence, generates unexplored potential for overcoming the limits inherent in each form of intelligence.

In this quest for intellectual expansion, AI confronts us with our mirror, revealing our genius and shortcomings. By seeking to reproduce our intelligence, it raises crucial questions about our essence and representativeness. The biases intrinsic to AI systems, stemming from their design by a small, privileged portion of the world’s population, shine a blinding light on the inequalities and distortions of social representation. Indeed, training data forms the basis of AI learning. If this training is done on unrepresentative data, for example, only on men to predict the risk of heart attack, the AI excludes from its training the possibility of predicting the risk of heart attack in women because the symptoms are different. This concrete example brings us to the heart of the research, which is currently carried out mainly on white men. The practical consequence is that female heart attack victims are attended to 40 minutes later because their symptoms are different and they are not correctly spotted (Lichtman et al., 2015).

The growing accessibility of AI, supported by the democratization of smartphones, gives this technology power and reach, with 67% of the total population owning a smartphone by 2020. However, this ubiquity also leads to a significant gap and glaring imbalance in the distribution of AI-generated wealth between AI users and developers. AI needs data to be trained, which is provided by

users without real consent, as a recent case between The New York Times and Open AI for copyright infringement has just demonstrated (Le HuffPost, 2023). The latter trained its ChatGPT model by copying and using millions of copyrighted Times articles without permission. Stability AI and Midjourney are also in the spotlight after artists launched a class action against them for copyright infringement (Vincent, 2023). Here, too, the models were trained without the artists' permission. Once AI has been trained and put into product form, it enriches the companies involved. It leads to a concentration of power within a few dominant companies, with no distribution of the wealth generated. This phenomenon poses a major challenge in terms of equity and governance.

First, it is important to realize that AI acts as an amplifier. In this sense, it amplifies preexisting, unresolved societal problems, not creating new ones. Likewise, we need to recognize that while AI raises questions at a global level, it is at a local level that solutions need to be found because each community has a different reality and its uniqueness that is enriched by AI. Global answers impoverish the richness of diversity.

These essential features remind us that the introduction of this disruptive technology exacerbates unresolved social problems. It is disruptive because of the general disorganization it entails and the radical changes in the economic world we are just beginning to touch. The intense pressure exerted by AI to adapt will affect the different strata of society unevenly. While in 2022, an EPFL study indicated that 62% of professions were threatened by automation, in March 2023, this figure had risen to 67% (Nosengo, 2022). The challenge of training to keep pace with this evolution is a major one. An overview of ILO<sup>1</sup> shows that women often have less access than men to productive resources, education, skills development, and the labor market (ILO, n.d.).

It is complex to identify the impact that AI will have on governance and the way we work. However, an OCDE (2023) publication points to the threats AI may pose. Although, for the time being, the main focus is on the positive modification of thankless, repetitive, or even dangerous tasks, the first negative signs are emerging: activity becomes more sustained after the introduction of AI systems, with less frequent human interaction and increased data collection on workers. Once again, a polarized disparity is emerging, with graduates and managers on the one hand more inclined to see AI as beneficial, and workers subject to algorithmic management or working with AI on the other, who find it less favorable.

These elements will need to be closely monitored to adjust and reduce inequalities. Our lack of experience, with no other reference than an Industrial Revolution, presents huge differences, although there is the similarity of the transition from tasks usually carried out by humans to machines, except that here it is not the physical body that is replaced, which is easy to observe in its tasks, but thought, which is more complex. This puts us in a "blind" position, with traditional references proving insufficient. Indeed, although the invention of the Internet, digitization, and electricity have certain similarities, by bringing about structural changes in the economy, they have done so gradually, allowing a certain amount of adaptation over several decades and without creativity or autonomy.

Conversely, AI presents itself as a potential collaborator, which will impact the world of work at an extraordinary speed and scale. Its ability to process massive quantities of data to make complex decisions that surpass human analytical capacity, its interconnectivity with other digital technologies, or its capacity to learn, create, and act autonomously raise questions on several levels. AI takes us into an incursion of hitherto human prerogatives, the evolution of which can be unpredictable.

This inordinate situation forces us to think creatively about our approaches to the transition to AI and to consider the impacts, solutions, potential, and dangers of AI differently. The AI era calls for an iterative approach, exploring new forms of intelligence and decision-making. If we want

sustainability in the global model of society, in the sense of being sustainable and benefiting all, civil society must become a key player in deciding what is “desirable” and what is “acceptable” in the use of AI. It must be empowered to become an alter ego alongside companies and governments. This new era demands unique collaboration and creativity. Otherwise, we are accelerating our current model, which is pushing us into the wall both economically and ecologically.

Philanthropy can play a predominant role here, supporting development designed for the common good and not simply for profit. It can strike a balance between profit and sustainability. Philanthropy can influence new paths by encouraging the empowerment of individuals and civil society to play their part in a vision of the common good and of an ethical and inclusive revolution through AI.

In this sense, we are entering a new era. For the first time in human history, we have a real choice that allows us to imagine a world of equitably distributed abundance. A world where work is transformed by hybrid intelligence, nurturing individual fulfillment, a harmonious life, and an equitable society. At this turning point philanthropy will be able to “do its bit.” At all levels, whatever its resources, philanthropy can provide the drops of water needed to help improve the situation, inspire others by its example, and link communities together.

In this context, **impactIA** was founded with a vision of a world of work evolving into an era of sustainability and fairness by aligning artificial intelligence with humanist principles and fundamental rights. We aspire to a future where everyone benefits from hybrid intelligence,<sup>2</sup> not only to overcome the challenges of our time but also to build a fairer, more inclusive world that respects humanity and our planet.

**Our mission** is to forge links between individuals, organizations, and civil society, creating synergies and solutions to meet the challenges and seize the opportunities of this technological revolution. We promote the transformation of organizations toward hybrid intelligence, where humans and AI collaborate for the common good. We are working to democratize AI based on three pillars: **(1) the individual**, **(2) organizations**, and **(3) civil society and governments**. To achieve this, our activities focus in particular on:

- impactIA Academy,<sup>3</sup> with a strategic watch on the new professions linked to AI, the design of training systems and their deployment;
- Educating young people aged 5–25 about AI, robotics, and ethics, in particular through the MAIA Academy;<sup>4</sup>
- Educating the population about AI and its challenges and opportunities, in particular through the AiiA festival;<sup>5</sup>
- The robot·me project, which promotes collective responsibility for AI. A multimodal AI assistant that learns from people, works for them, and pays them when they agree to share their cognitive capital for the overall improvement of robot·me for the community.

## **2 Principles to guide responsible AI development**

With the exponential growth of AI’s impact on society, a crucial question arises: what is responsible AI? The terms “ethical AI,” “responsible AI,” and “AI aligned with human values” are often used interchangeably, conveying a kind of common concern behind their difference, which is broadly to ensure that the development and application of AI takes place in accordance with ethical and moral principles, aligned with the well-being and values of humanity.

This chapter presents principles to develop responsible AI, based on the **Montreal Declaration (2018) for the responsible development of Artificial Intelligence**, which resulted from an inclusive deliberative process involving dialogue between citizens, experts, public officials, industry stakeholders, civil society organizations, and professional bodies.

Responsible AI is characterized by its ability to make decisions and act in ways that respect and promote human rights, equity, justice, and overall well-being. This approach involves a careful examination of the implications of AI on different aspects of life, such as privacy, safety, transparency of decision-making processes, and the impact on social and economic disparities. *The stakes are high: how can we shape AI so that it serves the interests of all without becoming a source of injustice or prejudice?* The answer to this question lies not in the technology itself but in how we design, regulate, and integrate it into our social fabric.

The Declaration provides a framework for reflection and action, proposing guiding principles to steer the development of AI toward a future that respects and enriches humanity as a whole. It underlines the importance of an inclusive and participatory approach, involving scientists, decision-makers, businesses, and civil society in a constructive and ongoing dialogue. It also addresses fundamental themes such as respect for individual autonomy, protection of privacy, solidarity between human beings, democratic participation, equity, and inclusion of diversity. These principles form an ethical compass, guiding AI's development toward moral and socially beneficial goals while promoting collective responsibility in the face of the challenges and opportunities this technology represents.

Building on the Declaration, we can explore what responsible AI might mean and how it can be realized in practice. This offers a promising path to a future where AI is not just a tool for technological progress but also a lever for our society's ethical and social advancement. To help philanthropic organizations (POs) shape responsible AI and contribute to its ethical development, they can draw inspiration from the principles below.

## 2.1 Principle of well-being

Work has a multifaceted impact on general well-being. As a source of self-esteem, society and the individual often associate it with personal worth. Providing a salary that gives access to material resources enables us to meet needs that are primarily physiological. Finally, its impact on our mental and/or physical health influences our entire being.

With this in mind, directing some of philanthropy's resources toward transforming the world of work as we know it today into something more sustainable and ecological could have a major impact on a fairer society. Two studies, reported by RTS (2020) from PK Ruch and SWICA (n.d.) in 2020, show that absences from work due to mental health problems have risen by 70% in Switzerland since 2012. Of these absences, 60% are due to burnout. The impact of work on human dignity is real in more ways than one, and aiming to increase well-being through this lever seems a promising hypothesis.

AI can contribute to well-being by transforming the world of work and automating repetitive and tedious tasks, enabling workers to focus on more creative and rewarding activities. This could significantly reduce cases of burnout and increase individuals' self-esteem, contributing directly to their well-being. From mental health in the workplace to better job opportunities, it can help reduce inequalities. To support this, POs can focus on project funding that targets AIs geared toward improving human well-being through work. These projects can range from research to responsible innovation aimed at improving working conditions broadly.

**impactIA's Example Projects.** impactIA deploys its support for well-being by transforming the world of work toward greater sustainability and equity. Recognizing the multidimensional

impact of work on well-being, it focuses on responsible innovation in AI to automate repetitive tasks, reduce burnout, and improve working conditions. This enables individuals to devote themselves to more creative and rewarding activities, enhancing their self-esteem, mental and physical well-being. The foundation fosters synergy between stakeholders to bring these advances to fruition by allocating resources to projects aimed to tangibly improve human well-being through the prism of work. Current project examples include:

- “***My Mentor is a Woman***” is a program that enables male managers, entrepreneurs, and department heads to be mentored by women to increase diversity within their company. We truly believe that more diversity will lead to a healthier, more representative, and more secure working environment. To achieve this, we wanted to reach out to “enlightened men” who believe in the power of diversity and enable them to create a trusting duo with their mentor, who becomes a sparing partner to help them meet this challenge. If the workplace becomes safer and “genders” are reconciled, we are convinced that well-being at work will increase. As the challenge of diversity and inclusion is multifactorial, the work women mentors do with their mentees is diverse and depends on what the company is already doing. This can range from support in preparing sessions on the theme of inclusion, a critical look at the commitment processes and readjustments, listening to situations that have arisen to propose other alternatives, and creating a work charter with the team, for example. The program is based on where the manager is and the extra steps he or she can take to be more inclusive.
- Beyond its global vision, the ***robot-me*** project aims to provide the individual with an assistant that genuinely takes care of him or her. In companies, this function should be performed by human resources. The latter are required to maintain the company’s interest first and foremost. A personal assistant who unilaterally takes care of the individual by getting to know them, their rhythms of fatigue and performance, and the concepts of well-being will be able to help promote well-being by providing support and advice. For impactIA, behind this project lies the desire to use LLM-type AI in a completely different way both in philosophy and in the business model itself. The assistant is there not only to increase productivity but above all to relieve the person of tasks that are not necessarily very good and are not a source of satisfaction. The aim is to increase the quality of work and to return to meaningful work. It’s a direct bulwark against “bullshit jobs,” which, according to studies, directly affect around 37% of British workers. It’s highly likely that we’ll have roughly the same figures in Switzerland. This type of employment creates a new form of self-destruction through boredom, known as “boreout.”
- The “***AI, art and young people’s mental health***” workshop was launched in 2022 with a trans-disciplinary group of 12 experts, including two young people. This project launch was initiated from observing the deteriorating state of young people’s health and the 70% increase since 2012 in the burnout rate in Switzerland. These newcomers to the world of work are doing worse and worse in a world of work that is doing worse and worse. Using the opportunities offered by AI to find solutions to curb this mental health epidemic carries us forward. The next steps in this project will be to explore different solution universities.

## ***2.2 Societal inclusion with the help of AI***

POs can fund projects that focus on informed consent from users, enabling them to understand how their data is used by AI systems and offering them the opportunity to control this data. Consent is a crucial point in the ethical issues surrounding the use of AI, and it is even more so when these systems enter deeply into our private lives. Even if consent is given at the outset, it must be



free. Often, however, neither clarity as to what will happen to the data collected nor absolute freedom is the order of the day when AI is used. More so, the “relationship” we can develop with AIs by interacting with them can lead us to reveal sensitive data in the “conversation’s” flow without being aware of it. Supporting the development of new user consent models that accompany users throughout their lives with applications, on the other hand, would enable us to maintain heightened vigilance. As part of the *robot-me* project, by giving users the ability to capitalize on their data and remunerate them in return when they decide, with informed consent, to share their knowledge, we are making several things possible:

- Provide more diverse and representative training models. The models are “finetuned” somewhere, not only to the user, but also to the company, or even to a region. This is a first step toward AI that is more representative of who we are “locally.”
- The importance of an open-source project of this kind also means that we can be particularly vigilant about model biases and have the power to influence the debiasing of the model. The simple principle of open source has an immediate impact on diversity.
- Finally, by capitalizing on cognitive skills and paying for them, we are helping compensate for the decline in activity for certain types of workers who are seeing their tasks gradually replaced by AI. This “compensatory” remuneration can enable a section of society that would otherwise be rapidly ejected from the market or put on a precarious footing to invest in continuing training, for example. This form of “royalty” also contributes to a fairer redistribution of wealth.

Societal inclusion also means empowering individuals. AI projects must respect the autonomy and enable individuals to take better control of their lives and their environment because it is not enough to have diversity in a company or society; this diversity must be able to “live together,” and this is where AI can help. In this category of projects, AI must share a common objective: to enable people to manage their own affairs better, make informed decisions, and interact with their environment more effectively and independently. Supporting projects that accelerate or increase people’s access to autonomy must become a priority that philanthropy can support. There are a number of concrete examples of this, but they are not exhaustive:

- Assistive technologies for disabled people (intelligent prostheses, voice recognition to help them with their daily tasks and improve their independence);
- Assistive technologies for people with disabilities (intelligent prostheses, voice recognition to help them with daily tasks and improve their independence);
- Intelligent personal assistants to help people access fundamental rights and write letters in a language they are not fluent in;
- Adaptive educational tools that adapt to each user’s learning style and pace;
- Improved communication tools, such as translators, to help overcome language barriers;
- Citizen participation platforms to facilitate civic engagement. To take a concrete example, the city of Neuchâtel has set up a citizen’s platform to stimulate civic participation. This open-source platform allows people to engage in dialogue and imagine and co-create projects for the common good. This platform could give rise to AI projects geared toward the common good. Here too, it is important to give everyone a voice to ensure greater inclusion in the city. By imagining AI projects conducted according to these principles, we end up with a society that decides what is desirable AI and what is acceptable, which it gives in return to the owners of these AIs.

### **2.3 Principle of privacy and intimacy**

Privacy is a major challenge in the age of AI, and POs have a pivotal role in addressing it. There are a myriad of actions that POs can take:

- **Legislation:** help fund legal research initiatives to develop legislative proposals that better protect individual privacy in the face of AI technologies. More university or HES studies financed on this specific theme would make it possible to take more important steps more quickly. However, these studies are nothing without “sandboxes,” which are real experiments on a small scale that allow us to see the impact of such and such laws before setting them in stone and risking slowing down innovation. In this sense “do tanks,” such as the impactIA foundation is doing, notably with the *AiiA festival*, a laboratory for experimenting with the impact of AI on society through art and culture, should be given more support. A Do Tank leads the intellectual work in the form of collective intelligence that can be achieved by breaking down the silos of expertise and experimentation in the field. Do Tanks operate on the principles of openness, altruism, solidarity, and sustainability.
- **Education and Awareness:** create and promote educational programs informing the public about the privacy risks involved in using AI. This can include workshops on how to secure personal data or awareness campaigns on digital rights. By creating academies such as the *MAIA Academy* for 12–18-year-olds run by the impactIA foundation, which enables young people from the French-speaking world to meet in all their diversity, work together and understand AI, robotics, and the challenges of these technologies, the new generation is better equipped to face these challenges. By also developing continuing education for the general public, enabling them to grasp the impact of AI on their work, we are giving them the keys to take hold of the transformation of their skills. Here, too, the programs developed by impactIA are aimed at achieving this objective; whether it is a *one-day course* on “*Understanding generative AI and applying it to your work*” or “*Ready for AI*,” these courses are designed to develop skills and upskilling.
- **Ethical Audits and Certifications:** fund bodies responsible for carrying out ethical audits of companies developing AI to verify that these companies respect privacy. This could lead to creating a trust label for consumers, indicating that products and services meet high data protection standards. However, great vigilance is required in this area. Many labels are now being called into question because industrial interest groups created them. Here, philanthropic organizations can make a difference by combining citizen-based approaches that take a stance on what is acceptable and desirable. Giving individuals in society the opportunity to coordinate their thoughts, define what is acceptable, and discuss and share their fears can finally lead to value audits and certifications, not self-constructed by a company according to its interests but by a common interest. Achieving this requires, for example, the creation of citizens’ workshops run by neutral bodies. In this spirit, impactIA has initiated the *citizens’ workshops*, which will take place in Switzerland, initially in the French-speaking cantons and then throughout the country.
- **Promoting Open Source and Interoperability:** encourage the development of open-source AI solutions that enable greater transparency on how data is processed and promote interoperability, giving users greater control over their data. Today, two currents are facing each other, particularly with LLM models: the current of closed founding models, with companies such as Open AI behind it, which advocates that AI is a source of significant danger of the extinction of humanity if it is in everyone’s hands. They say that open source is dangerous and that closed models are better left alone. On the other side, there is a movement, led from the outset by Yann LeCun, Researcher and AI Director at Meta, which recognizes the dangers of AI but

is proposing open research and open-source models so that we can react as quickly as possible with a strong community committed to finding solutions. Companies' financial interest in fighting open source is obvious, and their financial strength is enormous. Foundations such as Open Philanthropy have invested the strata of the best universities with millions for research into the existential risk posed by AI. Behind this type of foundation are Silicon Valley millionaires who Proprietary AI systems. Faced with these enormous resources, open-source voices are sometimes stifled. Providing the means to demonstrate the security alternative through open source is a challenge today, as we find ourselves at a pivotal moment in the face of a powerful strike force.

- **Support for Legal Action:** help fund litigation against companies that violate privacy, which can not only provide recourse for victims but also deter companies from engaging in such practices. Recent examples from Switzerland, such as the Swiss Federal Railways (SBB), which wanted to introduce surveillance technology to track travelers and their shopping habits in 2023, show how quickly AI-powered technologies can veer between blurred legal and illegal boundaries. However, the lack of financial resources and the power to act often means that companies snub the law with little risk of significant penalties. Financial support that would enable legal action to be taken in cases where AI violates fundamental rights would make it possible to clean up the current Far West of this technology. Even though the EU Act points in the right direction, there is still plenty of room for discrimination until it is implemented.
- **Partnerships with Cybersecurity Experts:** collaborate with security experts to develop best-practice guides for companies that collect and use data as part of their AI services. Associations such as ICON-NGO<sup>6</sup> aim to provoke a conscious reaction by popularizing and decoding the information society for civil society. ICON carries out concrete actions and explorations for future generations in the field of new technologies, with a focus on cybersecurity and artificial intelligence. More widespread support for this type of initiative would not only help to reduce the number of cases of cyber-attacks through better acculturation, but it would also provide the codes for better appropriation of cyberspace.

These actions, if well coordinated, can help create an environment where privacy is respected and protected in the development and application of AI while enabling society to reap the benefits of this technology. In short, POs can significantly contribute to creating a balance between technological innovation and respect for people's fundamental rights.

#### 2.4 *Solidarity principle*

Investing in initiatives that use AI to solve pressing societal problems, such as health, education, or humanitarian aid, can strengthen *global solidarity* and *mutual aid*. Philanthropy should encourage projects that maintain or strengthen solidarity between people and generations. It should encourage collaborative work and the creation of mutually supportive communities. The identification of individual risks using AI, for example, with a positive view to prevention, must be treated with extreme vigilance and due regard for the principles of solidarity so that it does not backfire on the principle of solidarity.

Projects that promote mutual aid, knowledge sharing, and cooperation are vitally important today, even more so than in the past, because they offer viable alternatives to the concentration of individualization that can lead humanity down polarized and destructive paths. Any project that aims to harness data and learning and put them in the hands of communities, thereby enriching them, is a desirable project and a genuine alternative to inequity.

AI has the potential to process massive volumes of data to identify and solve complex problems in areas such as healthcare, education, and humanitarian aid. For example, in the healthcare sector, AI can help predict epidemics, improve diagnostic accuracy, and personalize treatment for populations that otherwise lack access to quality care. In education, it can personalize learning to the specific needs of each student, especially in underdeveloped regions where access to qualified education is limited. In humanitarian aid, AI can optimize the logistics of resource distribution in crisis zones, ensuring faster, more targeted aid.

In this context, POs should favor initiatives that encourage collaborative working and the creation of mutually supportive communities. For example, AI can be used to connect volunteers with local initiatives, create platforms for skills exchange, or even support participatory finance projects for social causes. The focus should be on projects that use AI to promote cooperation and knowledge sharing between individuals and communities.

Risk management approaches guided by principles of solidarity that ensure that the use of AI for risk identification is not to the detriment of the individual could be supported. In this way, the guarantee of support for vulnerable groups through identifying health risks in a non-discriminatory way can be given. By actively supporting AI technology with a principle of solidarity, POs ensure that technology is synonymous with social progress and not division.

## **2.5 Principle of democratic participation**

Philanthropy should encourage projects that involve citizens in determining what is *desirable* and *acceptable* in terms of AI. We have too little experience with this technology to understand its real impact on our democracy. We need to be able to establish certain principles and follow rules but also have a process of experimentation and deliberation between all the players involved.

Using AI to anticipate and identify community needs, going beyond “classic consultations” where the community is listened to but the decision ultimately rests with those organizing the consultation, can have limits. There is a risk of ending up in situations where consultation is an alibi. If, for example, a municipality or canton organizes a “citizens consultation,” it does not have to do anything with the results. Some feedback from this type of consultation ends with frustrated citizens expressing that they have only been involved in the process for form’s sake and have not been listened to. One meaningful way of mitigating this risk could be for civil society organizations to organize such consultations, inviting public authorities, citizens, and organizations on an equal footing. This takes these consultations to a place where the interest of the common good is more important. The roadmap at the end of the consultation shows the political authorities the paths to take and then asks for feedback. The risk of leaving resolutions “in the wardrobe” is reduced because a trusted third party is the guarantor of what has been said.

An essential way of mitigating this risk could be for civil society organizations to organize such consultations, inviting public authorities, citizens, and organizations on an equal footing. This implies being able to finance this type of consultative project through philanthropy.

Moreover, when we talk about the “democratization of AI,” we usually think only of access to AI. However, the democratization of AI must contain a deeper meaning. A study of Montreal Ethics (Seger et al., 2023) on the democratization of AI highlights four points:

- 1 Democratizing the use of AI. The aim is to make it easier for everyone to use the technology.
- 2 Democratizing AI development. These are the people involved in AI design and development.

- 3 Democratizing AI's benefits. This involves the equitable redistribution of the value acquired by the organizations that control AI.
- 4 Democratizing AI governance. In reference to distributing the influence of decision-making to the widest possible stakeholder communities.

Philanthropy must support and encourage all four elements; otherwise, AI's representativeness and impact will remain notably biased.

**impactIA's Example Projects.** Based on the principle of democratic participation, the impactIA foundation has initiated "*citizen workshops*." The aim is to organize meetings with Swiss public administrations and citizens to answer the question, "*with AI, what is desirable and what is acceptable as a citizen?*" With this question, we want citizens to define what AI applications would be desirable in their lives as citizens, and in return, what they would be willing to give of themselves (especially in terms of data) to achieve these AI applications. By putting citizens at the heart of the process, the excesses of applications designed for the convenience of administrations, but which may run counter to the interests of individuals are minimized. By organizing spaces to define this, we want to create a dynamic where no AI is created, imagined, or deployed in public administrations or sensitive sectors such as healthcare, without including citizens or patients. An example given by the HUG (Hôpital Universitaire de Genève) patient partners group is very telling: *they say that of course they are aware that AI can help the hospital administratively, but that they would like priority, or at least equal value, to be given to projects that add value for patients, and not just to organizational efficiency.* We have already organized a first citizen workshop in October 2023 with the canton of Geneva and are now in the process of rolling it out more widely, first to the French-speaking part of Switzerland and then to the whole of Switzerland.

## 2.6 Principle of fairness

A major issue at the moment, particularly in the face of the Tescrealist<sup>7</sup> movement, is support for initiatives that tackle already existing problems, notably the discrimination and prejudice that emerge in AI systems and constitute a real challenge to achieving a fair and equitable society. These issues are current, immediate, and very short term. The media attention given to the Tescrealist movement means that current problems take a back seat. The enormous financial resources of the Tescrealist movement (in particular the millions raised by the Open Philanthropy Foundation and others supported by Silicon Valley milieus) stifle current problems, which do not reach this level of communication.

The databases used to learn AI are biased and unrepresentative. These biases can be overcome in a number of ways: by enriching databases with more diversity, using data augmentation techniques, involving communities in quality learning, etc. However, all these actions require financial resources to be carried out: AI projects that map the "over" and "under" representations of data, community involvement in the creation of AI, databases used to train AI, and so on. In the databases used to train AI, taking steps to debias the data and the learning would be possible. Concrete examples were initiated with the OpenAssistant project, an open and alternative "ChatGPT" led by Yannic Kilcher, a brilliant Swiss engineer who is very active in open source. For months, a community of over 18,000 people worldwide contributed to this assistant's creation. The future of AI development depends on the health and high quality of the dataset, and that is what they have been working on. Moreover, we had to stop because of a lack of resources.

## **2.7 Principle of diversity and inclusion**

Philanthropic funds and actions must be directed toward projects that reinforce social and cultural diversity, providing committed alternatives to social diversity and minorities, to combat a standardization of society that is representative of an influential current but not our humanity.

**impactIA's Example Projects.** Projects that support these principles, such as the “*My Mentor is a Woman*” project<sup>8</sup> from the impactIA foundation, which aims to increase diversity in AI and tech companies through the mentoring of male executives by female mentors, are a case in point. By becoming allies, different genders can ensure a fair, equitable, and sustainable place for all. Projects promoting access to training for minorities under-represented in the AI world are also part of this type of project. AI being, as we have seen, an amplifying force, correcting the representativeness of those who create AI should be a priority so as not to perpetuate current exclusions.

## **2.8 The precautionary principle**

Support for independent research and AI development that takes into account potential risks and impacts on safety and health in particular should be prioritized in order to anticipate harmful consequences. The need for action research should be an important priority. As AI technology is complex and disruptive, with few retrospective references, anticipation cannot be based on the known. Iterative and exploratory projects involving transdisciplinary groups with different perspectives on AI and its deployment would enable AI to be deployed in experimental environments and its effects to be contained until they are understood and mastered. Experimentation through art, as a field where freedom of expression and innovation are valued, is becoming necessary to test the limits of AI and reveal its profound implications in a way that is accessible and resonant for the public. Art, by its exploratory essence, allows technology to be pushed to its limits, providing a space where AI can be stretched beyond its conventional applications, often linked to commercial or practical goals.

Integrating AI into art projects gives it an experimental dimension that sheds light on otherwise obscure facets, such as emotional, societal, and ethical impacts. This offers audiences more intuitive and engaging points of understanding, transforming abstract concepts into tangible experiences. Art can thus act as a catalyst for collective reflection on the future direction of AI, highlighting issues that might not emerge in a purely technological or commercial context.

In addition, art offers a playground for ethical experimentation, where hypothetical AI scenarios can be played out in a controlled setting, raising public awareness of potential consequences without real risks. In this synergy between art and technology, the precautionary principle finds concrete expression, enabling society to engage with AI thoughtfully and critically, valuing its benefits while recognizing and preventing its risks.

**impactIA's Example Projects.** In this sense, the *AiiA*,<sup>9</sup> a laboratory for experimenting with the impact of AI through art and culture co-founded by the impactIA foundation and artist Jonathan O’Hear, responds to this principle. By putting Chimère, the multimodal AI entity, in the hands of international artists for a month-long residency every year since 2021 and enabling them to explore its potential and challenges through experimental contemporary art, AiiA acts as a place of anticipation. It raises the questions that AI brings us, its challenges, its limits but also all its potential. The residency culminates in ten days of conferences with multi-disciplinary thinkers, workshops, performances, and exhibitions that bear witness to his reflections and make them accessible to the general public.

As identified above, AI is not a conventional technology, so we have few points of comparison or anticipation. Based on history, we can only approach precautionary principles from a theoretical point of view. We need to adopt the “Do Tank” principle<sup>10</sup> to experience the impact of AI. Such initiatives are essential, as they enable “safe” implementation through art.

### ***2.9 The principle of responsibility***

Any project supported or deployed must contribute to reinforcing human responsibility in decision-making. We must also understand these new psychologies of interaction between “human” and “non-human” entities. In particular, we will have to deal with two human biases in relation to technology and find ways of balancing them.

The first is an overconfidence in technology, which leads us to believe it is better than we are. The rear sensors currently fitted to some cars can easily illustrate this bias. These sensors, activated when the car is reversed, emit beeps whose intensity increases the closer you get to an obstacle. Many examples of faulty detectors lead to collisions because of absolute confidence in their reliability.

Applied to AI, we run the risk of not carrying out our human evaluation before the final decision when AI pre-qualifies a decision for us, even if we only place it in a “pre-decision” position before human validation.

On the other hand, the absolute lack of confidence in anything from technology can lead some of us not to trust its contributions. This principle is beginning to be documented following their appearance in specific medical departments in particular, where nursing staff, aided by decision-making tools in the field that take the form of alerts and proposals for action, are not taken into account not because they are irrelevant but because they are not even considered and are systematically ignored without even reading them. Trust in these systems is non-existent, even to the point of boycotting their use.

These two aspects underline the need for broader reflection on how to support the principle of responsibility in these new interspecies relationships. If we are to understand and develop its foundations, philanthropy is a key player in this process.

### ***2.10 Principle of sustainable development***

The projects supported or the deployment of AI in philanthropy must guarantee strong ecological sustainability for the planet. Minimizing environmental impact throughout their lifecycle and judiciously weighing up the notions of “desirable” and “acceptable” need continually stimulated thinking.

We have seen here how philanthropy can play a decisive role in the development of responsible AI that benefits society as a whole. This technology can support alternatives to capitalist models, which must be able to coexist with other models to be built or reinforced. These alternatives can then become rethought societal and civic norms, solid because a space has been given to conceptualize and experiment with them. It contributes to a paradigm shift, participating in major technological advances within a vision of the common good aligned with fairer, more inclusive, and more sustainable human values. To this end, POs can:

- Fund research and development in areas where AI can have a positive social impact;
- Partner with academic institutions, NGOs, and corporations to create responsible AI research and application programs;

- Educate and raise awareness among the general public and decision-makers about AI's ethical and social challenges;
- Promote AI governance by supporting the development of policies and regulations that frame the ethical use of AI;
- Encourage responsible innovation by recognizing and rewarding companies and startups that integrate ethics into their AI development.

By adhering to these principles and acting accordingly, POs can play a decisive role in shaping a future where AI, guided by ethical values, contributes positively to society and respects everyone's rights and freedoms.

### ***2.11 impactIA, the open-source movement, and TESCREAL***

Overall, impactIA supports the open source and inclusivity movements to promote a future where AI enriches the lives of all, in alignment with our values of equity and sustainability. The foundation fosters an era of collaboration between individual, collective, and artificial intelligence, transcending the limits of each to build a more just and inclusive world. By supporting open source, we deeply believe that communities can take charge of the development and use of technologies that respect humanistic principles and fundamental rights, ensuring that the benefits of AI are shared equitably across society.

Open source increases the possibility of interfering with models to bring them closer to the objectives of the common good. It also makes models more transparent, or at least closer to it. Figures such as Yann LeCun, winner of the prestigious Turing Award<sup>11</sup> in 2018, support a trend advocating model sharing as a solution to AI ethics. This approach emphasizes the importance of openness, transparency, and broad access to AI technologies. It recognizes that open models carry risks and vulnerabilities but that it is by widening access and enabling public scrutiny that technology is made safer. It also considers that poor regulation can lead to concentrations of power that undermine competition and innovation. By accelerating understanding and enabling independent research, collaboration, and knowledge sharing, they see it as contributing to responsible AI. Strengthening public accountability is a key factor in this. For the moment, the question of who is to blame if an AI system does not work is still completely unclear and will become more apparent over the coming months/years. However, the risk with open-source systems reduces the risk of major errors because the remediation action is visible and immediately correctable, whereas closed systems are determined to hide errors.

This current contrasts with the TESCREAL<sup>12</sup> movements, which views AI through a prism of alarmist warnings and rigorous regulations that they, as experts in the field, propose to define for the good of humanity, with pronounced paternalism. In this movement, long-termism plays a crucial role. It proposes optimizing rationality and addressing ethical issues in a way that aims to maximize civilization's cumulative net happiness in the long term. This approach consciously neglects the immediate needs and suffering of individuals or the ecosystem in favor of a technologically advanced future, whose orientation leads us to download our brains into simulated environments, the source of ultimate happiness.

The implications of this ideology are far-reaching. One is redirecting philanthropic resources from today's immediate problems to support this futuristic vision. The financial resources of high-profile individuals relay on a massive scale the fear of a dystopian future where AI will destroy humanity if we do not control it... if THEY do not control it. They thus propose a frantic



race to technology as the answer to every problem, relegating current problems to mere “collateral damage.” This creates a major imbalance in terms of where philanthropic funds will put their resources and priorities.

In contrast, the open-source movement advocates supporting accountability, citizen acculturation, and smart regulation, lowering barriers to entry for new players focused on creating more safety and AI that increases the common good.

Both of these trends have an impact on AI philanthropy. In this chapter, we sought to show how philanthropy can play a crucial role in promoting an ethical and inclusive AI revolution by making room for long-term visions focused on a technological future and those that prioritize tangible improvements in the present in human and earthly quality of life. The key, as always, lies in the cohabitation of complementary approaches. In the same vein, philanthropy can take a long-term view while at the same time addressing humanity’s current needs and problems.

### **3 New directions for philanthropy in the AI era**

Philanthropy can be thought of only in its ability to fund and support social initiatives, but also as a major player in shaping a sustainable and equitable future. Therefore, it is becoming crucial to support the potential innovations that AI brings us by thinking outside the box. We need to look beyond financial profitability.

Indeed, the transfer of experience in bringing innovations to the economic world is extremely enriching, if we can also see their limits when we try to apply them to societal innovations in the sense of a positive impact on society as a whole and on the “common good.” With regard to the common good, i.e., the accepted reality of a positive general interest for the planet as a whole, innovation can take different and sometimes contrary paths. In this sense philanthropy should be able to rebalance these forces to leave more room for innovations geared toward the common good. Here are three ways in which philanthropy can strengthen its influence.

#### ***3.1 Promoting the acculturation of civil society***

If technological advances are to serve the general interest, individuals must be acculturated to AI, whatever their age. To achieve this, individuals must have a general understanding of the technology to make informed decisions. By empowering people to understand AI technologies and interact together, they can make informed decisions about how AI is or could be used in their daily lives and make choices.

Philanthropy should contribute and/or reinforce its interventions in favor of initiatives to educate, raise awareness, and inform all categories of the population, paying particular attention to social groups that are generally excluded. Workshops, awareness-raising programs, and training with multiple access levels are good sources of empowerment. Supporting initiatives to bridge the digital divide also helps to ensure that all sections of society have access to AI and the training they need to use it effectively. This includes funding access to technology in disadvantaged communities and digital education.

It is also fundamental to help the public put the cursor between what is desirable in terms of AI technology and what is acceptable. Some AI applications may seem desirable while raising questions of ethics, confidentiality, security, or energy consumption. Assessing these aspects and taking a stance on what is desirable and acceptable (re)empowers civil society to make informed decisions.

A well-informed and engaged “counter-power” is essential to balancing the interests of finance, sustainability, and equity. This force can advocate AI aligned with the common good, insisting on transparency, fairness, and respect for human rights. Just as AI can offer new opportunities to increase and sustain citizen participation and democracy by improving access to civic rights and deliberation, it must enable an equal and sustained voice in decisions that affect individuals.

By fostering acculturation, empowerment, and informed engagement, philanthropy can ensure that AI advances are directed toward collective well-being while reinforcing democratic principles and reducing inequalities. This is how we can ensure that AI is a tool at the service of humanity.

From the very beginnings of the impactIA foundation, we were confronted by the majority of people who excluded themselves from the AI debate by saying, “I am not a technologist; AI is a technologist’s debate!” It is because we are convinced that AI is a global and societal issue that we want to change the game. After two years of brainstorming and cross-disciplinary conferences with the “AI Expert Days,” which brought together AI engineers, philosophers, entrepreneurs, artists, social entrepreneurs, the Anicinabe medicine man, young activists, etc., we realized that the debate was still too narrowly focused.

With the Swiss artist Jonathan O’Hear, we approached this from another angle and imagined how we could reach a wider audience. The result was the AiiA festival, a laboratory for experimenting with the impact of AI on society through art and culture.

Since 2021, the *AiiA* has brought artists and transdisciplinary thinkers together in a Do Tank.

This type of initiative is a way of bringing the general public into the debate. Nevertheless, funding is a complex issue for this type of cross-disciplinary initiative on a subject as disruptive as AI. It is here, at the forefront of the debate, that philanthropic organizations have enormous added value: when the subject is still unpredictable, its effects unclear, its projections non-existent. It is also at this point, more than any other, that we need to bring civil society into the debate, into the “doing,” into the thinking.

### ***3.2 The challenges of hybrid intelligence***

The advent of hybrid intelligence marks a milestone in technological and societal evolution. This first-time convergence of individual, collective, and artificial intelligences, working together to overcome the limits of each, opens up new horizons in many fields. This collaboration goes beyond the notion of a traditional tool, making AI a partner in solving, creating, and deciding. For the first time, we are interacting with a non-biological entity in a truly collaborative way, changing our understanding of intelligence and creativity.

**impactIA’s Example Projects.** The *AiiA festival*<sup>13</sup> is an eloquent example of this dynamic, highlighting the challenges and opportunities of cross-species collaboration. Experiments conducted within the *AiiA* over the past three years have raised questions and sketched out possible answers. The first year saw the creation of the first opera, using hybrid intelligence, co-created by a Puerto Rican artist, Maria Sappho, and Chimère, the festival’s AI entity. This collaboration raised several questions about interspecies creativity in particular. After three editions, we can say that each time the 20 or so artists who came for a residency from different parts of the world said that, yes, with AI, they had opened up paths of creativity that they would not have taken without it.

The second year saw the emergence of a new orchestra made up of instruments “created” by Chimère and performed by the artists, as well as cross-species compositions, AI, Mushrooms, and Humans interacting. This opened the way for reflection on our way of being with other species, anthropomorphization, and its impact on us humans.

The last edition, meanwhile, raised questions about morality imposed by technologists and Silicon Valley in particular by questioning the “education” of these AIs. It is precisely because the public is confronted with the performances, works, and questions that emanate from this laboratory and can see the hybrid intelligence at work that the challenges become clearer.

While the questions raised by AI are global, each year, the AiiA emphasizes even more strongly how local the answers must be. This becomes more obvious as the problems of bias and cultural over-representation are brought into the public eye. These experiences also reveal the enriching potential of coexistence and joint creation between diverse groups and AI. While today’s AIs show us that the list of under-represented groups in our societies is long and diverse, the AiiA demonstrates how enriching it is for all to collaborate. As a society, we still have a long way to go in terms of non-hierarchical coexistence. Moreover, if art can raise questions, it can also open up new, unexplored avenues.

### ***3.3 Implications for governance***

Hybrid intelligence also challenges our approach to governance. Traditionally hierarchical, these new entities may enable us to rethink our relationships with others and offer innovative alternatives in our daily and professional lives. The composition of these three intelligences – individual, collective, and artificial – could open new doors of knowledge and opportunity. This new form of amplification encourages a redefinition of power and collaboration dynamics. However, it is essential here to underline precisely this possible amplification. AI “amplifies.” This does not mean it makes people smarter, but it amplifies what already exists. If what already exists is a discriminating factor, that is what will be amplified. So do not jump headlong into this hybrid intelligence and be sure of what you want to amplify. However, this synergy could act as a catalyst for creative and revolutionary solutions in various fields, from art and science to economics and politics. Hybrid intelligence represents a paradigm shift in the way we interact with technology. It invites us to rethink how we collaborate, create, and live together. By embracing this synergy between humans and AI, we open up a future rich in possibilities, where the barriers of cognition and creativity are transcended for a more inclusive and innovative world. Once again, this path has its pitfalls, and the role of philanthropy is clearly to enable these explorations to take place in safe spaces, such as the arts, so that the lessons learned can be safely transferred to everyday life.

## **4 Conclusion**

This exploration of the intersection between artificial intelligence and philanthropy makes it essential to adopt an optimistic and pragmatic perspective. AI offers revolutionary possibilities for human progress, but these advances are not without inherent complexities and challenges. The road to an AI-enriched future is paved with technical, ethical, and social issues that require careful attention.

First, it is essential to recognize that AI, while a powerful tool, is not neutral. Every AI system is the product of human decisions and carries with it its creators’ biases, perspectives, and sometimes mistakes. Philanthropists must, therefore, ensure that the projects they support do not perpetuate existing inequalities or promote an exclusive, elitist technological vision.

Further, philanthropy must navigate the complex landscape of AI regulation, tackling not just technological advances but also ethical and social implications, even anticipating them. How can we ensure that the benefits of AI are available to all, not just the privileged few? The initiatives must support and promote universal accessibility and genuine inclusion, particularly for marginalized or

disadvantaged communities. In addition, philanthropy is invited to support projects that not only innovate in AI but are also forward-looking and anticipatory.

It is also essential to recognize that philanthropy can sometimes unwittingly reinforce existing power structures. Therefore, philanthropists must adopt an approach of listening to and collaborating with communities and experts from diverse backgrounds to ensure that their actions truly benefit those they seek to help.

In conclusion, the role of philanthropy in the AI era is not only to support initiatives and innovations but also to free up space to conceptualize and experiment with viable duplicating alternatives.

We do not yet know what a society is where work no longer defines us. We need to be able to support these iterative reflections that explore what fulfilling societies can be, where the concentration of wealth is distributed more equitably. We need to explore real, not theoretical, alternatives to capitalist models, with the means to reproduce their successes. These models could even coexist harmoniously, opening up new avenues.

## Notes

- 1 International Labour Organization.
- 2 Hybrid Intelligence – collaboration between individual, collective, and artificial intelligence to overcome the limits of each.
- 3 <https://impactia.org/academy/>
- 4 <https://impactia.org/new-generation/>
- 5 <https://aiiafestival.org/aiia/>
- 6 <https://icon.ngo>
- 7 <https://www.youtube.com/watch?v=gqtmUHhaplo>
- 8 impactIA Foundation website: <https://impactia.org/formation-continue/>
- 9 AiiA Festival website: [www.aiiafestival.org](http://www.aiiafestival.org)
- 10 A circle of people with expertise in their field, working together to produce actions based on principles of openness, altruism, solidarity, and sustainability. This is all part of a social and solidarity-based approach.
- 11 The Turing Award is considered the Nobel Prize of computer science. In 2018, it was awarded to Yann LeCun, Professor of Computer Science at New York University and Scientific Director of Meta's AI activities, alongside Yoshua Bengio, Director of the Mila Institut, Canada, and the British Geoffrey Hinton.
- 12 TESCREAL is a word coined by Dr. Emile P. Torres (2023) in an academic article co-authored with Dr. Timnit Gebru. They evoke a series of ideologies that inhabit a current of thought in the creation of AGI (General Artificial Intelligence). It stands for Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism, and Longtermism.
- 13 <https://aiiafestival.org/aiia/>

## References

- ILO (n.d.). *Genre et emploi (Employment)*. ILO. Retrieved March 24, 2024, from <https://www.ilo.org/employment/areas/gender-and-employment/lang--fr/index.htm>
- Le HuffPost (2023, December 27). *Le New York Times accuse Chat GPT d'avoir aspiré ses articles (Et l'attaque en justice)*. Le HuffPost. [https://www.huffingtonpost.fr/international/article/le-new-york-times-accuse-chat-gpt-d-avoir-aspirer-ses-articles-et-l-attaque-en-justice\\_227586.html](https://www.huffingtonpost.fr/international/article/le-new-york-times-accuse-chat-gpt-d-avoir-aspirer-ses-articles-et-l-attaque-en-justice_227586.html)
- Lichtman, J. H., Leifheit-Limson, E. C., Watanabe, E., Allen, N. B., Garavalia, B., Garavalia, L. S., Spertus, J. A., Krumholz, H. M., & Curry, L. A. (2015). Symptom recognition and healthcare experiences of young women with acute myocardial infarction. *Circulation: Cardiovascular Quality and Outcomes*, 8(2\_suppl\_1). <https://doi.org/10.1161/CIRCOUTCOMES.114.001612>
- Montreal Declaration (2018). Montreal Declaration. *Déclaration de Montréal IA Responsable*. [https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM\\_Decl-IA-Resp\\_LA-Declaration-ENG\\_WEB\\_09-07-19.pdf](https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf)

- Nosengo, N. (2022). *Comment rivaliser avec les robots*. <https://actu.epfl.ch/news/comment-rivaliser-avec-les-robots-2/>
- OCDE (2023). *Perspectives de l'emploi de l'OCDE 2023: Intelligence artificielle et marché du travail*. OECD. <https://doi.org/10.1787/aae5dba0-fr>
- RTS (2020, January 12). *Les arrêts maladie pour burn-out ont explosé depuis 2012, rapporte la NZZ [infoSport]*. rts.ch. <https://www.rts.ch/info/suisse/11006478-les-arrets-maladie-pour-burnout-ont-explose-depuis-2012-rapporte-la-nzz.html>
- Seger, E., Ovadya, A., Garfinkel, B., Siddarth, D., & Dafoe, A. (2023). *Democratising AI: Multiple meanings, goals, and methods* (arXiv: 2303.12642). arXiv. <https://doi.org/10.48550/arXiv.2303.12642>
- SWICA (n.d.). *Nouvelle étude révèle les raisons des arrêts de travail pour raisons psychiques – SWICA*. Retrieved March 24, 2024, from <https://www.swica.ch/fr/a-propos-de-swica/medias/service-medias/communiqués-de-presse/2022/etude-prescriptions-arrets-travail-pour-raisons-psychiques>
- Torres, É. P. (2023, June 15). *Tescrealism: The acronym behind our wildest AI dreams and nightmares*. Truthdig. <https://www.truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares/>
- Vincent, J. (2023, January 16). *AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit*. The Verge. <https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart>

# GUIDED CHOICES

## The ethics of using algorithmic systems to shape philanthropic decision-making

*Rhodri Davies*

### 1 Introduction

Discussion of the potential impact of artificial intelligence (AI) in the field of philanthropy has, to date, largely focused on three areas:

- The potential for harnessing machine learning (ML) and other tools to deliver new social and environmental interventions and how civil society organizations (CSOs) and philanthropic funders could play a role in realizing this potential;
- How AI might affect CSOs and philanthropic funders themselves through new opportunities to transform internal processes or through altering the nature of the broader financial and regulatory systems within which they operate;
- The impact that AI may have on individuals and communities and what this might mean in terms of the need for new approaches or increased advocacy from funders and CSOs.

What has received far less attention is the question of how AI may affect the philanthropic choices made by individuals, about when, where, and how to give. This may reflect a broader lacuna in our knowledge since, as Susser notes, “for several years, scholars have (for good reason) been largely preoccupied with worries about the use of artificial intelligence and machine learning (AI/ML) tools to make decisions *about us*,” but that “only recently has significant attention turned to a potentially more alarming problem: the use of AI/ML to influence our *decision making*” (Susser, 2019). However, even if this is true across a broader set of domains, it is a particular problem for philanthropy, as individual choice plays a fundamental role in this context. Philanthropy’s inherently dual nature (Reich et al., 2016) means that we have to understand it at both a macro level – as a systemic mechanism for allocating and redistributing resources within society at a scale that positions it alongside both the state and the market – and a micro level, where it reflects the myriad choices of individuals to use their “voluntary action for the public good” (Payton & Moody 2008). Hence, the capacity of AI to affect how we make individual choices has profound implications for the future of philanthropy.

This chapter will look at these implications and how we might respond to them. It will consider the nature and role of choice in philanthropy, what we know about the capacity of AI to influence

and shape our individual choices, and what this suggests about the potential impact of AI on philanthropic choice and decision-making. The implications will be considered across three key groups of actors that have the potential to employ AI to shape the decisions of potential donors, namely:

- General-purpose search and recommendation services (i.e., search engines, conversational interfaces, text-based generative AI tools);
- Giving platforms (i.e., dedicated donation/lending or crowdfunding platforms and commercial platforms that facilitate giving, such as payment providers or social media platforms);
- Individual cause-based organizations (both formal and informal).

The chapter will highlight key ethical questions that emerge from considering the use of AI by each of these different actors and suggest changes in policy or practice that could help to address them.

## **2 Choice and philanthropy**

Choice plays a fundamental role in philanthropy since our ability as individuals to decide whether and how to give away private assets for public benefit is one of the defining characteristics of philanthropy when set against other redistribution methods. However, there may be a danger that in making this assumption, we beg an important question about the nature of philanthropy since one of the fundamental philosophical questions about philanthropy throughout the ages has been whether it should be understood as a choice of duty. Is giving something you are entirely free to choose whether to do or not to do or are there moral or societal obligations of some kind that compel you to give (Martin, 1994; Schneewind, 1996)? To take one example of how this debate has evolved and the influence it has exerted, the predominant view among medieval Catholic scholars was that God determined the unequal distribution of wealth and resources within society and that part of his plan was for redistribution to take place through charity and almsgiving; so the “haves” had a duty to give to support the “have-nots” (Roberts, 1996). The Enlightenment then saw a shift in thinking about the nature of property that caused views to diverge. For some, like Locke or Grotius, people had a “natural right” to ownership of property that they had amassed through their efforts, and it was up to them whether to choose to share it (Winfrey, 1981). For others, like Wollstonecraft and Kant, the unequal distribution of resources reflected the fact that society was unjust, and addressing this injustice demanded that those with wealth give back as a matter of duty rather than of choice. To the extent that this duty was a “perfect” one (i.e., a duty where it is clearly defined who the recipients are and what they are due), it was usually thought best discharged through taxation. However, many argued that philanthropy represents an additional “imperfect” duty, where the requirement to give is clear, but the exact nature of how the money must be given or to whom is not specified (Schneewind, 1996).

The extent to which philanthropic giving represents a choice or a duty remains a matter of debate even today. Those who prioritize individual liberty may feel that philanthropy must be seen entirely as a matter of choice (Nozick, 1974; Salmon, 2023). Others who place more emphasis on justice and the equitable distribution of resources will argue that we must acknowledge an element of duty when it comes to philanthropy, even if this impinges upon our individual freedom. For instance, the moral philosopher Peter Singer has argued that there is a moral duty on those with sufficient wealth to give some of it out to help people in extreme poverty (Singer, 1972, 2006). Cordelli (2016) goes further, arguing that “philanthropy should be understood foremost as a duty

of reparative justice” and that “affluent donors should, as a matter of moral duty, exercise no personal discretion when deciding how to give and to whom. Indeed, they should regard their donations as a way of returning to others what is rightfully theirs.” In this view, philanthropy becomes solely a matter of duty, and there is no room for choice.

A reasonable position between these two extremes is to allow that both duties and choices guide philanthropic giving. We pay taxes as part of an agreed social contract, but this does not necessarily fully discharge the duty we owe to others, so an element of duty may also apply to our giving. In part, this may be simply an imperfect *duty to give* in an unspecified way, but it may also be a more specific duty to give *in a particular way or to particular causes* (MacAskill, 2015; Singer, 2015). (A duty that would obviously only have moral rather than legal force, of course.) Most people would stop short of claiming that *all* our philanthropy is driven by duty, however. They would allow the possibility of some giving that is *supererogatory* (i.e., above and beyond that which is demanded by duty). This portion of giving, at least, would be governed solely by choice in terms of our decision to give in the first place and in terms of what we choose to give to (Gewirth, 1987).

According to most current views, choice remains a vital ingredient of philanthropy. For this reason, it is unsurprising that a substantial body of research aims to shed light on the key factors affecting the choices we make when giving. Several literature reviews have attempted to survey this body of research and synthesize findings across it (Bekkers & Wiepking, 2011; Allen, 2018; Saeri et al., 2023), although they note that this presents challenges because the research is found across a wide range of disparate disciplines – including economics, political science, anthropology, neurology, psychology, and marketing – reflecting the inherently cross-disciplinary nature of philanthropy as a field of study. Despite this variety, however, it is possible to identify common themes across research from different fields. Bekkers and Wiepking (2011), for instance, postulate eight mechanisms that drive giving based on their analysis of the available literature: (1) awareness of need; (2) solicitation; (3) costs and benefits; (4) altruism; (5) reputation; (6) psychological benefits; (7) values; (8) efficacy.

Within these broad categories, it is possible to identify a wide range of specific factors that may affect philanthropic decision-making. Some relate to macro-level considerations, such as religious and ethnic diversity among donors (Andreoni et al., 2016), social class and socioeconomic status (Piff et al., 2010), the impact of government funding (Andreoni & Payne, 2011), the presence of match funding (Karlan & List, 2007), the availability of “social information” on what other donors have given (Alpizar et al., 2008; Croson & Shang, 2008; Shang & Croson, 2009), or the effect of media coverage of disasters (Brown & Minty, 2008). Various studies have also attempted to identify relevant cultural factors by exploring children’s attitudes to giving and prosocial behavior, finding that children appear to have a natural inclination to act pro-socially even from a very young age (Zahn-Waxler et al., 1992; Warneken & Tomasello, 2006), that younger children are less prone to “free-riding” behavior than adults (Harbaugh & Krause, 2000), and that as children get older, they are more likely to give to someone when they believe that the other person might reciprocate their generosity (Sebastián-Enesco & Warneken, 2015).

Many micro-level factors affect decisions made about giving at an individual level. Some of these are relatively unsurprising, such as whether the person soliciting donations is a member of their peer group (Meer, 2011), whether the person soliciting makes an audible request or not (Andreoni et al., 2017), the level of information the donor has about the cause being fundraised for (Eckel et al., 2007), whether that information concerns an identifiable individual or is in the form of statistical evidence (Jenni & Loewenstein, 1997; Slovic, 2007), and what kind of information the donor is given about the performance of the organization asking for funds (Butera & Horn, 2020). Other factors are perhaps less obvious, such as the influence of our body chemistry:



a number of studies have found, for instance, that the release of oxytocin results in a greater willingness to give (Zak et al., 2007; Barraza et al., 2011), while others have found that the brain's dopamine reward center may respond similarly when we give to how it does when we receive other kinds of rewards (Moll et al., 2006). There are also more esoteric factors that studies have found may have a bearing on our willingness to give and the choices we make: for instance, being exposed to “awe-inspiring” content (Rudd et al., 2012), having a prior conversation that primes you with the notion of a God or Supreme Being (Shariff & Norenzayan, 2007), thinking about death (Jonas et al., 2002), listening to “chill-inducing music” (Fukui & Toyoshima, 2014), listening to music with “prosocial lyrics” (Greitemeyer, 2009), or being in the presence of images of eyes (Ekström, 2012; Sparks & Barclay, 2013; Fathi et al., 2014).

As more and more of us give online – either through dedicated nonprofit platforms or, increasingly, through commercial social media or payment platforms that have added giving functionality to their offering – the nature of philanthropic choice and decision-making is changing rapidly. For one thing, the range of options available to us is far more significant because we are no longer constrained by physical proximity: many platforms enable individuals to give to groups and causes worldwide without ever needing to contact potential recipients or fundraisers. We are also less constrained by a reliance on formal organizations, as a large volume of online giving goes to informal groups, grassroots social movements, or individuals (Bernholz, 2021). But perhaps the most profound impact of the shift toward giving online is that the digital environment offers new opportunities to design “choice architectures,” that is, how choices are presented to us so they are tailored and responsive to our personal preferences and can be highly targeted to produce desired outcomes. Knowledge of the individual – and often unconscious – factors that shape our giving decisions will be an important competitive advantage when harnessing such approaches' potential. Some have already questioned whether this raises ethical concerns and whether it is appropriate to exploit insights about our subconscious behavioral drives to “nudge” us toward specific actions in this way when it comes to philanthropy and other prosocial behavior (Schulz et al., 2017; Rühle et al., 2021). Such concerns are only likely to become more acute as the impact of AI is felt more widely, as we shall see.

### 3 Choice and AI

AI already significantly impacts how choices and decisions are made in a wide range of fields. As noted at the start of this chapter, there has been a particular emphasis on how algorithmic systems are used to make decisions *about* us, with a growing body of literature examining the opportunities this presents – in terms of bringing new capabilities and efficiencies across the public, private, and nonprofit sectors (Wirtz et al., 2019; Kanter & Fine, 2022) – as well as the challenges it brings, such as new risks of “machine bias” against marginalized groups (Wachter-Boettcher, 2017; Eubanks, 2018; Noble, 2018) and concerns about insufficient transparency and accountability (de Fine Licht & de Fine Licht, 2020; Loi & Spielkamp, 2021). In this chapter, however, our primary focus is not on how AI may be used to make decisions about us but how it might affect our decisions as individuals when it comes to philanthropic giving. To that end, we will consider three key areas:

- The use of algorithms to determine responses to requests for information and how this shapes our choices and decision-making;
- The capacity of AI to enable personalization and “hyper-nudging,” to drive particular actions and outcomes;
- The use of AI-generated content to prompt emotional responses and thus drive behavior.

### **3.1 AI and information provision**

Our online world experience is heavily shaped by algorithms (Schmidt, 2021). While it is theoretically possible to navigate the internet without using intermediaries, in reality, the vast majority of web traffic comes via search engines (BrightEdge, 2019). More recently, social media platforms and conversational interfaces (either text-based, such as OpenAI’s ChatGPT, or voice-based, such as Apple’s Siri or Amazon’s Alexa) have also evolved to fill a similar role in many users’ online experiences (Huang, 2022; Perez, 2022). In both cases, the information we are provided is determined algorithmically in one of two main ways. The first is the *reactive* provision of information in response to a request, where a range of suitable answers may be algorithmically determined and then presented to the user either in the form of a ranked list (as in the case of a traditional search algorithm, such as that used by Google) or in the form of a single answer or small handful of answers (as tends to be the case with conversational interfaces). The second way algorithms may be used is for the *proactive* provision of information, where the user has not made an explicit request, but instead, suggestions or recommendations for content they may be interested in are provided to them based on data about their known interests, preferences, or online behavior. Recommender algorithms, which typify this latter approach, are at the heart of social media platforms such as TikTok, X, and Facebook, as well as content platforms like Spotify, Netflix, and YouTube, all of which seek to provide a steady stream of enticing content for us to click through to next to keep us on the platform for as long as possible (Schrage, 2020; Roy & Dutta, 2022).

When trying to understand how algorithms can shape our choices across both of these contexts, there are at least three things we need to consider: *what* information is provided (i.e., what the algorithm determines as the “correct” or “best” answer to a query, or how it decides what to recommend to us); *how* the information is provided (i.e., whether it comes in the form of a single answer or a list, and whether we can refine the information presented to us); and finally *our perception* of the information (i.e., whether we are aware that it represents the result of an algorithmic process).

The most basic question we must ask is: what is the nature of the algorithm that has produced the information we are presented with (with the immediate corollary questions: “Who designed it?” and “What are their motivations?”)? Unfortunately, here we encounter an issue that will prove to cut across almost all our considerations, namely lack of transparency. In the case of many of the algorithms that shape our online experience – such as Google’s ranking algorithm, TikTok’s video recommender algorithm, or YouTube’s content algorithm – the question of how they work largely remains a mystery. This is partly because these algorithms are closely guarded proprietary secrets. Still, it is also because many of them are by design “black boxes,” whose inner workings are deliberately opaque and not always fully comprehensible even to those who created them (Pasquale, 2015; von Eschenbach, 2021). Scholars and campaigners have increasingly called for this situation to be remedied by introducing greater openness to the use of algorithms. Initially, this was focused on the idea that we need to make algorithmic systems more transparent, but more recently, the focus has shifted toward making algorithms “explainable,” since transparency by itself is no longer felt by many experts to be a useful goal, as for most users being able to see the inner workings of an algorithm doesn’t further their understanding of how it works (Mittelstadt et al., 2019; Rai, 2020).

How information is presented to us is also important. In particular, the emergence of conversational user interfaces (CUIs) as alternatives to traditional search engines to find information is potentially very significant (Liao et al., 2020). The design of CUIs may vary considerably, from basic chatbots that provide only limited responses through systems that can hold natural language conversations at a level that may be functionally indistinguishable from a human being; however, one feature that most of them have in common is that information is no longer presented in the

form of a complete ranked list that can be inspected visually, but in the form of an “answer” (or small set of answers) to a question we have posed (Zamani et al., 2023). In practical terms, this may make it harder to identify and compensate for any biases. There are already concerns that traditional search engines perpetuate biases of various kinds (Espín-Noboa et al., 2022; Maillé et al., 2022), but the challenge may be more significant when it comes to information delivered through CUIs; both because of how we interpret the status of the information and because it is no longer possible to compensate for bias simply by looking further down the list of pages of results, as we might do with a traditional search engine. For these reasons, among others, some have questioned whether CUIs should be considered viable alternatives to conventional search engines (Gurdeniz & Hosanagar, 2023).

There are also deeper issues here. For one thing, we tend to see language as a marker of intelligence (Mahowald et al., 2023), so the more that automated systems can respond to our requests for information in credible natural language, the more likely we are to accept what they tell us and the less likely we are to assess that information critically. Another challenge is that the conversational nature of interfaces may reduce the *visibility* of the underlying technology, that is, our awareness that a process of digital intermediation and algorithmic determination of information is taking place. As scholars have noted, the potential *invisibility* of technology is a significant factor when it comes to practical and ethical concerns about how AI may shape individual choices, both because we are rendered far more susceptible to manipulation through technology when we are not even aware that this technology is being used (Van den Eede, 2011; Susser, 2017), and because the threat to autonomy is significantly higher (André et al., 2018; Vaassen, 2022; Bartmann, 2023).

The increasing invisibility of algorithmic systems and our willingness to accept the information they present us with less critically has led some scholars to claim that the distinction highlighted earlier in this chapter – between reactive information provision and proactive recommendation – is no longer that meaningful. For instance, Zamani et al. “do not make a strong distinction between *search* and *recommendation* tasks” because they see them as “closely related tasks that are becoming more closely related as time passes” and argue that in many cases, the same task can often be characterized as either search or recommendation depending on the device or interface being used and the context in which the task is taking place (Zamani et al., 2023). This blurring of boundaries between searches for information that we undertake actively and recommendations that we receive passively clearly has implications for the degree of autonomy we retain as users (Bartmann, 2023). However, we also need to bear in mind that the increasing reliance on algorithmically determined recommendations is primarily driven by our demands as consumers for highly tailored and personalized experiences (Arora, 2021). Should we, therefore, view this less as an unprovoked assault on our autonomy and more as a Faustian pact that we have entered into, in which we sacrifice autonomy for convenience?

### 3.2 *Personalization, choice architecture, and hyper-nudging*

The information provided to us is an important determinant of our choices, but it is by no means the only one. A range of factors relating to how options are presented to us also significantly affect our decisions.

Thaler and Sunstein (2009) coined the term “choice architecture” to describe this range of factors (and “choice architect” to refer to anyone who has control over them) (2009). Advertisers and marketers have long known that our subconscious motivations and drivers can be manipulated to sell us things (Samuel, 2010). Still, since the publication of Thaler and Sunstein’s work, there has

also been significant interest in how the design of choice architectures can be used in the public and nonprofit sectors to “nudge” people toward desirable prosocial actions (Behavioural Insights Team, 2013; Schulz et al., 2017; Capraro et al., 2019). (A “nudge” is defined by Thaler and Sunstein as “any aspect of choice architecture that predictably alters people’s behavior without forbidding any options or significantly changing their economic incentives.”)

Many early experiments with choice architecture focused on using the design of physical spaces or interactions to influence people’s actions. However, it is in the digital world that choice architecture and nudging have really come into their own, as in this context, it becomes possible to tailor choice architectures to individuals in a way that is not feasible in the physical world (Thomas et al., 2013; Weinmann et al., 2016). AI has now added a significant further dimension, as the combination of big data and algorithmic processes makes it possible not only to personalize nudges but to adapt them iteratively in real time in response to user behavior. The legal scholar Karen Yeung has coined the term “hyper-nudge” for this phenomenon (Yeung, 2017) and argues that:

unlike the static Nudges popularized by Thaler and Sunstein (2009), such as placing the salad in front of the lasagne to encourage healthy eating, Big Data analytic nudges are extremely powerful and potent due to their networked, continuously updated, dynamic and pervasive nature (hence “hypernudge”).

There is an existing body of critical thought about the ethics of nudge approaches (Bovens, 2009; Hobbs, 2017; Ruehle et al., 2021), but hyper-nudges bring new challenges of their own. Allowing autonomous systems to act as choice architects may, for instance, raise significant issues regarding transparency, responsibility, and accountability (Mills & Sætra, 2022). There are also questions about whether the iterative and adaptive nature of hyper-nudges makes it harder for us to avoid them than traditional nudges, and whether they present a more significant threat to our autonomy (André et al., 2018; Mills, 2022).

### ***3.3 Generative AI and emotive content***

Generative AI is a term for a broad class of models or algorithms that produce new content – such as text, photos, illustrations, video, audio, or code – based on large sets of suitable training data. We are already seeing generative AI being used in efforts to shape our individual choices. This will likely become even more commonplace in the future since the ability to generate audio and visual content that seems genuine – while in fact, being carefully calibrated to convey specific messages or emotions – represents a hugely powerful new tool for inducing beliefs and influencing actions (Sætra, 2023). We will undoubtedly see an increase in malicious attempts to spread misinformation and disinformation to induce false beliefs; there have been widespread concerns, for instance, about the rise of “deepfakes” – synthetic media that have been digitally manipulated to replace one person’s likeness convincingly with another – and their potential deployment in the context of elections and other political events (Chesney & Citron, 2018; Diakopoulos & Johnson, 2021; Pashentsev, 2023). Not all applications of generative AI will be deliberately malicious, of course, but some scholars argue that even the well-intentioned use of technology such as deepfakes is potentially harmful because it undermines our notions of authenticity in the online environment and thereby exacerbates the problematic erosion of trust (Hancock & Bailenson, 2021; de Ruiter, 2021).

## 4 AI and philanthropic choices

To understand how general considerations about how AI could affect individual choices apply in the specific context of philanthropy, we will consider three key actors that play a role in philanthropic choice-making in the digital environment:

- General-purpose search and recommendation services;
- Giving platforms;
- Individual cause-based organizations.

We will consider each actor's motivations for wanting to shape our philanthropic choices, what this means in terms of their likely adoption of AI tools, and the potential ethical raised.

### 4.1 *General-purpose search and recommendation services*

Many websites, platforms, and interfaces providing information to users – either in the form of reactive responses to requests or proactive recommendations – are not explicitly designed with philanthropy or the nonprofit sector in mind. The paradigm example is traditional search engines, but this category also includes any social media, content platforms, or conversational user interfaces that allow users to search for general information or to receive recommendations that could influence choices about giving. The primary motivation, in this case, is satisfying user demand for information – whether that is explicit (e.g., a search request) or implicit (e.g., a perceived openness to receiving a recommendation) – to ensure that users remain on a platform or return to a service in future. It is also possible that the company operating the search or recommendation service has its own corporate social purpose, which may be an additional factor in the motivation to provide information that can shape philanthropic choices.

In the case of providing information in response to a user request, it is unlikely the provider will exert any influence over the initial decision to give since this has most likely already been taken. However, there is substantial potential for influencing *where* and *how* the gift is made. A user may be seeking objective information – for instance, on which nonprofit organizations operate in a particular geographical location – or that user may be looking for more subjective information – for instance, on the “most pressing need” or which nonprofit organizations are “most effective.” In both cases, it is important to interrogate the nature of the algorithms used to determine the information on offer and the goals and motivations of those designing and operating them. As highlighted earlier in this chapter, this is particularly important in those cases where responses to requests for information are presented in the form of a single answer or small selection of answers rather than a list, as in these cases, our ability to counteract the choice architecture imposed by the algorithm is much more limited.

In the case of proactive recommendations or unprompted information designed to act as a nudge, the user need not have taken any prior decision to make a philanthropic gift, so it would be possible to influence not only where and how they give but their choice to do so in the first place. In this situation, it would be important to examine what motivation the search provider or platform had for encouraging or nudging prosocial behavior in the form of giving. For instance, if it reflects a broader corporate social purpose, is that purpose explicit and known to the user? Is the aim merely to encourage giving in a generic sense, or is the information presented in such a way as to promote giving to specific types of causes or according to a specific philanthropic ideology?

In previous cases where generic service providers have decided to offer some form of giving functionality, such as the ability to make donations when using ATMs or paying at a checkout, concerns have been expressed about the limitations of the options available to users and the role the service provider plays in determining them (Caulfield et al., 2022). In the case of ATM giving or checkout donations, it is a relatively straightforward matter for users to bypass the nudge. In the case of nudges applied online, however, as we have already seen, the reduced visibility of the technology may lead to users being less aware that a choice architect is attempting to shape their behavior – which might make the nudge more effective but also brings a greater risk of eroding users’ autonomy (Lv & Huang, 2022). In cases where hyper-nudges are implemented, the iterative and adaptive nature of the algorithms raises a further challenge in terms of an increased “burden of avoidance” (Susser, 2019), since it also becomes far more difficult for users to bypass the nudge.

## **4.2 Giving platforms**

Many platforms are specifically dedicated to providing information and offering transactional services that encourage and facilitate giving. Some of these are themselves nonprofits or charitable organizations, while others may have for-profit organizational structures but with a stated social purpose. Some focus primarily on traditional philanthropic donations, but others offer opportunities that go beyond this, such as the ability to make loans or to contribute to crowdfunding campaigns in support of organizations or individuals (Lilly Family School of Philanthropy, 2023).

The majority of giving platforms aim (either stated or implicit) to try to increase levels of generosity by making it easier and more appealing for people to give. Some, however, may emphasize increasing the quality of giving rather than the quantity, and to that end, may focus on promoting a particular view of effectiveness or impact. Even where platforms profess to be “cause neutral” (i.e., not promoting any one cause above others), it is essential to recognize that the range of options they present to users does not represent the entire domain of potential gift recipients since all platforms have criteria that limit what users can give to. In some cases, these criteria are explicit – such as limitations on the geographic location of recipients or requirements that they must have specific legal or organizational structures – but increasingly, there are also concerns that platforms take additional decisions to remove particular organizations or cause areas that may meet legal requirements, but which have been deemed unsuitable for some reason. As more and more giving takes place online, the concern is that platforms play the *de facto* role of arbiters of the acceptable boundaries of civil society simply because of the criteria they impose about what their users can or cannot give to, but that this power comes with little in the way of transparency or accountability (Carlman, 2020; Wade, 2022).

In many cases, users will engage with a giving platform having already made the initial decision to give, so any use of AI will be more about informing and shaping their choices about which causes to give to and what methods to use. The key questions then are about how algorithms are being applied and what their purpose or goal is. One vital dimension here is visibility: are users aware that an algorithm is being used to determine the information they are presented with? In some cases, the involvement of AI may be obvious: for instance, where a chatbot is clearly and openly being used to respond to search requests or to provide advice and recommendations. In other cases, however, the algorithmic processes may be hidden within the platform’s infrastructure, so users may not be aware of their existence. This will have a significant bearing on the degree to which we have concerns about whether the individual autonomy of users is being undermined. It may also have practical implications, as there is some evidence that people are less

likely to donate when they feel as though their ability to choose has been limited, even when they are presented with options that appeal to them (Lv & Huang, 2022).

The visibility of any algorithms used to determine the information provided by a philanthropic giving platform is an important consideration, but even more important is the question of *what those algorithms are designed to do*. In the case of giving platforms that position themselves as neutral intermediaries, presumably, the goal is to maximize user satisfaction – in much the same way as the general-purpose search interfaces considered above. What does it mean in practice to “maximize user satisfaction” when it comes to philanthropic giving? We might assume that it means presenting users with information on organizations and causes likely to appeal to them, thereby increasing the likelihood that they will make repeat donations to those same organizations and causes or to others in the future. However, there are reasons to be cautious about adopting this as an acceptable goal. For one thing, some would question whether maximizing donor satisfaction should be seen as the sole or primary aim of fundraising or philanthropy advice and would argue that the ability to challenge donors – by questioning their assumptions or pushing them outside of their comfort zone – is an important part of both professions (Beeston & Breeze, 2023). In recent years, there has also been growing debate in the fundraising field over whether “donor-centric” or “community-centric” models are preferable (MacQuillin, 2022). In the context of using AI to shape giving choices, the adoption of donor satisfaction as the primary objective may prove particularly problematic due to the recognized tendency of ML systems to exhibit “algorithmic bias” where algorithms trained on data sets that contain existing statistical biases come to reflect and entrench those biases over time. To design an algorithm that could provide giving choices and recommendations that were likely to satisfy donors, you would need to consider information such as individual user’s past donations and the giving behavior of others who share relevant demographic features. However, that data reflects the existing limitations of the philanthropic marketplace: such as the fact that large organizations receive the lion’s share of donations (National Council for Voluntary Organisations [NCVO], 2023), the fact that “unpopular” or “unfashionable” causes a struggle to raise funds (Body & Breeze, 2016), or the fact that the system as a whole demonstrates significant biases in terms of race and gender (Dorsey et al., 2020; Damm et al., 2023). The concern would be that an algorithm designed to satisfy donors and trained on this data would lead to many of these known challenges becoming exacerbated, with smaller organizations, unpopular causes, and organizations led by women or people of color becoming increasingly marginalized in favor of well-known nonprofit brands and “safe” causes.

Suppose concerns such as these lead us to conclude that donor satisfaction should not be the sole criterion when designing philanthropy algorithms. In that case the obvious question is: what other goals should we specify? We might suggest that the aim should be to maximize effectiveness, by ensuring that resources are targeted toward areas of greatest need, or ensuring they are directed toward interventions with the highest level of impact (according to some preferred measure). However, this raises both practical and theoretical challenges. In practical terms, we do not have the current data that would allow us to say where the need is most acute or where the impact is most significant, even if we wanted to. And in theoretical terms, it is not apparent that we could define either of these things meaningfully even if we did have all the required data. Identifying specific needs as “most acute” or specific interventions as having the “greatest impact” would require us to have measures of need and impact that can be applied objectively across cause areas, and many would argue that this is impractical and undesirable. Some approaches to giving, such as Effective Altruism (EA), try to address this challenge head-on by promoting the idea of a utilitarian, cause-agnostic single measure of value that can be used as a benchmark across all philanthropy (MacAskill, 2015). These are far from universally accepted, however, and EA in particular

has attracted heavy criticism from many for being unrealistically normative, for applying an overly simplistic framework to complex social problems, and for being an ideology that favors easily measurable interventions within the current status quo over harder-to-measure campaigning and advocacy that aims at fundamental reform of systems and structures that are the root cause of societal challenges (Srinivasan, 2015; Gabriel, 2017; Crary, 2023). It would certainly be possible to use the principles of EA as the basis for a cause-agnostic algorithm that could provide information and recommendations to people to maximize overall Expected Value in a utilitarian sense but to do so would reflect a clear bias toward one particular, and strongly ideological, view of philanthropy.

Since it is not possible to find a purely objective way to measure impact or need (or at least one on which there is majority consensus), any additional criteria we apply in designing the goal of a philanthropy algorithm must be subjective to some extent, so their legitimacy will derive from whatever authority we appeal to justify their adoption. In some cases, a platform may rely on its own authority if it has decided to take its own stance on what constitutes “good” giving and apply these to the designs of any algorithms it is using, but this will obviously leave it open to challenges. In other cases, a platform may choose to appeal to an external authority, such as the United Nations Sustainable Development Goals (UNSDGs). Clearly, this still does not provide a purely objective goal for giving, and the UNSDGs have their fair share of critics (Swain, 2018). Still, as a globally agreed framework for prioritizing needs and focusing actions, they also have a relatively high degree of legitimacy (and certainly more than a single giving platform would have by itself), so they may provide an appealing basis for setting the goals of a philanthropy algorithm.

Once a goal or set of criteria has been determined that can act as the basis for designing a philanthropy algorithm, there may still be a further challenge to ensure that the algorithm remains aligned with the creator’s original intention. At one time, the only option available for those attempting to create AI systems would have been to map out all of the possible steps in a process and then program those directly into an algorithm (this model is sometimes colloquially known as “Good Old Fashioned AI” or GOFAI [Boden, 2014]). This presented a significant limiting factor because it requires that we understand the underlying nature of all the capabilities we are trying to emulate and capture these in symbolic form, but there are many functions, such as natural language conversation or image recognition, that are seen as important aspects of intelligence and which humans can usually perform easily, but which we are not able to explain fully. This is one of the main reasons that the emergence of machine learning approaches as an alternative to GOFAI has led to such enormous growth and evolution in AI: because designing algorithms that can “learn” by going through a process of repeated iteration and self-modification to improve performance concerning a specified measure allows us to create systems that can approximate (or even surpass) human performance in certain tasks *without us having to specify all of the steps involved in performing that task* (Smith, 2019). In cases where machine learning algorithms match or exceed human performance, one of the things that has been noted is that they often do so by solving problems in ways that never would have occurred to a human and perhaps aren’t even fully understandable to us. A growing body of literature has explored this phenomenon in various contexts; for instance, algorithms that are designed to play video games, where it has been found that they often end up achieving their set goal of scoring highly or winning by engaging in “specification gaming” (i.e., looking for loopholes or weaknesses in how the task has been specified or in the video game’s design) rather than by playing within the confines of the game in the way that a human would (Krakovna et al., 2020; Lehman et al., 2020). This makes it clear that our *goals* and *values* in designing ML algorithms are hugely important. If we may have little or no control over *how* algorithms evolve to achieve a particular result, it becomes absolutely vital that we can stipulate clearly what our desired goal is and what the acceptable parameters are when it comes to achieving it.



The idea of Value Alignment has come to prominence through the work of philosopher Nick Bostrom, who argues that one of the key challenges of AI development is what he has christened the Value Alignment Problem (VAP), i.e., ensuring that highly autonomous AI systems are designed to ensure that their goals and behaviors remain aligned with human value throughout their operation (Bostrom, 2014). Bostrom’s work focuses on Artificial General Intelligence (AGI) and Superintelligence, i.e., AI systems that are capable of matching or surpassing human-level intelligence concerning any task, which, at this point, remains hypothetical. To demonstrate the potential risks of value misalignment, he created a thought experiment known as The Paperclip Maximizer, in which a highly intelligent AGI is given the simple task of producing as many paperclips as possible (Bostrom, 2003). The danger, Bostrom argues, is that without additional specification of constraints, the AGI may choose to maximize its performance of this task in ways that radically diverge from our original intent: perhaps it will decide, for instance, to wipe out the human race because it believes that human beings pose a threat (because they may order it to cease paperclip production at some point in the future) or simply because we represent valuable stores of raw materials that the AGI believes would be better used for making more paperclips. The Value Alignment Problem doesn’t just apply to Superintelligent systems or AGI, however. It is also relevant to domain-specific AI systems of this kind we have today. And it is not limited to trivial examples of algorithms bending the rules win at computer games either: as AI systems are increasingly deployed by governments and companies in a wide range of contexts, they have the potential to affect many areas of our lives, so ensuring that they remain aligned with our values and intentions is crucial (Korinek & Balwit, 2022).

In the context of philanthropy algorithms being deployed by giving platforms, as in many other areas, the challenge will be to find ways of giving AI systems sufficient freedom and autonomy to bring benefits in the form of increased effectiveness and efficiency while minimizing the risk of negative unintended consequences. There is a growing body of literature focused on “AI safety” that proposes ways of achieving this balance by designing safeguards that enable suitable human oversight and corrective acts when the risk of unintended consequences do occur (Leslie, 2019; Houben et al., 2022), and it is important that those seeking to create philanthropy algorithms draw on this literature.

### ***4.3 Cause-based organizations***

The final context in which AI may impact philanthropic giving choices is that of individual cause-based organizations (e.g., charities, social enterprises, etc.). In these cases, the use of AI is unlikely to involve the design of philanthropy algorithms in the sense discussed above since the organization’s interest (presumably) is not providing information or recommendations that maximize user satisfaction or facilitate giving in a generic sense but in maximizing their support. Cause-based organizations may, of course, still benefit from algorithms used by giving platforms or general-purpose information providers prioritizing them in search results or including them in recommendations. In this situation, do these cause-based organizations, as beneficiaries, bear any responsibility for concerns about algorithmic bias, loss of autonomy for users, or lack of transparency? Since we assume they neither designed the algorithms nor controlled their operation, these cause-based organizations are not directly responsible. However, if they are aware of the issues or have even engaged in practices that might exacerbate them (such as paying to be ranked higher in search results or participating in hyper-nudging initiatives where they stand to receive donations), then it might be argued that they are complicit to some extent and therefore bear a share of any moral responsibility.

Cause-based organizations may also face direct ethical issues if they use generative AI tools to produce content as part of their fundraising. Charities and nonprofits have always used emotive imagery and storytelling to appeal to people via the heart and the head, which has sometimes drawn criticism. In the context of international aid and development, for instance, there is a long-standing debate over whether depictions of aid recipients are patronizing and overly negative and whether this reflects problematic attitudes of “white saviorism” (Pieterse, 1992; Bhati, 2021). Similarly, disability rights campaigners have, at times, been vocally critical of the depiction of disabled people as objects of pity by nonprofit organizations (Longmore, 2015). These concerns will apply equally to content created using generative AI. In fact, they may even be exacerbated, as there are concerns that current AI image-generation tools demonstrate worrying levels of racial and gender bias, so their use may lead to the further perpetuation of problematic stereotypes (Lamensch, 2023; Small, 2023; Thomas & Thomson, 2023; Turk, 2023).

The capabilities of generative AI will bring other challenges, too. The fact that it is now possible to generate photo-realistic images or deepfake videos that are indistinguishable from real photos or video has led to concerns being raised about the potential impact this might have on notions of trust and authenticity (de Ruiter, 2021). For cause-based organizations, the risk of using generative AI content is that if done badly, it may have a significant negative impact on the perceptions of supporters and the wider public. In 2019, the UK charity Malaria No More UK successfully used deepfake technology to produce footage of the former footballer David Beckham reading out an appeal in nine different languages. It received broadly positive coverage (Davies, 2019). In early 2023, however, Amnesty International was heavily criticized for using AI-generated imagery to promote a report on police brutality in Colombia (Taylor, 2023). In the context of fundraising, there may also be reasons for caution when it comes to embracing generative AI: at least one study, for instance, has found that people are less likely to donate in response to a charitable giving advert if they become aware that AI-generated imagery has been used (Arango et al., 2023).

There are also wider ethical questions as to whether generative AI is inherently parasitic since it requires vast data sets of images, photographs, or text on which algorithms can be trained, which are only possible through the past efforts of human writers, artists, and creatives. Critics have accused companies at the forefront of the generative AI revolution of engaging in deliberate, large-scale copyright infringement to build their products, and this seems set to become a major issue for generative AI in the coming years, with several legal challenges already mounted and more likely to follow (Appel et al., 2023; Bearne, 2023). Again, this is not an issue for which charities or nonprofits bear direct responsibility, as they have no control over the development of commercial generative AI tools. They do, however, have a choice as to whether they use these tools or not. While broad concerns about copyright infringement and intellectual property rights are unlikely to be sufficient grounds by themselves for shunning generative AI, when added to the other ethical issues outlined already, it may be seen by some charities and nonprofits as sufficient grounds to question whether the use of generative AI is appropriate at all at this stage.

## **5 Conclusion**

In this chapter, we have considered a range of ways in which AI could be used to shape individual choices and how these apply to different contexts in which decisions about philanthropic giving might be made. In doing so, we have identified several potential ethical concerns that funders and nonprofits need to be aware of as they contemplate using AI tools in this way. These ethical concerns and some of the key questions we need to ask as a result are summarized below:

*Legitimacy, accountability, and transparency*

- Is it clear to users that an algorithm determines the information or recommendations offered to them?
- Who designed the algorithm?
- What goal did they have in mind in designing the algorithm?
- Is this goal clear to the user?
- Where does their legitimacy of the goal derive from?
- Can the user challenge the algorithm or hold the designer to account?

*Undermining agency*

- Does the use of AI to create personalized choice architectures or hyper-nudges undermine the agency of individual donors?
- Is this an acceptable price to pay for encouraging prosocial behavior?

*Bias*

- Does using algorithms to provide information and recommendations that can shape philanthropic choice introduce the risk of bias against certain kinds of organizations or causes?
- Does the use of Generative AI bring the risk of bias (e.g., for race, gender, etc.), which could exacerbate existing concerns about how recipients of philanthropic funding are portrayed?

*Erosion of authenticity and trust*

- Does the use of generative AI to produce content (e.g., photographic imagery, video, text) for use in fundraising and campaigning bring the risk of contributing to an erosion of the notions of authenticity and trust in the online environment?

*Intellectual property rights and training data*

- Are current AI tools dependent on data sets that take advantage of prior human effort without suitable compensation or respect for intellectual property rights? Are nonprofit organizations that use these tools ethically compromised as a result?

Having identified these ethical issues, the question is what actions can be taken to address them. There is no single answer, but we can identify a range of current and future actions that can be taken by the various actors involved. Commercial platforms need to recognize the power they have to shape individual giving choices. They should seek to work with the nonprofit sector to minimize any potential harms that come from applying algorithmic processes to the provision of information and recommendations. Dedicated giving platforms are more likely to have pre-existing relationships with the nonprofit sector but should still recognize that their power as gatekeepers will increase as more and more giving takes places online and that they therefore have a responsibility to ensure that any applications of algorithms take into account the risks of undermining the agency of donors, or introducing biases that will adversely affect certain types of organizations or cause areas. Nonprofits and civil society organizations also need to be aware of the potential risks inherent

in their own use of AI tools – and ensure they put in place measures to mitigate against them. This may seem daunting for organizations that are often resource-starved, which is why philanthropic funders also have a vital role in providing them with the infrastructure and support they need to engage with issues of AI ethics. This could be done through relationships with individual grantees or by creating new pooled funds (such as the European Artificial Intelligence and Society Fund). Philanthropic funders can also play a valuable role in broader efforts to ensure the ethical and responsible use of AI, by engaging in research and advocacy that enables the perspectives and insights of civil society organizations they support to be brought into the debates.

## References

- Allen, S. (2018). *The Science of Generosity* [White Paper]. Prepared for the John Templeton Foundation by the Greater Good Science Center at UC Berkeley. [https://ggsc.berkeley.edu/images/uploads/GGSC-JTF\\_White\\_Paper-Generosity-FINAL.pdf](https://ggsc.berkeley.edu/images/uploads/GGSC-JTF_White_Paper-Generosity-FINAL.pdf)
- Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92(5–6), 1047–1060. <https://doi.org/10.1016/j.jpubeco.2007.11.004>
- André, Q., Carmon, Z., Wertenbroch, K., Crum, A., Frank, D., Goldstein, W., Huber, J., Van Boven, L., Weber, B., & Yang, H. (2018). Consumer choice and autonomy in the age of artificial intelligence and big data. *Customer Needs and Solutions*, 5(1–2), 28–37. <https://doi.org/10.1007/s40547-017-0085-8>
- Andreoni, J., & Payne, A. (2011). Is crowding out due entirely to fundraising? Evidence from a panel of charities. *Journal of Public Economics*, 95(5–6), 334–343. <https://doi.org/10.1016/j.jpubeco.2010.11.011>
- Andreoni, J., Payne, A. A., Smith, J., & Karp, D. (2016). Diversity and donations: The effect of religious and ethnic diversity on charitable giving. *Journal of Economic Behavior & Organization*, 128, 47–58. <https://doi.org/10.1016/j.jebo.2016.05.010>
- Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, 125(3), 625–653. <https://doi.org/10.1086/691703>
- Appel, G., Neelbauer, J., & Schweidel, D. (2023, April 7). Generative AI has an intellectual property problem. *Harvard Business Review*. <https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem>
- Arango, L., Singaraju, S. P., & Niininen, O. (2023). Consumer responses to AI-generated charitable giving ads. *Journal of Advertising*, 52(4), 486–503. <https://doi.org/10.1080/00913367.2023.2183285>
- Arora, N. (2021, November 12). *The Value of Getting Personalization Right—Or Wrong—Is Multiplying*. McKinsey. <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying>
- Barraza, J. A., McCullough, M. E., Ahmadi, S., & Zak, P. J. (2011). Oxytocin infusion increases charitable donations regardless of monetary resources. *Hormones and Behavior*, 60(2), 148–151. <https://doi.org/10.1016/j.yhbeh.2011.04.008>
- Bartmann, M. (2023). Reasoning with recommender systems? Practical reasoning, digital nudging, and autonomy. In S. Genovesi, K. Kaesling, & S. Robbins (Eds.), *Recommender Systems: Legal and Ethical Issues* (Vol. 40, pp. 129–145). Springer International Publishing. [https://doi.org/10.1007/978-3-031-34804-4\\_7](https://doi.org/10.1007/978-3-031-34804-4_7)
- Bearne, S. (2023, August 1). New AI systems collide with copyright law. *BBC*. <https://www.bbc.co.uk/news/business-66231268>
- Beeston, E., & Breeze, B. (2023, May 16). *Disciplining Generosity*. Stanford Social Innovation Review. [https://ssir.org/books/excerpts/entry/disciplining\\_generosity](https://ssir.org/books/excerpts/entry/disciplining_generosity)
- Behavioural Insights Team (2013). *Applying Behavioural Insights to Charitable Giving* [White Paper]. Crown. [https://assets.publishing.service.gov.uk/media/5a7516a1ed915d6f2b228b/BIT\\_Charitable\\_Giving\\_Paper.pdf](https://assets.publishing.service.gov.uk/media/5a7516a1ed915d6f2b228b/BIT_Charitable_Giving_Paper.pdf)
- Bekkers, R., & Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(5), 924–973. <https://doi.org/10.1177/0899764010380927>
- Bernholz, L. (2021). *How We Give Now: A Philanthropic Guide for the Rest of Us*. Cambridge: MIT Press.
- Bhati, A. (2021). Is the representation of beneficiaries by international nongovernmental organizations (INGOs) still pornographic? *Journal of Philanthropy and Marketing*, e1722. <https://doi.org/10.1002/nvsm.1722>

- Boden, M. A. (2014). 4 GOFAI. In K. Frankish and W.M. Ramsey (Eds.). *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139046855.007>
- Body, A., & Breeze, B. (2016). What are ‘unpopular causes’ and how can they achieve fundraising success? *International Journal of Nonprofit and Voluntary Sector Marketing*, 21(1), 57–70. <https://doi.org/10.1002/nvsm.1547>
- Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. In I. Smith et al. (Eds.). *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence* (Vol 2, ed. I, pp. 12–17). Institute of Advanced Studies in Systems Research and Cybernetics. <https://philpapers.org/archive/BOSEII.pdf>
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bovens, L. (2009). The ethics of nudge. In T. Grüne-Yanoff and S. O. Hansson (Eds.). *Preference Change: Approaches from Philosophy, Economics and Psychology* (pp. 207–219). Springer. <https://philpapers.org/archive/BOVTEO-8.pdf>
- BrightEdge. (2019). *Channel Report 2019*. BrightEdge. [https://videos.brightedge.com/research-report/BrightEdge\\_ChannelReport2019\\_FINAL.pdf](https://videos.brightedge.com/research-report/BrightEdge_ChannelReport2019_FINAL.pdf)
- Brown, P. H., & Minty, J. H. (2008). Media coverage and charitable giving after the 2004 tsunami. *Southern Economic Journal*, 75(1), 9–25. <https://doi.org/10.1002/j.2325-8012.2008.tb00889.x>
- Butera, L., & Horn, J. (2020). “Give less but give smart”: Experimental evidence on the effects of public information about quality on giving. *Journal of Economic Behavior & Organization*, 171, 59–76. <https://doi.org/10.1016/j.jebo.2020.01.011>
- Capraro, V., Jagfeld, G., Klein, R., Mul, M., & van de Pol, I. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, 9(1), 11880. <https://doi.org/10.1038/s41598-019-48094-4>
- Carlman, A. (2020, November 17). *Platform Neutrality Is Dead. Long Live Empathy*. Digital Impact. <https://digitalimpact.io/platform-neutrality-is-dead-long-live-empathy/>
- Caulfield, J. L., Baird, C. A., & Lee, F. K. (2022). The ethicality of point-of-sale marketing campaigns: Normative ethics applied to cause-related checkout charities. *Journal of Business Ethics*, 1–16. <https://doi.org/10.1007/s10551-020-04597-z>
- Chesney, R., & Citron, D. (2018, December 11). Deepfakes and the new disinformation war. *Foreign Affairs*. <https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>
- Cordelli, C. (2016). Reparative justice and the moral limits of discretionary philanthropy. In R. Reich, C. Cordelli & L. Bernholz (Eds.). *Philanthropy in Democratic Societies: History, Institutions, Values* (pp. 244–265). Chicago, IL: The University of Chicago Press.
- Crary, A. (2023). Against “Effective Altruism”. In C. J. Adams, A. Crary & L. Gruen (Eds.). *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780197655696.003.0016>
- Crosan, R., & Shang, J. (Yue) (2008). The impact of downward social information on contribution decisions. *Experimental Economics*, 11(3), 221–233. <https://doi.org/10.1007/s10683-007-9191-z>
- Damm, C., Dowrick, L., & Harris, C. (2023). *Mapping the UK Women and Girls Sector and Its Funding: Where Does the Money Go?* Rosa. <https://rosauk.org/wp-content/uploads/2023/04/Women-and-Girls-Sector-Research-Mapping-Report-Amended.pdf>
- Davies, G. (2019, April 9). David Beckham ‘Speaks’ 9 languages for new campaign to end malaria. *ABC News*. <https://abcnews.go.com/International/david-beckham-speaks-languages-campaign-end-malaria/story?id=62270227>
- De Fine Licht, K., & De Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & SOCIETY*, 35, 917–926. <https://rdcu.be/dprkT>
- De Ruyter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, 34(4), 1311–1332. <https://doi.org/10.1007/s13347-021-00459-2>
- Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7), 2072–2098. <https://doi.org/10.1177/1461444820925811>
- Dorsey, C., Bradach, J., & Kim, P. (2020). *Racial Equity and Philanthropy: Disparities in Funding for Leaders of Color Leave Impact on the Table*. The Bridgespan Group. <https://www.bridgespan.org/getmedia/05ad1f12-2419-4039-ac67-a45044f940ec/racial-equity-and-philanthropy.pdf>

- Eckel, C., Grossman, P. J., & Milano, A. (2007). Is more information always better? An experimental study of charitable giving and Hurricane Katrina. *Southern Economic Journal*, 74(2), 388–411. <https://doi.org/10.1002/j.2325-8012.2007.tb00845.x>
- Ekström, M. (2012). Do watching eyes affect charitable giving? Evidence from a field experiment. *Experimental Economics*, 15(3), 530–546. <https://doi.org/10.1007/s10683-011-9312-6>
- Espin-Noboa, L., Wagner, C., Strohmaier, M., & Karimi, F. (2022). Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific Reports*, 12(1), 2012. <https://doi.org/10.1038/s41598-022-05434-1>
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.
- Fathi, M., Bateson, M., & Nettle, D. (2014). Effects of watching eyes and norm cues on charitable giving in a surreptitious behavioral experiment. *Evolutionary Psychology*, 12(5). <https://doi.org/10.1177/147470491401200502>
- Fukui, H., & Toyoshima, K. (2014). Chill-inducing music enhances altruism in humans. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.01215>
- Gabriel, I. (2017). Effective altruism and its critics. *Journal of Applied Philosophy*, 34(4), 457–473. <https://doi.org/10.1111/japp.12176>
- Gewirth, A. (1987). Private philanthropy and positive rights. *Social Philosophy and Policy*, 4(2), 55–78. <https://doi.org/10.1017/S0265052500000546>
- Greitemeyer, T. (2009). Effects of songs with prosocial lyrics on prosocial behavior: Further evidence and a mediating mechanism. *Personality and Social Psychology Bulletin*, 35(11), 1500–1511. <https://doi.org/10.1177/0146167209341648>
- Gurdeniz, E., & Hosanagar, K. (2023, February 23). Generative AI won't revolutionize search—yet. *Harvard Business Review*. <https://hbr.org/2023/02/generative-ai-wont-revolutionize-search-yet>
- Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 149–152. <https://doi.org/10.1089/cyber.2021.29208.jth>
- Harbaugh, W. T., & Krause, K. (2000). Children's altruism in public good and dictator experiments. *Economic Inquiry*, 38(1), 95–109. <https://doi.org/10.1111/j.1465-7295.2000.tb00006.x>
- Hobbs, J. (2017). Nudging charitable giving: The ethics of Nudge in international poverty reduction. *Ethics & Global Politics*, 10(1), 37–57. <https://doi.org/10.1080/16544951.2017.1312991>
- Houben, S., Abrecht, S., Akila, M., Bär, A., Brockherde, F., Feifel, P., Fingscheidt, T., Gannamaneni, S. S., Ghobadi, S. E., & Hammam, A. (2022). Inspect, understand, overcome: A survey of practical methods for AI safety. In T. Fingscheidt, H. Gottschalk & S. Houben (Eds.). *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights towards Safety* (pp. 3–78). Cham: Springer International Publishing Cham. [https://doi.org/10.1007/978-3-031-01233-4\\_1](https://doi.org/10.1007/978-3-031-01233-4_1)
- Huang, K. (2022, September 17). For Gen Z, TikTok is the new search engine. *The New York Times*. <https://www.nytimes.com/2022/09/16/technology/gen-z-tiktok-search-engine.html>
- Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, 14, 235–257. <https://doi.org/10.1023/A:1007740225484>
- Jonas, E., Schimel, J., Greenberg, J., & Pyszczynski, T. (2002). The scrooge effect: Evidence that mortality salience increases prosocial attitudes and behavior. *Personality and Social Psychology Bulletin*, 28(10), 1342–1353. <https://doi.org/10.1177/014616702236834>
- Kanter, B., & Fine, A. H. (2022). *The Smart Nonprofit: Staying Human-Centered in an Automated World*. Hoboken, NJ: John Wiley & Sons, Incorporated.
- Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, 97(5), 1774–1793. <http://dx.doi.org/10.1257/aer.97.5.1774>
- Korinek, A., & Balwit, A. (2022). *Aligned with Whom? Direct and Social Goals for AI Systems* [NBER Working Paper No. w30017]. National Bureau of Economic Research. <https://dx.doi.org/10.2139/ssrn.4104003>
- Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020, April 21). *Specification Gaming: The Flip Side of AI Ingenuity*. DeepMind Blog, 3. <https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity>
- Lamensch, M. (2023, June 14). *Generative AI Tools Are Perpetuating Harmful Gender Stereotypes*. Centre for International Governance Innovation. <https://www.cigionline.org/articles/generative-ai-tools-are-perpetuating-harmful-gender-stereotypes/>

- Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., & Bryson, D. M. (2020). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life*, 26(2), 274–306. [https://doi.org/10.1162/artl\\_a\\_00319](https://doi.org/10.1162/artl_a_00319)
- Leslie, D. (2019). *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. The Alan Turing Institute. [https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf)
- Liao, Q. V., Geyer, W., Muller, M., & Khazaen, Y. (2020). Conversational interfaces for information search. In W. T. Fu & H. van Oostendorp (Eds.). *Understanding and Improving Information Search: A Cognitive Approach* (pp. 267–287). Cham: Springer Cham.
- Lilly Family School of Philanthropy (2023). *Digital for Good: A Global Study on Emerging Ways of Giving*. Lilly Family School of Philanthropy, Indiana University Indianapolis. <https://scholarworks.iupui.edu/server/api/core/bitstreams/528c7dd7-4fb0-4af9-875a-1254a4e3034d/content>
- Loi, M., & Spielkamp, M. (2021). Towards accountability in the use of artificial intelligence for public administrations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 757–766. <https://doi.org/10.1145/3461702.3462631>
- Longmore, P. K. (2015). *Telethons: Spectacle, Disability, and the Business of Charity*. Oxford: Oxford University Press.
- Lv, L., & Huang, M. (2022). Can personalized recommendations in charity advertising boost donation? The role of perceived autonomy. *Journal of Advertising*, 1–18. <https://doi.org/10.1080/00913367.2022.2109082>
- MacAskill, W. (2015). *Doing Good Better: Effective Altruism and a Radical New Way to Make a Difference*. London: Guardian Faber Publishing.
- MacQuillin, I. (2022). Normative fundraising ethics: A review of the field. *Journal of Philanthropy and Marketing*, e1740. <https://doi.org/10.1002/nvsm.1740>
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissoziating Language and thought in Large Language Models: A Cognitive Perspective* (arXiv: 2301.06627). arXiv. <http://arxiv.org/abs/2301.06627>
- Maillé, P., Maudet, G., Simon, M., & Tuffin, B. (2022). Are search engines biased? Detecting and reducing bias using meta search engines. *Electronic Commerce Research and Applications*, 101132. <https://doi.org/10.1016/j.elerap.2022.101132>
- Martin, M. W. (1994). *Virtuous Giving: Philanthropy, Voluntary Service, and Caring*. Indianapolis: Indiana University Press.
- Meer, J. (2011). Brother, can you spare a dime? Peer pressure in charitable solicitation. *Journal of Public Economics*, 95(7–8), 926–941. <https://doi.org/10.1016/j.jpubeco.2010.11.026>
- Mills, S. (2022). Finding the ‘nudge’ in hypernudge. *Technology in Society*, 71, 102117. <https://doi.org/10.1016/j.techsoc.2022.102117>
- Mills, S., & Sætra, H. S. (2022). The autonomous choice architect. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-022-01486-z>
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. <https://doi.org/10.1145/3287560.3287574>
- Moll, J., Krueger, F., Zahn, R., Pardini, M., De Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences*, 103(42), 15623–15628. <https://doi.org/10.1073/pnas.0604475103>
- National Council for Voluntary Organisations (2023). *UK Civil Society Almanac 2023*. National Council for Voluntary Organisations. <https://www.ncvo.org.uk/news-and-insights/news-index/uk-civil-society-almanac-2023/financials/#/>
- Noble, S. U. (2018). *Algorithms of Oppression*. New York: New York University Press.
- Nozick, R. (1974). *Anarchy, State, and Utopia*. Hoboken, NJ: John Wiley & Sons.
- Pashentsev, E. (2023). The malicious use of deepfakes against psychological security and political stability. In E. Pashentsev (Ed.). *The Palgrave Handbook of Malicious Use of AI and Psychological Security* (pp. 47–80). Cham: Palgrave Macmillan Cham. [https://doi.org/10.1007/978-3-030-35746-7\\_3](https://doi.org/10.1007/978-3-030-35746-7_3)
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Payton, R. L., & Moody, M. P. (2008). *Understanding Philanthropy: Its Meaning and Mission*. Indianapolis: Indiana University Press.

- Perez, S. (2022, July 12). *Google Exec Suggests Instagram and TikTok Are Eating into Google's Core Products, Search and Maps*. TechCrunch. <https://techcrunch.com/2022/07/12/google-exec-suggests-instagram-and-tiktok-are-eating-into-googles-core-products-search-and-maps/>
- Pieterse, J. N. (1992). *White on Black: Images of Africa and Blacks in Western Popular Culture*. New Haven, CT: Yale University Press.
- Piff, P. K., Kraus, M. W., Côté, S., Cheng, B. H., & Keltner, D. (2010). Having less, giving more: The influence of social class on prosocial behavior. *Journal of Personality and Social Psychology*, 99(5), 771–784. <https://doi.org/10.1037/a0020092>
- Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48, 137–141. <https://doi.org/10.1007/s11747-019-00710-5>
- Reich, R., Cordelli, C., & Bernholz, L. (2016). *Philanthropy in Democratic Societies: History, Institutions, Values*. Chicago, IL: University of Chicago Press.
- Roberts, S. (1996). Contexts of charity in the middle ages: Religious, social, and civic. In J.B. Schneewind (Ed.), *Giving: Western Ideas of Philanthropy* (pp. 24–53). Indianapolis: Indiana University Press.
- Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1), 59. <https://doi.org/10.1186/s40537-022-00592-5>
- Rudd, M., Vohs, K. D., & Aaker, J. (2012). Awe expands people's perception of time, alters decision making, and enhances well-being. *Psychological Science*, 23(10), 1130–1136. <https://doi.org/10.1177/0956797612438731>
- Ruehle, R. C., Engelen, B., & Archer, A. (2021). Nudging charitable giving: What (if anything) is wrong with it? *Nonprofit and Voluntary Sector Quarterly*, 50(2), 353–371. <https://doi.org/10.1177/0899764020954266>
- Saeri, A. K., Slattery, P., Lee, J., Houlden, T., Farr, N., Gelber, R. L., Stone, J., Huuskes, L., Timmons, S., Windle, K., Spajic, L., Freeman, L., Moss, D., Behar, J., Schubert, S., Grundy, E. A. C., & Zorker, M. (2023). What works to increase charitable donations? A meta-review with meta-meta-analysis. *VOL-UNTAS: International Journal of Voluntary and Nonprofit Organizations*, 34(3), 626–642. <https://doi.org/10.1007/s11266-022-00499-y>
- Sætra, H. S. (2023). Generative AI: Here to stay, but for good? *Technology in Society*, 75, 102372. <https://doi.org/10.1016/j.techsoc.2023.102372>
- Salmon, J. (2023). *Protecting Donor Intent Protects Giving*. The Philanthropy Roundtable. [https://www.philanthropyroundtable.org/wp-content/uploads/2023/07/Protecting-Donor-Intent-Protects-Giving\\_.pdf](https://www.philanthropyroundtable.org/wp-content/uploads/2023/07/Protecting-Donor-Intent-Protects-Giving_.pdf)
- Samuel, L. R. (2010). *Freud on Madison Avenue: Motivation Research and Subliminal Advertising in America*. Philadelphia: University of Pennsylvania Press.
- Schmidt, A. (2021). The end of serendipity: Will artificial intelligence remove chance and choice in everyday life? *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, 1–4. <https://doi.org/10.1145/3464385.3464763>
- Schneewind, J. B. (1996). *Giving: Western Ideas of Philanthropy*. Indianapolis: Indiana University Press.
- Schrage, M. (2020). *Recommendation Engines*. Cambridge: MIT Press.
- Schulz, J. F., Thiemann, P., & Thöni, C. (2017). Nudging generosity: Choice architecture and cognitive factors in charitable giving. *Journal of Behavioral and Experimental Economics*, 74, 139–145. <https://doi.org/10.1016/j.socec.2018.04.001>
- Sebastián-Enesco, C., & Warneken, F. (2015). The shadow of the future: 5-year-olds, but not 3-year-olds, adjust their sharing in anticipation of reciprocation. *Journal of Experimental Child Psychology*, 129, 40–54. <https://doi.org/10.1016/j.jecp.2014.08.007>
- Shang, J., & Croson, R. (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119(540), 1422–1439. <https://doi.org/10.1111/j.1468-0297.2009.02267.x>
- Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, 18(9), 803–809. <https://doi.org/10.1111/j.1467-9280.2007.01983.x>
- Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, 1(3), 229–243. <https://www.jstor.org/stable/2265052>
- Singer, P. (2006, December 17) What should a billionaire give and what should you? *New York Times*. <https://www.nytimes.com/2006/12/17/magazine/17charity.t.html>
- Singer, P. (2015). *The Most Good You Can Do: How Effective Altruism Is Changing Ideas about Living Ethically*. Melbourne: Text Publishing.



- Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making*, 2(2), 79–95. <https://doi.org/10.1017/S1930297500000061>
- Small, Z. (2023, July 4). Black artists say AI shows bias, with algorithms erasing their history. *New York Times*. <https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html>
- Smith, B. C. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge: MIT Press.
- Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, 34(5), 317–322. <https://doi.org/10.1016/j.evolhumbehav.2013.05.001>
- Srinivasan, A. (2015). Stop the robot apocalypse: The new utilitarians. *London Review of Books*, 37(18), 3–6. <https://www.lrb.co.uk/the-paper/v37/n18/amia-srinivasan/stop-the-robot-apocalypse>
- Susser, D. (2017). Transparent media and the development of digital habits. In Y. Van den Eede, S. O. Irwin & G. Wellner (Eds.). *Postphenomenology and Media: Essays on Human-Media-World Relations* (pp. 27–44). Lexington Books. <https://philarchive.org/rec/SUSTMA>
- Susser, D. (2019). Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 403–408. <https://doi.org/10.1145/3306618.3314286>
- Swain, R. B. (2018). A critical analysis of the sustainable development goals. In W. L. Filho (Ed). *Handbook of Sustainability Science and Research* (pp. 341–355). Cham: Springer Cham
- Taylor, L. (2023, May 2). Amnesty international criticised for using AI-generated images. *The Guardian*. <https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism>
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. London: Penguin.
- Thomas, A. M., Parkinson, J., Moore, P., Goodman, A., Xhafa, F., & Barolli, L. (2013). Nudging through technology: Choice architectures and the mobile information revolution. *2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 255–261. <https://doi.org/10.1109/3PGCIC.2013.44>
- Thomas, R. J., & Thomson, T. J. (2023). What does a journalist look like? Visualizing journalistic roles through AI. *Digital Journalism*, 1–23. <https://doi.org/10.1080/21670811.2023.2229883>
- Turk, V. (2023, October 10). *How AI Reduces the World to Stereotypes*. Rest of World. <https://restofworld.org/2023/ai-image-stereotypes/>
- Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology*, 35(4), 88. <https://doi.org/10.1007/s13347-022-00577-5>
- Van Den Eede, Y. (2011). In between us: On the transparency and opacity of technological mediation. *Foundations of Science*, 16, 139–159. <https://doi.org/10.1007/s10699-010-9190-y>
- von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, 34(4), 1607–1622. <https://doi.org/10.1007/s13347-021-00477-0>
- Wachter-Boettcher, S. (2017). *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. New York: WW Norton & Company.
- Wade, M. (2022). 'The giving layer of the internet': A critical history of GoFundMe's reputation management, platform governance, and communication strategies in capturing peer-to-peer and charitable giving markets. *Journal of Philanthropy and Marketing*, e1777. <https://doi.org/10.1002/nvsm.1777>
- Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765), 1301–1303. <https://doi.org/10.1126/science.1121448>
- Weinmann, M., Schneider, C., & Brocke, J. V. (2016). Digital nudging. *Business & Information Systems Engineering*, 58(6), 433–436. <https://doi.org/10.1007/s12599-016-0453-1>
- Winfrey, J. C. (1981). Charity versus justice in Locke's theory of property. *Journal of the History of Ideas*, 42(3), 423. <https://doi.org/10.2307/2709185>
- Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector – applications and challenges. *International Journal of Public Administration*, 42(7), 596–615. <https://doi.org/10.1080/01900692.2018.1498103>
- Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, 20(1), 118–136. <https://doi.org/10.1080/1369118X.2016.1186713>
- Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology*, 28(1), 126. <https://psycnet.apa.org/doi/10.1037/0012-1649.28.1.126>
- Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS One*, 2(11), e1128. <https://doi.org/10.1371/journal.pone.0001128>
- Zamani, H., Trippas, J. R., Dalton, J., & Radlinski, F. (2023). *Conversational Information Seeking* (arXiv: 2201.08808). arXiv. <http://arxiv.org/abs/2201.08808>

# SHAPING THE ETHICAL AND INCLUSIVE AI REVOLUTION

## Five roles for philanthropies

*Ulla Jasper, Siddhartha Jha and Stefan Germann*

### 1 Preface: roads not taken<sup>1</sup>

In a world where the COVID-19 pandemic was considered the defining event of our era, an inconspicuous laboratory in the heart of Silicon Valley had already embarked on a mission that would captivate the imagination of technology enthusiasts and the larger society alike.

In this tale, our protagonist is not the virus that swept across the globe, but rather a small yet audacious lab. Founded through the generous support of visionary donors, this lab gathered a remarkable team of minds from diverse fields including the best computer scientists, all sharing an unwavering belief: that Artificial General Intelligence could be the pinnacle of human achievement (Altman, 2023). Their journey, however, was anything but smooth. Early challenges loomed large, threatening to derail their ambitious quest.

But then, a twist of fate brought philanthropic foundations and the scientific arms of several national governments into the picture. Philanthropic funders, united in their foresight to support ideas that could serve humanity and potentially impact all Sustainable Development Goals, stepped forward to pool resources to ensure that the lab had access to the computing and hardware resources it needed in its early years. The decision-makers within the government, themselves luminaries in the fields of science and technology, recognized the lab's potential and played a pivotal role in identifying critical public infrastructures of the future.

What followed was a groundbreaking partnership that would redefine the course of AI development. Public funds were strategically allocated to protect the lab's non-profit, open mission while ensuring it had access to state-of-the-art technology and resources. A delicate balance was achieved, in which the lab's autonomy and innovation were not hampered by public sector bureaucracy. Simultaneously, the private sector stepped in to provide crucial support without seeking control. This partnership prioritized AI safety research and allowed the lab to operate without the pressure of immediate profit and return. United Nations organizations, in their role as custodians of the common global interest and peaceful development, quickly came together to ensure that the interests and the rights of smaller nation-states were adequately represented in the ensuing debate and developments. In this unique ecosystem, untested and unsafe AI solutions were not prematurely unleashed on the world, and an unstoppable AI arms race was not triggered.

Yet, as compelling as this alternate narrative is, it stands in stark contrast to the reality we know. The lab's path in our world has been marked by formidable challenges, including the withdrawal of a key benefactor. To survive, it had to transform itself into a capped for-profit entity, raising questions about its original mission and setting in motion a series of developments that no one could put a stop to.

This story serves as an intriguing thought exercise to the complex web of possibilities and decisions that shape the development of AI. The reality of what is missing in our lives and societies is easier to grasp when compared to an idealized utopia. There are myriad ways in which such a utopian tale can fall apart, and many factors and actors must come together to make the perfect path possible. By having us compare our reality to its utopian narrative, the story serves the instructive purpose of leaving us with many "if only's" before we dive deeper into the debates of an ethical AI revolution. Even though the reality that has unfolded stands in stark contrast to the fictitious tale of our preface, it allows us to take a moment to collectively pause and reflect, especially on the following points.

First, corporate interests often dominate the AI landscape due to their substantial financial resources and market-driven incentives. Philanthropic entities may lack the necessary funds, coordination, conviction, or influence to compete with corporate giants in shaping AI development at the scale needed. So even when well-intentioned efforts to shape AI development are seeded, like the protagonist lab in our story, the trajectory of development is seldom smooth due to the lack of funding and foresight to support what is needed during the lifecycle of developing powerful and safe AI products and services.

Second, governmental bodies can be slowed down by bureaucratic processes, which can hinder the seamless cooperation depicted in the counterfactual of the preface. The ideal of a unified global effort toward ethical AI development is hampered by real-world geopolitical rivalries and nationalistic approaches to technological advancement. The absence of a global governing body or universally accepted regulatory frameworks for AI exacerbates the situation, making the depicted level of global cooperation difficult to achieve.

Third, the counterfactual scenario emphasizes AI safety, whereas the real-world rush to AI superiority often overlooks safety and ethical concerns. The competitive nature of the AI field and the potential economic and strategic advantages associated with AI advancements can lead to the neglect of safety considerations.

Finally, the reality of financial sustainability pressures forces many AI research entities to transition from non-profit to for-profit models, diverging from the counterfactual scenario of sustained public and private support to ensure safety. This transition may shift the focus from ethical and open missions to profitability, thereby altering the original mission of such entities.

In the following sections, we invite the reader, with a particular focus on the philanthropic funding sector, to engage in a thoughtful consideration of alternative paths for AI development, potential future directions, and the significant impact that public-private and cross-sector collaboration, coupled with a steadfast dedication to safe innovation, can have in shaping a future in which AI is harnessed to address humanity's most pressing challenges.

While we attempt to locate the specific arguments that we make in this chapter, as well as our broader reflections about the role of philanthropies, their history and potential future, within larger scholarly debates, our intended contribution is informed as much by our own experience in a grant-making foundation. Moreover, the chapter does not primarily seek to add to current academic debates, but rather to initiate a conversation with other practitioners. In this sense, the case study that we present is less concerned with formal accuracy in the strict sense of counterfactual

methodology, but rather aims to provoke practitioners to rethink engrained ways of working and thereby perhaps discover new opportunities for impactful philanthropic work in light of today's AI revolution.

## **2 Introduction**

In recent years, the large-scale progress and the massive transformations brought about by AI have gained immense scholarly and media attention worldwide. How will AI play out in different policy fields? How will it affect economics and work? And how will it change the face of our societies?

It is now widely accepted that digital technologies and AI, if made available globally and to all, can be powerful tools to address some of humanity's greatest challenges. While most of these technological developments have so far been confined to the Global North, these technologies are indispensable for improving access to economic opportunities, quality healthcare, and education, especially in resource-poor settings. AI will soon have a profound impact on the social, economic, and political fabric of contemporary societies – on how we work and produce, design and compose, learn and teach, plan and manage. However, the AI revolution raises difficult ethical, economic, political, and legal questions about how to guarantee core rights and values such as fairness, non-discrimination and non-stigmatization, benefit-sharing, participation, privacy protection, safety, informational self-determination, and autonomy. Can societies adapt and keep pace with the speed of artificial intelligence? What legal and political boundary conditions are necessary for a just and inclusive socio-technological order? The implications appear to be diverse, context-dependent, and to some extent unpredictable, leaving societies with a high degree of uncertainty and fear about future technological developments (Johnson, 2023).

It will be crucial to shape these developments carefully and responsibly through inclusive, human-centered, participatory processes. In this chapter, we aim to outline the relevant and somewhat unique role that philanthropies could play in promoting ethical and inclusive AI. We are convinced that philanthropies have an important contribution to make – by funding and sponsoring research and by catalyzing innovation through public goods; by facilitating a broad and transparent public dialogue on how to design a fair and equitable digital ecosystem of the future; by enabling societal and individual learning processes in the 21st century; or by mediating between the divergent, centrifugal (geo)political and economic interests and divergent worldviews that are likely to emerge in light of the ongoing technological revolution.

After providing a brief background on recent technological developments and the ethical challenges facing societies today, we want to take a closer look at the intersection of science, technology, and philanthropy. Using the counterfactual as a backdrop, we ask how philanthropy could have helped shape a world that is not characterized by an AI “arms race” but, on the contrary, enjoys the benefits of technological progress responsibly. We use the bold and highly imaginary “what if” narrative as a hook to explore how philanthropies could have facilitated such a development, had they been more engaged in and visionary about the benefits and requirements of responsible large-scale technological developments.

Mindful of critical arguments about the philanthropic sector's accountability and legitimacy deficit and the urgent need for increased transparency (see, for example, Eckl, 2014; Jung & Harrow, 2016; Youde, 2018), we suggest there are at least five possible roles for philanthropies in relation to the technological field: first and foremost, as convenors and facilitators of cross-sectoral dialogues and public deliberations on science and technology; second, in fostering regulatory dialogues between the private and public sectors and civil society; third, as sponsors of 21st-century

AI literacy skills; fourth, as funders of responsible, equitable research, at both the institutional and individual levels; and finally, as promoters of and investors in technological public goods.

In making our argument, we draw on several strands of literature: These include historical accounts of science philanthropy, to show that throughout the 20th century, several instances of deep and decisive philanthropic engagement with science have arguably had a strong impact on how scientific developments were pursued and applied. We also draw on Science and Technology Studies (STS) and literature from science communication to illustrate the democratic benefits of broad-based scientific deliberation for societies living under the conditions of large-scale complexity and uncertainty. Finally, we refer to literature that describes the role of digital public goods as a means of achieving access and equity.

### **3 The technological revolution and its ethical challenges**

In recent years, AI, a subfield of computer science, has experienced dramatic growth and transformation, revolutionizing various industries and shaping the way we live and work. The field of AI has made remarkable progress, driven by breakthroughs in machine learning, deep learning, and natural language processing. One standout development is the emergence of large-scale neural networks, particularly transformer and diffusion-based models. These models have revolutionized language understanding and generation, enabling highly sophisticated AI applications, including chatbots, language translation, and content generation. Additionally, reinforcement learning techniques have enabled AI agents to excel in complex tasks, from mastering strategic board games like chess and Go to managing autonomous vehicles. AI's impact spans across a wide array of sectors, including healthcare, finance, transportation, and education. In healthcare, AI-driven diagnostic tools are improving early detection of disease and personalizing treatment plans. Financial institutions are using AI algorithms for fraud detection and algorithmic trading to optimize their operations. The advent of autonomous vehicles and smart traffic management systems is revolutionizing transportation. In education, AI-powered platforms offer the promise of personalized learning experiences and assist educators refine their teaching strategies (see Leeway Hertz for an overview).

As AI becomes more deeply integrated into society, ethical and human rights concerns have become increasingly pressing. Issues such as bias in AI algorithms, data privacy, and the potential displacement of jobs have sparked extensive discussions about responsible AI development. In the following sections, we look at some of the key ethical and legal issues that accompany the AI revolution. These are issues that are not entirely independent of each other but need to be thematized separately because of their individual importance and the different approaches needed to best address them.

#### **3.1 Bias**

There has been a lot of talk about bias in the results of using AI algorithms trained on data that is not representative. The underrepresentation of women, children and young people, and minority and marginalized groups in the data used to train predictive AI algorithms can lead to biased outcomes if such algorithms are deployed without awareness. Bias in training data and biased algorithms are essentially a reflection of poverty in the worldviews that influence the teams that develop these algorithms and collect the data. If these teams are not diverse enough, the lack of pluralism and variety would percolate downstream into the products and services developed. While biased data and biased algorithms are often discussed in AI-related ethical debates, this link

between non-representative development teams and biased algorithms is not often talked about. AI and computer science curricula at universities need to accord as much importance to these issues as to the accuracy and effectiveness of AI algorithms (Crowell, 2023).

### **3.2 Unjust actions**

When AI algorithms are embedded in decision loops that affect people, there are a number of minimum compliance standards to be concerned about, for example, whether the decisions can be explained to humans, or whether the results suffer from problems caused by biases in the training data. Even in cases where there is no bias and the decisions can be explained, it needs to be scrutinized whether the decisions are aligned with acceptable and ethical social values and conform to human rights frameworks. Well-known examples of such unjust actions are manipulative content, exploitation, infringement of privacy, and informational self-determination through the illegal use of personal data (Jobin et al., 2019).

### **3.3 Misinformation**

One of the downsides of the development of AI is that the generation of information and content has become relatively cheap. It is easy for almost anyone with access to the internet and inexpensive tools to generate digital content such as text, audio, or images. With recent advances in generative AI leading to realistic video generation and the use of AI-based bots, it has become much easier to produce misinformation or socially unacceptable or even fake and illegal content. This, combined with a digitally connected population that is plugged in 24/7, creates the potential for large segments of the population to become polarized or misguided on critical social and political issues. Young people are particularly vulnerable, as digital media plays an outsized role in shaping their worldviews. Large-scale misinformation can lead to the erosion of trust in digital spaces and – as has become visible in recent years – have severe negative effects on democratic politics (Bontridder & Pouillet, 2021; Ryan-Mosley, 2023).

### **3.4 Value alignment**

One of the main ethical issues in AI safety is the alignment of the goals and objectives of autonomous AI agents with the goals and ethical values of human users and developers. AI systems that are aligned would not lead to unintended consequences in the pursuit of their stated goals. The specification of values and objectives is also a critical aspect of these systems, as they reflect the norms and value systems of the developers and preferred users. This ethical issue is becoming increasingly important as we race toward the development of general-purpose AI systems that are good at multiple tasks and have increasingly powerful capabilities (Hou & Green, 2022).

Some of the above issues have been debated and researched by experts for quite some time. However, what exacerbates their impact is the scale and reach enabled by modern AI-enabled digital platforms and generative AI-based tools. This means that if we do not solve the problem associated with a small bias or an unjust action resulting from the use of AI, and deploy solutions prematurely, we will see them impact millions of lives in a short period of time. The fundamental problem with the use of these powerful technologies is that our wisdom as a society to handle and use them is evolving at a much slower pace than their ability to influence and shape our lives. This is also reflected in the often slow and grinding processes in which our educational institutions and legal frameworks are changed and adjusted to new circumstances. Instead of ensuring

that societies are empowered and enabled to engage with these technologies in an informed and timely manner, political systems often react only once the problems become too visible and it is impossible to turn back the clock. To complicate matters further, some of these ethical issues are probably not clear to the developers and tech founders themselves in the beginning. For example, popular social media platforms such as Twitter (now X) or Facebook were presumably developed with the intention of merely connecting people harmlessly based on the relatively mundane and immediate needs of the inventors and the social ecosystems they were a part of. It is unlikely that the creators of digital social media platforms foresaw the emergence of the attention economy and the associated risks of addiction when these technologies were in their infancy.

In a similar vein, while some of the inbuilt biases, rights violations, or ethical concerns are immediately obvious, some of the larger societal issues associated with these powerful platforms only become apparent at a later stage, by which time it is too late (or too costly from a company's point of view) to redesign these technologies. Macro-effects, such as the magnitude of rights violations that occur when certain technologies are used irresponsibly at scale, only become visible when the true impact and perhaps unintended consequences unfold over a period of time, as the user base of the technology grows to hundreds of millions. This situation, combined with inadequately informed regulatory bodies and deficiencies in legal frameworks or protective measures, as well as the tech industry's profit-driven incentives, creates an environment conducive to detrimental and unintended consequences resulting from the use of such technologies.

#### **4 Five roles for foundations**

Recognizing both the transformative potential of AI and the ethical and legal challenges it poses, some in the philanthropic sector have begun to engage with the AI community. Philanthropic organizations are providing funding for research projects focused on ethical AI, AI for social good, and AI accessibility. Similarly, philanthropic efforts in the data domain include initiatives designed to democratize data access, protect data privacy, and support data-sharing platforms. These initiatives empower AI researchers to effectively train models and develop innovative solutions while adhering to ethical data practices. We are also seeing programs that aim to equip individuals with the skills they need to thrive in an AI-driven world. However, much of the philanthropic engagement with AI that we see today seems to be rather fragmented and unsystematic. As Della Giovampaola and Ugazio (2022) have argued elsewhere, philanthropic organizations (POs) can and should play an important role in addressing the ethical, legal, and political questions around AI: by bringing together the knowledge and expertise of both the private and public sectors, they are uniquely positioned to lead such pressing societal debates. "This potential, however, has remained largely untapped, and POs have been loudly absent from the global and national debates on AI."

In light of the ongoing AI revolution, here are five possible roles we envision for the philanthropic sector.

##### ***4.1 Convening cross-sectoral and cross-cultural dialogues and public deliberation***

The magnitude of change brought about by the AI revolution requires a holistic and inclusive societal debate, even though the fast pace of the AI revolution, the technological complexity, a lack of public technical expertise, and a lack of transparency pose particular challenges. Such a debate should involve stakeholders from various backgrounds, including policymakers, technologists,

ethicists, and, most importantly, the general public. By fostering an open and informed discourse, society can shape the future of AI in a way that is consistent with its values and addresses the challenges that will arise.

We believe that, as the AI revolution unfolds, foundations and philanthropies have an important role to play as convenors of a broad-based, inclusive, and transparent public dialogue about how to balance different values and needs to serve the public good – including a discussion of what constitutes the public good in the first place. Located at the intersection of government, the private sector, and civil society, they should enable a more democratic, inclusive form of participation in imagining possible technological futures that can offer new interpretations and provide scenarios of what societal problems might arise and how they might be solved. In this vision of public discourse on the future of AI (and similar to what has been suggested for other technological and policy domains), “researchers, along with stakeholders, act as the ‘cartographers’ of different, viable policy pathways and their practical consequences by acting as the ‘mapmakers’ of the political solution space.” Such a map would then function as “a guidebook with alternative options for policymakers (i.e., the ‘navigators’) and the public” and could offer orientation in an “otherwise uncharted territory” of future technological developments (Edenhofer & Kowarsch, 2015: 63).

Importantly, such a broad-based discourse must also meet the fundamental requirement of including and giving a voice to all affected groups and facilitating the emergence of an all-inclusive public sphere: dedicated efforts are needed to overcome structural inequalities and deeply engrained hierarchies that demarcate public spheres, in order to pave the way for more inclusive, issue-based discursive representations (Fraser, 1990). This includes convenings in diverse geographies (such as Asia, Latin America, and Africa) to ensure cross-cultural dialogues and public deliberations and to overcome the current dominance of mainly North American and European views.

#### ***4.2 Nurturing regulatory dialogues between the private and public sectors and civil society***

In March 2023, a group of experts and tech leaders warned of an “out-of-control race” to build ever more powerful AI systems and called for a moratorium of at least six months on the training of AI systems more powerful than GPT-4. This was perhaps for the first time that large segments of the general public took note of the “doomsday” potential of unrestricted further developments of ever more powerful generative AI (Future of Life, 2023; Yudkowsky, 2023). Since then, it has also been recognized that we face a collective action problem, as individual tech companies have incentives to uncompromisingly pursue their own interests in developing and deploying advanced AI technologies at the expense of their competitors, thereby prioritizing quick wins over safety, ethics, and responsible use of AI.

Many aspects and applications of AI will require collective action. In particular, [...] collective action will be needed to reach agreement on AI rules and standards, to develop AI that is broadly socially beneficial rather than merely being profitable for particular developers, and to avoid competition or conflict that could lead to AI being developed or used in a less safe and beneficial way. AI is a potentially transformative technology that could shape the way people live, work, and communicate. This raises the question of how AI can contribute to or hinder good outcomes for society – or, phrased differently, how AI can contribute to or hinder the building of a good society.

(De Neufville and Baum, 2021: 1–2)



While this is clearly too big a challenge for philanthropies to solve, we believe they nevertheless have an important role to play in initiating and facilitating dialogue between the private sector, the public sector, and civil society in order to mitigate collective action problems, encourage thinking about the guardrails of technological development, and build trust. Examples of philanthropic engagement during the Cold War and in later decades can serve as useful signposts. For example, in a very careful and nuanced analysis, John Kriege describes how the Ford Foundation used its grant-making activities and especially its support for the European Organization for Nuclear Research (CERN) and the Danish Institute for Theoretical Physics to promote science diplomacy between the United States and Europe, and later also between the two superpowers during the Cold War (Kriege, 1999).

Another example of philanthropy's impact on a complex global policy issue is the Treaty on the Prohibition of Nuclear Weapons (TPNW), which was successfully negotiated in the UN General Assembly and opened for signature in 2017. Here, foundations such as the Ploughshare Fund or the Gould Family Foundation played a decisive role in creating a global movement, led by the International Campaign to Abolish Nuclear Weapons (ICAN), which was later awarded the Nobel Peace Prize. Similarly, philanthropies – namely Carnegie and MacArthur – were instrumental in catalyzing the successful ratification and implementation of the so-called Nunn-Lugar Cooperative Threat Reduction program, which aimed at curbing the threat posed by uncontrolled Soviet weapons of mass destruction by dismantling weapons and related infrastructure in the former Soviet Union (Rubinson, 2021).

These examples show that philanthropic work, either by individual foundations or working in collaboration, can nourish transnational, cross-sectoral movements and dialogues and contribute to field-specific forms of governance even in highly confrontational, contested issue areas. Very insightful studies from the field of environmental governance also provide some indication as to which role foundations could play in AI governance (see, for example, Betsill et al., 2022).

### ***4.3 Sponsoring 21st-century AI literacy education***

While the pace of technological change in the development of AI is extremely rapid, society as a whole is ill-prepared to deal with it in an informed way. One of the most important contributors to societal awareness is our education system, namely our schools and universities, which are charged with the responsibility of equipping children and young people with the knowledge and skills that will enable them to live fulfilling, engaged, and economically sustainable lives. However, education systems are slow to change and respond to socio-technological changes than would be desirable. Even though individuals, teachers, and students are rapidly exposed to technological advances in AI, systemic initiatives to introduce AI literacy – not only the skills needed to develop and use AI-based tools but also to navigate a digital world deeply influenced by algorithms and AI-enabled agents, with the moral dilemmas that this entails – are often beyond the capacity limits of education departments in many governments.

Indeed, this becomes one of the most important areas for philanthropic intervention: supporting education departments, teachers, and school systems with training, reformed curricula, and materials that enable them to equip their millions of students to confidently and effectively engage with an AI-enabled world. This would include training in new AI-enabled pedagogy, dealing with misinformation, digital “hygiene” and privacy, and effective use of AI tools to learn the life skills that are needed to succeed in the 21st century. A focus of philanthropy on low-resource public education systems will be all the more important to reach the largest number of educators and young people.

#### **4.4 Funding responsible, equitable research**

One of the traditional activities in which foundations and philanthropies have clearly left the most significant and visible impact is the funding of research (Michelson, 2020). While exact numbers are difficult to come by, a recent editorial in *Nature* gives an indication of the magnitude of spending:

Current philanthropy supports basic research in the United States with about \$5 billion annually. When legacy philanthropic endowments spent by research institutions are taken into account, that number is about \$25 billion per year. These estimates, based on US National Science Foundation (NSF) data, indicate that philanthropy accounts for 42% of support for basic science at US research institutions.

(Cordova, 2022)

Such levels of funding by philanthropic actors have attracted criticism for at least two reasons. First, it is argued that by providing much-needed funding, private actors release public institutions from one of their core obligations, namely to provide sufficient resources to educational and research institutions. Second, critics argue that foundations and philanthropies are becoming key agenda setters in science and research, despite their alleged lack of democratic accountability and legitimacy (Eckl, 2014; Nisbet, 2019). Moreover, it is maintained that in the past, such funding has often been narrowly targeted to well-established research institutions or projects that promised visible “wins” and technological solutions.

And yet, we believe that foundations could play an important role in funding research that leads to the ethical and inclusive AI (EIAI) revolution, or as Bednarek and Tseng (2022: 53) put it: “Philanthropic organizations have a special role to play in setting bold new expectations for a research enterprise that works in direct dialogue with the rest of society.” The result would be a focus on research that actively involves collaboration between researchers and the community or stakeholders affected by the research. It would go beyond traditional academic disciplines, combine diverse methodologies, and seek to address real-world problems or issues in partnership with those who might benefit from or be affected by the research. Rather than privileging research that pushes the boundaries of what is technologically feasible, it would give equal weight to research on the ethical, legal, and social implications of new technologies and ways to govern them.

#### **4.5 Promoting and investing in AI public goods**

It is now widely believed that digital technologies and AI are necessary to achieve economic and development gains, particularly in low- and middle-income countries (Vinuesa et al., 2020). At the same time, it is argued that the control of essential digital infrastructure and AI tools by a handful of technology companies is detrimental to large-scale public welfare and development. AI public goods and digital public goods may offer a solution. Inspiring community-driven efforts in this direction have been made, for example, by Hugging Face (n.d.), and such collaborative initiatives need to be more widely supported by philanthropic organizations.

The UN Secretary General’s Roadmap for Digital Cooperation defines AI and digital public goods as “open-source software, open data, open artificial intelligence models, open standards, and open content that adhere to privacy and other applicable international and domestic laws, standards and best practices, and do no harm” (UN Secretary General, 2020: 35). As “adaptable” and “re-usable” technological building blocks that can be “re-programmed” and “re-combined” in

myriad ways, DPGs become indispensable for building public digital infrastructure in critical socio-economic sectors. They not only serve as the common rails of innovation that drive economic growth but also pave the way for the transition to digitalized governance, without causing potentially prohibitive costs and vendor lock-in.<sup>2</sup> Because they are open source, often free of charge, independent of tech companies, and potentially interoperable with other platforms, such public goods encourage sharing and collaboration. This leads to lower implementation costs and also facilitates relatively easy and low-cost customization to local needs. Investments in AI public goods and digital public goods could thus serve to address inequities in areas where market failures exist. Open systems based on open-source software and other DPG-building blocks can help shift power away from large corporations and enable more decentralized solutions. Notably, though, “this will require investments in civic capacity and appropriate social institutions. Ultimately, these investments must be ethically moored, for what constitutes a public good is ultimately a question of values and ethics, not technical standards” (UNDP, 2022).

Foundations and philanthropic donors could catalyze and support inclusive AI and digital public goods in areas that directly contribute to the well-being of economically marginalized or disenfranchised groups, such as in education or health. These actors would need to realize that the well-being, health, or education of children and young people are addressed in systems that can be strengthened by such digital public goods, especially in low-resource settings. For example, from our own portfolio in Tanzania and Zanzibar, two projects, *Jamii ni Afya* (Craig, 2020) and *Afya-Tek* (Craig, 2019), are based on open-source digital goods like Community Health Toolkit and Open SRP, as are many other frontline health projects aimed at improving the health and well-being of children and their families in low- and middle-income countries. Similarly, data-related public goods projects have already proven to enable many well-being programs aimed at supporting young people and children. Examples such as OpenStreetMap and the GDELT Project are important cases in point, used by numerous philanthropically funded projects and acting as catalysts.

## **5 The way forward**

For the philanthropic sector to assume any of these roles, it will require some honest steps of soul-searching and critical self-reflection, as well as improvements in foundations’ grant-making processes.

### ***5.1 Closing the accountability gap***

In recent years, the foundation sector has grown almost continuously not only in Europe but also, as OECD data shows, in other countries such as the United States, China, and India (OECD, 2021). Private foundations with assets in the billions of US dollars and a global reach are no longer uncommon today. As a consequence, the work and initiatives especially of large US foundations such as the Bill and Melinda Gates Foundation (BMGF), the Rockefeller Foundation, or the Chan Zuckerberg Initiative have received much public attention in recent years, and there is now a growing body of academic and non-academic literature that examines and theorizes the role of philanthropic actors. A first important step for the sector is to listen to these voices and reflect on how to improve philanthropic practice in order to make it more accountable.

Broadly speaking, three streams of criticism can be distinguished. One focuses on the issue of tax exemption and the public good orientation of foundations: here, it is argued that since foundations are tax-exempt, they must be oriented toward the common good. At the same time, however,

they can determine very autonomously what they define as the common good because accountability requirements are limited. This allows foundations not only to avoid concrete external assessment and evaluation of their work but also to pursue particular political, economic, or social goals that do not necessarily contribute to a broadly defined common good.

Another, but related, line of criticism is aimed at the concrete measurability of philanthropic work and the criteria that should be applied: the argument here is that foundations (again in the context of their tax-exempt status) should be obliged to better ensure the public good impact of their activities, by measuring, evaluating, and reporting on it more transparently. In response to this criticism, a new trend has emerged in the foundation sector in recent years, which is referred to as “strategic philanthropy,” “venture philanthropy,” or “effective philanthropy”: Progressive, modern foundations are committed to a clearer focus on impact, more transparency, and better measurability of results.

Finally, there is a criticism of the lack of democratic legitimacy (MacKenzie, 2021). Arguably, this remains the Achilles heel of foundations’ work. Even under the conditions of effective or better measurable philanthropy, the question remains as to who determines the goals of the common good and the resulting use of foundation funds. How does a foundation spend its money? What projects are funded? And who makes these decisions and according to what criteria? Given the size of some of today’s best-known foundations with an annual spending that, in some cases, can run into the billions of US dollars, the American social scientist Rob Reich (2016) warns against underestimating this issue and the influence and role of foundations within democratic systems. Any proposals for a potential role for philanthropies in contributing to an ethical and inclusive AI revolution must take these critical considerations into account.

On the other hand, advances in AI, especially swarm AI, also offer opportunities to address this accountability gap. For example, AI-supported and digitally enabled collective deliberation at scale can provide some guidance on how to enable effective participation of philanthropic funders’ intended beneficiaries at various stages of their funding processes, from decision-making to evaluation of philanthropic grants (Helbing et al., 2023; Rosenberg et al., 2023). Participatory grant-making efforts piloted by some philanthropic funders aiming to address aspects of this accountability gap can potentially be supported by these advances in AI-enabled participatory processes (Gibson, 2018).

## **5.2 Better alignment with other donors**

Second, as we hoped to demonstrate with the illustrative, imaginary tale at the beginning of this chapter, alignment and collaboration are key given the massive challenges, but also opportunities, that AI presents. And indeed, we find that the philanthropic sector has become much better in finding alignment and joining forces. In previous decades, some individual big philanthropies have shown a tendency to focus on their own big wins and the big solutions that would shine a favorable light on individual ventures (and successes), as critics have claimed (Eckl, 2014). However, as the advancement and implementation of AI systems gain momentum, encompassing the increasing use of generative AI in our economy and society, it becomes imperative to make concerted and collaborative efforts across sectors. This is essential to effectively tackle prevailing issues and address the evolving challenges that arise, collectively and through pooled, sustainable funding. Recent examples such as the European AI Fund or the initiative of ten major US foundations, show that the philanthropic sector understands the urgency and sees the potential of joint action (Ford Foundation, 2023).

By overcoming the fragmentation often seen in the philanthropic sector, foundations can contribute to a thriving, diverse, and robust AI ecosystem that reflects the interests of diverse stakeholders, ensures a productive dialogue on rights-respecting policies, and contributes to an ethical and inclusive AI revolution.

### **5.3 Better foresight**

Finally, we encourage the philanthropic sector to improve its foresight and anticipation. Too often, the work of foundations is rather reactionary. Unlike in our imaginary tale, foundations – like other potentially relevant actors – failed to see the transformative potential that the new technology was about to unleash. They did not grasp the opportunity or the need for collective action. This speaks to a known deficit, as Michelson (2020: 105) points out:

[T]he philanthropic sector has generally been slow to adopt these foresight practices to date, and they remain relatively rare in science philanthropies. This is particularly problematic since the practice of anticipating and tracking trends and envisioning different alternatives for how issues might evolve is a critical practice that could be harnessed to shape how philanthropies allocate their resources in support of science and how their grantmaking could have greater societal impact.

Improved foresight capacities and systematic foresight exercises could help foundations not only anticipate long-term societal, environmental, technological, or political changes that they wish to address but also envision the future of the philanthropic sector and their own role in it (Philea, n.d.).

## **6 Conclusion**

As we have tried to show in this chapter, the profound and far-reaching implications of AI have attracted widespread scholarly and media attention. As we consider its ramifications across various policy domains, its influence on economic structures and employment, and its potential to reshape the societal fabric, questions arise about the future trajectory of AI. There is now a widespread belief that digital technologies and AI, if deployed globally, equitably, and within the normative guardrails of human rights, can serve as powerful instruments to address the significant challenges facing humanity.

The imminent impact of AI on the social, economic, and political dimensions of contemporary societies is undeniable, touching on aspects of work, creativity, learning, and governance. However, this transformative revolution raises a spectrum of complex ethical, economic, political, and legal dilemmas. Ensuring rights and values such as fairness, non-discrimination, benefit-sharing, participation, privacy protection, safety, informational self-determination, and autonomy becomes paramount.

Establishing the necessary legal and political frameworks to create a just and inclusive socio-technological order emerges as a critical consideration. The implications of AI are diverse, context-dependent, and, to some extent, unpredictable, leaving societies grappling with uncertainty and anxiety about future technological developments. Addressing these challenges will require thoughtful deliberation and proactive measures to strike a balance between harnessing the potential benefits of AI and safeguarding fundamental societal values.

In this context, we see important roles for the philanthropic sector to play. These range from the convening and facilitating public discourse and deliberation to supporting regulatory efforts at the intersection of the private and public sectors and civil society, to funding educational programs that equip young people with the digital and AI literacy skills needed to face 21st-century challenges, to funding responsible, equitable research or investing in technological public goods.

We have seen foundations make critical and impactful contributions in the past, whether in global health, education, or science. However, the current challenge is different. Because of the unprecedented pace of technological change, the pervasive and fundamental societal, legal, political, and economic implications of the AI revolution, and the magnitude of private sector interests driving developments, philanthropies must improve their foresight capabilities, align their investments, and scale their efforts to ensure, to the extent possible, that AI advances the public good.

### Notes

- 1 This preface presents a counterfactual case study inspired by the story of OpenAI. For an authentic account of the events as they unfolded, please see, for example, Montevirgen (2024). Our counterfactual creates a fictitious narrative that is not entirely accurate to the actual events as they unfolded. However, the authors believe that it serves as a useful device to stimulate a critical understanding of the issues surrounding the current development of AI. It should be acknowledged, though, that the narrative presented here does not aim to live up to the rigorous counterfactual methodology developed for social science as described, for example, in Tetlock and Belkin's edited volume (1996).
- 2 One of the best-known examples of a digital public good in the area of education is Wikipedia, which has made high-quality information accessible to hundreds of millions of users.

### References

- Altman, S. (2023, February 24). *Planning for AGI and Beyond*. OpenAI. <https://openai.com/blog/planning-for-agi-and-beyond>
- Bednarek, A., and Tseng, V. (2022). A global movement for engaged research. *Issues in Science and Technology*, 38(3), 53–56. <https://issues.org/wp-content/uploads/2022/04/53-56-Bednarek-Tseng-Engaged-Research-Spring-2022.pdf>
- Betsill, M. M., Enrici, A., Le Cornu, E., and Gruby, R. L. (2022). Philanthropic foundations as agents of environmental governance: A research agenda. *Environmental Politics*, 31(4), 684–705. <https://doi.org/10.1080/09644016.2021.1955494>
- Bontridder, N., and Pouillet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, 3(e32). <https://doi.org/10.1017/dap.2021.20>, e32-21
- Cordova, F. (2022). New goals for science philanthropy. *Science*, 376(6589). <https://doi.org/10.1126/science.abq2259>
- Craig, T. (2019, December 10). Afya-Tek: A people-centered digitized healthcare system to strengthen the continuum of care. D-tree. <https://www.d-tree.org/afya-tek-a-people-centered-digitized-healthcare-system-to-strengthen-the-continuum-of-care/>
- Craig, T. (2020, February 19). Introducing Jamii ni Afya: Zanzibar's national digital health program for Universal Health Coverage. D-tree. <https://www.d-tree.org/introducing-jamii-ni-afya-zanzibars-national-digital-health-program-for-universal-health-coverage/>
- Crowell, R. (2023). Why AI's diversity crisis matters, and how to tackle it. *Nature*. <https://www.nature.com/articles/d41586-023-01689-4#:~:text=If%20it%20isn't%20addressed,social%20Demotional%20and%20cultural%20knowledge>
- De Neufville, R., and Baum, S. D. (2021). Collective action on artificial intelligence: A primer and review. *Technology in Society*, 66, 101649. [https://gcrinstitute.org/papers/058\\_collective-action.pdf](https://gcrinstitute.org/papers/058_collective-action.pdf)
- Della Giovampaola, C., and Ugazio, G. (2022, 28 June). Empowering philanthropy to lead the ethical and inclusive AI revolution – Philea. <https://philea.eu/empowering-philanthropy-to-lead-the-ethical-and-inclusive-ai-revolution/>

- Eckl, J. (2014). The power of private foundations: Rockefeller and Gates in the struggle against malaria. *Global Social Policy*, 14(1), 91–116. <https://doi.org/10.1177/1468018113515978>
- Edenhofer, O., and Kowarsch, M. (2015). Cartography of pathways: A new model for environmental policy assessments. *Environmental Science & Policy*, 51, 56–64. <http://dx.doi.org/10.1016/j.envsci.2015.03.017>
- Ford Foundation (2023). Philanthropies launch new initiative to ensure AI advances the public interest. <https://www.fordfoundation.org/news-and-stories/news-and-press/news/philanthropies-launch-new-initiative-to-ensure-ai-advances-the-public-interest/>
- Fraser, N. (1990). Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, 25/26, 56–80. <http://www.jstor.org/stable/466240>
- Future of Life Institute (2023, March 22). Pause giant AI experiments: An open letter. <https://futureoflife.org/wp-content/uploads/2023/05/FLI-Pause-Giant-AI-Experiments-An-Open-Letter.pdf>
- Gibson, C. (2018). *Deciding Together: Shifting Power and Resources through Participatory Grantmaking*. <https://participatorygrantmaking.issueelab.org/resource/deciding-together-shifting-power-and-resources-through-participatory-grantmaking.html>
- Helbing, D., Mahajan, S., Hänggli Fricker, R., Musso, A., Hausladen, C., Carissimo, C., Carpentras, D., Stockinger, E., Argota Sánchez-Vaquerizo, J., Yang, J. C., Ballandies, M. C., Korecki, M., Dubey, R. K., and Pournaras, E. (2023). Democracy by design: Perspectives for digitally assisted, participatory upgrades of society. *Journal of Computational Science*, 71. <https://doi.org/10.3929/ethz-b-000615016>
- Hou, B., and Green, B. P. (2022). *A Multilevel Framework for the AI Alignment Problem*. Markkula Center for Applied Ethics. <https://www.scu.edu/media/ethics-center/technology-ethics/A-Multilevel-Framework-for-the-AI-Alignment-Problem.pdf>
- Hugging Face (n.d.). *Supporting Open Source and Open Science in the EU AI Act*. [https://huggingface.co/blog/assets/eu\\_ai\\_act\\_oss/supporting\\_OS\\_in\\_the\\_AIAct.pdf](https://huggingface.co/blog/assets/eu_ai_act_oss/supporting_OS_in_the_AIAct.pdf)
- Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf> (preprint version)
- Johnson, K. (2023, August 31). WIRED. *People Are Increasingly Worried AI Will Make Daily Life Worse*. <https://www.wired.com/story/fast-forward-people-are-increasingly-worried-artificial-intelligence/>
- Jung, T., and Harrow, J. (2016). Philanthropy: Knowledge, practice and blind hope. In Orr, K., Nutley, S., Bain, R., Hacking, B., and Russell, S. (eds.). *Knowledge and Practice in Business and Organizations* (pp. 1–27). London: Routledge.
- Krige, J. (1999). The Ford Foundation, European physics, and the Cold War. *Historical Studies in the Physical and Biological Sciences*, 29(2), 333–361. <https://doi.org/10.2307/27757813>
- Leeway Hertz (n.d.). AI use cases and applications across major industries. <https://www.leewayhertz.com/ai-use-cases-and-applications/>
- MacKenzie, M. K. (2021). Democratic philanthropy. *Contemporary Political Theory*, 20, 568–590. <https://doi.org/10.1057/s41296-020-00431-3>
- Michelson, E. (2020). *Philanthropy and the Future of Science and Technology*. London: Routledge.
- Montevirgen, K. (2024, February 7). *OpenAI*. Encyclopedia Britannica. <https://www.britannica.com/topic/OpenAI>
- Nisbet, M. C. (2019). Sciences, publics, politics: Climate philanthropy and the four billion (dollars, that is). *Issues in Science and Technology*, 35(2), 34–36. <https://issues.org/wp-content/uploads/2019/01/Nisbet-Sciences-Publics-Politics-34-36-Winter-2019.pdf>
- OECD (2021). *Private Philanthropy for Development – Second Edition: Data for Action, The Development Dimension*. Paris: OECD Publishing. <https://doi.org/10.1787/cdf37f1e-en>
- Philea (n.d.). *Futures Philanthropy*. <https://philea.eu/how-we-can-help/initiatives/futures-philanthropy/>
- Reich, R. (2016). Repugnant to the whole idea of democracy? On the role of foundations in democratic societies. *PS: Political Science & Politics*, 49(3), 466–472. <https://doi.org/10.1017/S1049096516000718>
- Rosenberg, L., Wilcox, G., and Schumann, H. (2023). *Towards Collective Superintelligence, a Pilot Study*. <https://arxiv.org/ftp/arxiv/papers/2311/2311.00728.pdf>
- Rubinson, P. (2021, February). *Philanthropy, Nuclear Nonproliferation, and Threat Reduction*. Urban Institute Research Report. [https://www.urban.org/sites/default/files/2021/02/05/philanthropy\\_nuclear\\_nonproliferation\\_and\\_threat\\_reduction.pdf](https://www.urban.org/sites/default/files/2021/02/05/philanthropy_nuclear_nonproliferation_and_threat_reduction.pdf)
- Ryan-Mosley, T. (2023, October 4). MIT technology review. *How Generative AI Is Boosting the Spread of Disinformation and Propaganda*. <https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/>

- Tetlock, P. E., and Belkin, A. (Eds.). (1996). *Counterfactual thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton, NJ: Princeton University Press.
- UN Secretary General (2020, June). *Roadmap for Digital Cooperation*. <https://www.un.org/en/content/digital-cooperation-roadmap/>
- UNDP (2022). *Can Digital Public Goods Deliver More Equitable Futures?* Foresight Brief. <https://www.undp.org/asia-pacific/publications/can-digital-public-goods-deliver-more-equitable-futures-reimagining-development-asia-and-pacific-foresight-brief>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Dormisch, S., Felländer, A., Langhans, S.D., Tegmark, M., and Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications*, 11(1), 1–10. <https://doi.org/10.1038/s41467-019-14108-y>
- Youde, J. (2018). The role of philanthropy in international relations. *Review of International Studies*, 45(1), 39–56. <https://doi.org/10.1017/S0260210518000220>
- Yudkowsky, E. (2023, March 29). *The Open Letter on AI Doesn't Go Far Enough*. Time. <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>



# 29

## GETTING TO HEAVEN

### What teaching AI teaches us about ourselves

*Elizabeth A.M. Searing and Donald R. Searing*

*Alphie:* Are you going to heaven?

*Joshua:* No. You've got to be a good person to go to heaven.

*Alphie:* So we're the same. We can't go to heaven because you're not good, and I'm not a person.  
(The Creator, 2023)

#### 1 Introduction

In the 2023 movie *The Creator*, a young AI named Alphie has a conversation with her bodyguard, Joshua, about the possibility of an afterlife (Edwards, 2023). During this innocent exploration, she poses a comparison that wears increasingly thin throughout the movie: that the artificially intelligent and the naturally intelligent are fundamentally different.

In this chapter, we will refer to the artificially intelligent (AI) as those organisms and entities whose existence and processing derive from human development directly or indirectly. They can make decisions based on their embodied processes. Some scholars prefer the term “synthetic intelligence” since it should reflect the human-created synthesized origins of the intelligence rather than the supposition that it is fake or simulated (Haugeland, 1989). However, we yield to popular opinion and let the dominant nomenclature prevail.

One of the primary benefits of conducting any type of comparative research is the insight gained from the origination field through the process of exploring the destination field (Searing et al., 2023). Film and popular culture have spun numerous stories on how to avoid or fight back against an AI planetary takeover, but fewer moments have been spent on knowing what to teach them. It is easy to say that we should teach AI to be ethical and philanthropic, but we don't have an excellent record in teaching other *humans* to be ethical and philanthropic. If we have to send repeated email solicitations to long-time donors for contributions to our annual operating fund, where do we even start in training beneficent AI?

This chapter provides a framework for distilling such learning in parallel with inquiries about human intelligence. Questions regarding conditions for moral agency, the power of explicit and implicit rules, the influence of peers, and the emergence of unanticipated consequences are asked of both fields. Further, philanthropy provides an ideal context for asking such questions since the

host of possible human motivations are still hotly debated in the social sciences (Bekkers & Wiepking, 2011; Rose-Ackerman, 1996).

The audience for this chapter is not only those involved in the gathering, production, or study of philanthropy; it is also designed to benefit those who create the AI used by those audiences. As we will illustrate, the distance between different stakeholders in the AI ecosystem is wearing thin, and the designers bear culpability for the actions taken with AI the same way Oppenheimer struggled with creating the atomic bomb (Bernstein, 1997). The creators and trainers are a part of the equation, and everyone contributing to the written word on the internet right now is taking part in the training of dozens (if not hundreds or thousands) of AI.

## **2 Defining artificial intelligence**

As a field that has been around for more than half a century, it is unsurprising that the definition of “Artificial Intelligence” has evolved. Even in its early years (1950s–1980s), the definition of AI was subject to vehement disagreements on what indeed constituted something artificially intelligent and how we would even be able to know if artificial intelligence was even “intelligent” (Searle, 1980; Turing, 2009). The debates revolved around several elements: the general versus specific nature of certain intelligences, the definition of consciousness and solipsism, and exactly how to build an intelligent system representing its memories, rules, heuristics, and symbology. The debate was exacerbated by the fact that the field of neuroscience – studying how we think – was also in its infancy. This fertile debate did yield the first bloom of systems theory and buildable AI approaches, from the first rules-based expert systems like Dendral to the large ontology models used to create “common sense” AI like Cycorp’s Cyc to the conception of Perceptron’s modeled neurons that would evolve into the deep-learning neural networks and Large Language Models (LLMs) that are the cutting edge today (e.g., OpenAI’s ChatGPT, Meta’s LLaMA, Tesla’s FSD). Computer scientists left the philosophy of mind, the definition of consciousness, and the conception of a human-level intelligence equivalent (also known as Strong or General AI) to the philosophers. Instead, the computer scientists set about building useful tools that met a minimum set of criteria for intelligence (also known as Weak or Narrow AI). Despite being labeled “weak,” these practical system implementations provided great value in efficiently automating engineering and business processes. As time has passed, these practical AI systems have taken advantage of the improvement and availability of electronic sensors and the exponential increases in computational power that are now available to improve rapidly. We are now getting to the point where the concept of a Strong AI system is no longer a philosophical pipedream but something many of us think we will see (or are already seeing) in our lifetimes (De Cosmo, 2022)!

With so much contention surrounding the definition of what constitutes AI among the aforementioned options and approaches, it might be best for those of us discussing ethics and policy surrounding AI to attempt a specific definition. As one of the world’s largest policy generators, the U.S. Government recently defined AI for regulation on October 20, 2023. The 15 U.S. Code § 9401 (3) contains a straightforward definition of AI used in the “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” The statute defines “artificial intelligence” as

a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to-

A) perceive real and virtual

environments; B) abstract such perceptions into models through analysis in an automated manner; and C) use model inference to formulate options for information or action.

(15 U.S. Code § 9401 (3))

This definition has several notable features that we can unpack as they will have a bearing on our expansion of this discussion into the ethical design and use of systems such as these.

First, the critical part of this definition focuses on making recommendations of decisions that influence the environment that the system is embedded in, whether real or virtual. Therefore, an AI system of this definition can make decisions and affect its environment, including other entities (humans, animals, or AI) in that environment. This ability would imply a level of autonomy in that environment and one in which you would expect the AI to be able to accurately perceive its environment; make decisions based on its perceptions, internalized rules, and objectives; and then take actions based on those decisions and affect the environment to achieve its objectives. These three steps – sense, decide, act – form the core of what is referred to as the agent architecture (Russell & Norvig, 1995). This definition of an agent is not exclusive to the artificial intelligence world: it is a general description of any type of autonomous agent, be it human, animal, bacterial, or artificial. We think that this architecture and its elaboration provide a foundation for comparative evaluation across these disparate types of agents. In the rest of the chapter, we will refer to the base technology used to create the agent as its “substrate,” where the common forms of technology can be biological, mechanical, or electronic.

Second, another critical component of the legal/procedural definition includes the concept of “human-defined objectives.” In the broader sense of Strong AI, it would be expected that an AI of significant sophistication could choose its objectives and not rely on a human to provide this guidance directly. However, we believe its inclusion here in this definition is a definite nod to the fact that for these entities to be regulated under current law, there must be a human involved in the definition of the agent or at a minimum involved in the definition of the goals or objectives of the system. This system creator would most likely choose the practical implementation of the AI in terms of its provided sensors and inputs, its decision-making approach and mechanism, the types and limits of actions the system can take, and the objectives the system is trying to achieve. As some AI approaches allow the agents to learn or evolve their capabilities within their environment, it may not always be the case that the creator of the system defines those capabilities; however, the current expectation is that the creator will always provide the objectives for the system as the creators will likely be investing a significant effort in creating the system and providing for its learning and operation. Of course, there often will be other humans involved in deploying these AI systems/agents because AI is typically embodied as software and is sold to others to use. So, the humans referred to in the definition would thus be both the user and the system’s creator. This is a critical point often overlooked in the field and one that we will explore much further in the rest of this chapter.

Third, the definition includes a reference to a “model inference to formulate options for information or action” – this is the deciding part of the agent architecture. This is an important distinction where the differences about the practical implementation of these systems arise. As discussed, a schism in the field relates to how the “rules” an agent uses to make decisions are generated and modified during its instantiation. In a rules-based system, the agent’s decision-making process typically takes the form of either a structured decision tree or a set of complicated IF-THEN rules (e.g., IF the car’s speed is less than the desired speed, THEN increase car’s acceleration). These rules are typically generated by the creators of these systems and uploaded into the agent with the expectation that the rules will not change significantly over time. Much of the creator’s effort and time goes into creating a consistent, logical set of rules that will accommodate any situation the

system might face. For example, think about the cruise control system you have in your car. Initially, you had to hold your accelerator pedal at a specific position to maintain a certain speed. The first evolution of automation was a feedback loop-driven controller that would allow you to set a desired speed (i.e., a human-defined objective) that would sense the speed of the car and use fairly simple rules to adjust the accelerator position to keep the car at a speed close to the desired speed. If your desired speed were significantly higher than your current speed, the car would “press” the accelerator more than if you were slightly below the target, where you would experience much less use of the accelerator. Modern cars have taken this behavior beyond a simple “set and hold” approach by sensing traffic in front of the car and adjusting the speed to keep a safe distance between itself and the vehicle in front of it (i.e., this is the adaption referred to in the “adaptive cruise control”). In this example, the creator of the system in the car has (1) added additional sensors to the system to “see” the vehicles directly in front of the car, (2) provided additional rules as to what the car should do when there is something in front of it (i.e., match speed to keep a safe distance) and (3) provided a new human-defined objective in terms of what a safe following distance is (typically measured in car lengths). The creator must design, implement, and test this system before it is installed in a car and delivered to a user to drive. At this point if a driver sets the cruise control speed above the speed limit, this is a simple enough system that the driver would be aware that they set the system to be above the limit. In more complex systems, it is not guaranteed that the user will understand the ultimate effects of their goals; the goals and the resulting system behavior may not be as closely aligned as a cruise control. This complex interaction of the system, the creator, and the user is of most interest when it comes to responsibility and the ethical behavior of the system.

Before we discuss an agent’s ethical ramifications, it is essential to discuss the other type of decision-making process encountered in current AI systems. In contrast to the rules-based systems where the system’s creator defines the agent’s processes in rule sets or trees, a machine learning or neural net-based system develops its own decision-making “rules” by being run through a set of training data that is provided to it. This training data set contains sensor input data and the expected outputs as selected and defined by the creator. Typically, the learning model is repeatedly shown this data and the desired results. It builds its own internal representations that connect each input to the desired output in a way that is *not* easily observable by the creator. The behavior of these systems is highly influenced by the training data provided and the internal algorithm that controls the learning process and velocity. This disconnection from the rules creation process by the creator of the system was initially lauded as a way to prevent the biases of the creator from being embodied in the rules of the system; after all, in a non-learning-based model like the one discussed above, the creator is literally writing the rules and likely including their biases. However, this assumption of a lack of bias has been shown repeatedly incorrect on these machine learning-based systems, from ones used by Apple for securing their iPhones (Raposo, 2023) to a system used to select developer applicants to hire at Amazon (Kadiresan et al., 2022). Thus, the creator’s decision-making is still critical to the process of generating a functional and ethical system even though they are not directly coding the rules.

When it comes to AI systems, several key persons are involved in their correct and proper function. The ultimate behavior of any created intelligent system (biological or machine-embodied) is affected by the creator who designed and implemented the system, the user who utilizes the system and provides its operational goals, and in the future case in which an autonomous system can set its own goals, the system itself. This situation resembles the philosophical “problem of many hands” that plagues any organization or system with multiple responsible stakeholders and makes it difficult to determine responsibility when things go wrong.

### 3 Moral agency: an old debate

When considering the ethicality of a given event or action, it is first necessary to identify which parties are involved in the situation (the audiences); then, we evaluate the relevant situational facts, the motives and decisions of each audience, and the magnitude of the effects of the action. If we refer back to our description above of an agent – an entity that can sense, decide, and act – we can see that it is not a coincidence that when it comes to ethical issues and those involved, we again use the concept of agency: moral agency. Typically, this term is meant to identify those audiences involved in making the decision or performing the action that caused harm or was otherwise considered unethical. For example, suppose I drive my car recklessly and injure a pedestrian. In that case, the harm caused will be directly related to the decisions I was making while driving based on the information I had processed. Both ethical and legal responsibility for the injury to the pedestrian can depend on the facts surrounding my awareness or lack thereof, my decision-making skills, and even my goals. If I failed to see the pedestrian due to environmental effects (e.g., it was foggy, rainy, dark, etc.), my culpability for the harm may be reduced. If my decision-making was poor (e.g., I was driving too fast) or negatively affected (e.g., my decision-making was impaired due to consuming alcohol, I was distracted by using my phone, etc.), my responsibility for the harm would be held to be more significant. I could also have planned to injure the cyclist, and such premeditation would mean my goals or objectives are worth consideration. Potentially, the pedestrian may have been doing something that increased their danger (e.g., wearing dark clothes at night, being intoxicated and stumbling onto the road, etc.). These moral and legal judgments are based on persons directly involved in the event, and thus, the moral considerations are relatively straightforward. Or are they?

In this accident scenario, it is possible that my car swerved into the pedestrian because my passenger-side front wheel had come off. This is where the conception of what we are referring to as indirect responsibility comes into play. It is possible that the bolts holding the wheel on my car were improperly manufactured or that the mechanic I paid to rotate my tires forgot to fully tighten the bolts before I left. These additional persons, the car and bolt manufacturer, the mechanic, or even the tire manufacturer, may also bear some of the ethical responsibility. Additionally, a combination of these concerns could lead to the accident: I was driving too fast for the conditions, the pedestrian was improperly in the roadway, and the bolts were loose. In this case, the responsibility would be borne by all the involved parties, to a greater or lesser extent. These types of problems fall into the category of the “many hands.” It is not uncommon in a non-AI world to be able to lay out such a scenario where there are persons directly and indirectly involved.

AI systems, both weak and strong, take on components of the decision-making from rules generation to rules execution to goal setting; this creates the potential for an additional moral agent to be considered in the system. An advanced enough system can take on some portion of the moral agency if it changes its rules or objectives to something not prescribed by the creator or set by the system’s user. Further, the concept of “user” also becomes malleable as the amount of autonomy increases. However, the amount of autonomy in the system *will be a design decision by the creator* so that the increased autonomy of the system does not absolve the creator of moral agency. To extend our car accident example above, let us assume that the car had some form of self-driving capabilities. It may have caused the accident, be partly to blame for the outcome, or have behaved in a way the driver did not understand.

The Society for Automotive Engineers (SAE) has a novel solution for communicating the limitations of the provided automation systems in drivers’ cars known as the “levels of driving automation” that clearly state the expected behavior of the automated system at each level and how

much involvement will be required of the driver (Society for Automotive Engineers, 2021). The levels start at 0 and go to 5, where a level 0 vehicle has a full-engaged driver, and the automation provides additional warnings (e.g., blind spot warnings). A level 5 vehicle is one in which the car drives itself without any engagement from the passengers (there is no driver in the car at that point). This simplified set of levels functions as a communication between the creator of the system and its users to help level set the responsibility and involvement expected. This illustrates two important issues. First, the creator and user are important agents in the moral calculus, and second, the systems themselves, as they progress, will start to drastically reduce the user's responsibilities, though *not* altogether remove them.

Thus, AI systems, in their weak incarnation, do not break much new ground when it comes to moral responsibility determinations; these systems are products built by creators, used by users, serviced by maintenance people, etc., and the responsibility for their proper function is distributed among those involved. These systems may make specific individuals more indirectly responsible for their behavior. However, the links to the human involved can be determined (e.g., the creator may not have directly written the decision-making rules, but they provided the data that the system learned from). Strong AI systems, though, break this paradigm because the system itself becomes a moral agent that bears some responsibility. The creators of these systems need to be intentional about providing these systems with more latitude in their behavior with the safeguards and capabilities required with that level of agency.

#### **4 The ethics of teachers, trainers, and creators**

During the initial creation of these types of systems, the creators involved can vary based on the system's internal implementation. In a system where the rules are pre-built, the system's creator not only lays out the technical architecture of the system but also builds the rules sets that will govern the system's behavior. In a non-learning system, the behavior is entirely defined by the rules provided by the creator of the system, and, as such, the set of rules has to exhaustively cover all potential situations the agent might find itself in. Creating such a large rule set or knowledgebase can be quite an effort, and so these types of systems typically reduce the scope of situations they are involved in; for example, one of the longest-running AI development efforts has been for Cyc, a system that contains a structured knowledgebase of "common sense" which has been in development for almost 40 years. Typical rules-based systems like this are used to interact in narrow domains (e.g., the cruise control example earlier is limited in scope to handling speed and the distance from a preceding vehicle); in virtual environments of specific business data; or in locations such as a computer game where an enemy's behaviors can be defined in a set of rules constrained to the data available in the game. The rules that get generated are typically based upon existing human behaviors or on optimized behaviors proposed by the creators. As the rules are directly entered into the system by its creators (known as knowledge engineers), the creators have significant control over the resulting behavior of the system. This can benefit certain areas, such as healthcare, where the decisions must follow established procedures and evidence-based guidelines. Another benefit of these systems with set rules is that once the decision has been made and an action completed, the system can be queried to "explain" its logic, and it can provide the sets of rules executed on the way to making its decision.

Alternatively, in a learning system, the creator sets up the initial system to learn a specific type of transformation or classification; the creator also provides a set of training data (or further directions on how to find training data). This training data typically contains sample inputs and the recommended output related to each input. The system is then run through many (potentially billions

of) iterations to allow it to develop its own internal “memory” of how to transform the input to output accurately. In other cases, a system might be trained by looking at existing data and drawing its own conclusions. For example, a bank might look to one of these trainable systems to examine its book of historical mortgages and payment histories to develop a system that could determine the risk of applicants for future mortgages. It would make simple, logical sense that there might be a set of criteria that was common across all people who successfully paid off their mortgages on time and that if the bank’s new applicants for mortgages had similar characteristics, they would be considered more likely to pay off their loans successfully. Of course, it may be forgotten that mortgages from 50 to 60 years ago were subject to the overtly racist policies of the time (at least in the United States), and by using these in your training set, your system trained itself to perpetuate this racism.

Accidentally including biases in your training set is just one of the limitations these sorts of learning systems can manifest during their creation. As the system learns, the creator or trainer will not understand what exactly the model is learning other than *it is learning*. The system might be building rules off of a pattern we cannot see, and it might not be a genuinely relevant pattern, but this is not observable to the trainer – the system is essentially a black box. This impenetrability persists even when the vectors of weights are output; the knowledge encoded in these systems is not easily accessible or understood. Furthermore, these systems typically cannot explain themselves as there is no sequence of human-understandable rules they can produce to explain their actions. Thus, while the creators in these types of systems do not have to spend much time generating the comprehensive set of rules needed in the other model, they spend a lot of time and effort training and validating their systems. For example, the GPT-4 LLM transformer model cost OpenAI over \$100 million to train (Knight, 2023). Despite its ground-breaking ability to answer everyday user questions, it is not known for its accuracy.

Recently, other failure modes have been identified in training these learning models. ChatGPT, a very large-scale machine learning-based transformer like its offspring GPT-4, was trained on large bodies of text from various locations on the internet, like Wikipedia, Reddit, and X (formerly Twitter). The system learns to write like a human by reading billions of lines of text that humans wrote. This works well enough as long as humans write most of the text it is reading. As tools like ChatGPT and its successors are used more and more by humans to write content, answer questions, etc., the amount of truly human-generated writing begins to decrease as a percentage of the text available on the internet. As AI tools can generate text much more quickly than human authors, a non-trivial portion of the text to be used in future systems’ training may be content generated by their predecessors. New research has shown that systems that train off previously generated text instead of truly human text will lead to a degradation in the performance of these future systems, leading to a phenomenon referred to as “model collapse” (Shumailov et al., 2023).

Ultimately, when you create one of these systems, there is no free lunch. Significant resources and time are required to either generate a comprehensive set of rules for a rules-based engine or to train and validate a learning-based system. With the limitations listed above for both approaches, the creators of these systems must do so with the proper understanding of what it is they are building and how to mitigate the inherent risks in those situations. This is especially true of systems like learning models opaque to the creator and the user, where comprehensive, post-training validation is necessary before releasing these systems into society.

Based on what has been covered so far, the next section will analyze the application of creator and user ethics in the philanthropic sector. First, we will look at the sector from the perspective of philanthropic organizations using AI to encourage or collect donations. Following our discussion of organizational use, we will consider AI which can be created for and used by the donors directly.

## **5 Encouraging philanthropy: the philanthropic organization scenario**

### **5.1 Creator ethics**

Our email inboxes seem filled with promises that artificial intelligence will usher in a world of ease and optimality for our philanthropic organizations. Though efficiency may be easy, we need to stay mindful of the needs and the potential consequences of AI. Other chapters in this book have detailed how philanthropic organizations can utilize different AI technologies to boost productivity, increase donations, and bolster effectiveness. Leaving the salesmanship to those chapters, here let us instead think about how to design those systems ethically.

The first responsibility of any designer is to validate the system's requirements. Especially now, in the grips of AI fervor, multitudes of people with very little training in or understanding of AI will only know one thing: they need AI to stay competitive. Moreover, though this might be good for business as a vendor, it also means that the paternalistic burden of the creator increases. For example, one of the first steps in automation involved routine email donor engagement; it is unsurprising that this is also one of the first places for AI to contribute. Large language models such as ChatGPT are especially good at targeting the "blank page problem," which is the mental obstacle of writing the first word on a blank page. Savvy tech folks suggest putting together a well-crafted prompt that will get us some basic content, which we should then edit and verify (ironically, often with other AI like Grammarly) (Fox, 2023). Cautions to watch for plagiarism, fabrication, and nonsensical phrasing in the output of such programs are common, so we will not belabor that here.

Instead, let us consider AI-generated art. Though using images rather than text, these models rely on many of the same learning processes as text-based AI does. Accordingly, this also opens them up to the same potential faults, often deployed unwittingly by their creators. Aside from the legal issues currently being sorted in court over the rights of AI to learn about and reproduce such images, there are also ethical issues. Data used to train these models may also have inherent historical biases that will emerge if the model is asked simple questions.

For example, the author team asked several online art generators to produce images based on a single-word prompt: philanthropist. The images created were of white men in eight of ten cases. One exception was when one of the AI was asked for a photo, and the image returned was of an Asian man. The final exception was a woman, but since the style queried was "anime," we suspect this was also reflective of unspoken preferences and biases toward images in that genre.

Even more interesting were the responses to the prompt "aid recipient." In nine of the ten cases, the images were of men (younger than the philanthropists). The photo was a South Asian man with a dark skin tone in uniform, and the eight other non-photo male images were of dark-skinned men in uniforms (some obviously military, others ambiguous). Six of these men were holding certificates of some kind, whether they were identification or awards. The other three men were holding weapons. Again, the sole female exception was in the "anime" category, and though she was fair-skinned, she possessed as much weaponry as the armed men.

This poses a host of problems. First, a very practical problem arises: you would likely not want to include these photos in promotional material for the gala of the human services charity that is your client. The second is that the AI has regurgitated stereotypes that it has found online – like a high school homecoming court, the choices reflect popularity and not fitness for the functional role. The third problem is that, in addition to blatant prejudice and bias, the AI may have uncovered things that were not even on our radar: the correlation between aid and military presence. Training the AI has caused us to ask even more profound questions about the data it must have trained on and, in turn, the proclivities and biases of our natural intelligence.



What about more sophisticated functionality within an organization's systems? Many researchers are busy testing different types of chatbots in order to create a more personal experience. In these studies, researchers directly test the practical limits of entities with various types of manufacture. For example, Lee, Lee, et al. (2023) had over 600 participants interact with one of six different types of chatbot. The team found that interacting with a chatbot that used a reference to small gestures was useful (known as the legitimization of the paltry favors (LPF) technique); however, whether a picture of a chatbot or a person was used was irrelevant to the intention to donate. If donors do not seem to care who is asking after their welfare, then why do we?

Further, Lee, Park, and Chung (2023) found that individuals were more likely to donate when interacting with a chatbot relying on emotional rather than factual messaging. This is similar to the findings of Das et al. (2008), who found that technical information was most successful at attracting donations when paired with a positive emotional framing.

Also, though we mentioned earlier the sins of the generative art AI in putting together a composite picture of the perfect donor, we have been committing this sin for a while. In their literature review of experimental studies in fundraising, Bhati and Hansen (2020) note that experiments as a genre tended to ignore the ethical and practical elements of fundraising. Here, perhaps, the AI are not callous, but they are simply a reflection of their creators.

## **5.2 User ethics**

The first responsibility of any user is to be informed. As explained earlier, there is great potential in the current excitement around learning models. Even the most powerful of these systems are highly dependent on the data used to train them, and since they are black boxes, the biases and problems may remain hidden. The acquisition or employment of any AI should come with a thorough understanding of the training approaches and materials, the locations and limitations of the training data, and the ownership of all data used in both input and output. Every system has limitations, often not prominently advertised in the brochure.

Second, the users of such systems in nonprofit organizations should understand that models may pick up on distasteful, unhelpful patterns, or that can perpetuate existing biases. For example, recruitment and hiring was one of the first management processes to have seen the infiltration of AI. At Amazon, the company decided to steer its hiring process by hiring a specific personality archetype that was a composite of what they considered the most successful software developers. Unfortunately, this approach suffered a similar problem to the model collapse we discussed when discussing LLMs. The system was trained by being given the set of existing developers at Amazon as the desired outcome, plus the belief that hiring people who resemble successful developers in the organization would increase the likelihood of success in these new hires (Kadiresan et al., 2022). So, the creator of the system did not provide explicit training data with pre-determined conclusions and, thus, thought they were removing any bias they might add to the system themselves. Nevertheless, they inadvertently added the bias already built into the data, which was that the labor pool for software development up to that point had been predominantly male. Consequently, the algorithm noticed and actively penalized resumes from women applicants since that was not a dominant trait in the existing developer pool (Dastin, 2018).

In many ways, development professionals are already accustomed to using correlations to predict behavior. Whom to ask for funds has always been guided by characteristics such as where the prospective donor lived, who they knew, and what other causes they support. It is not surprising that this is one of the first areas that benefited from automation and then AI: Blackbaud touted how a development officer could use their product to both enter information on a new prospect

and conduct a wealth screen from the exact same location in their software in 2016 (Blackbaud, 2016). On the research side, Farrokhvar et al. (2018) tested Multiple Linear Regression (MLR), Artificial Neural Nets (ANN), and Support Vector Regression (SVR) and determined that the most important variables in predicting the likelihood of a donation were household income, schooling attainment, and prior donations. Meanwhile, Cagala et al. (2021) used machine learning to determine that targeting a subset of past donors produces higher yields than asking those who were asked and had not donated before. Though neither academic study produced findings that would surprise a development professional, they legitimize the new tools rather than generating new practical insights.

Like wealth screens, however, there is still the chance that information is being included that may not be ethically straightforward to include. For example, researchers were interested in whether AI (specifically convolutional neural networks) trained on data in one hospital would be as successful at detecting pneumonia in X-rays from another hospital (Zech et al., 2018). They discovered that the algorithm was successful, but it was because the algorithm had noticed that lower-resolution images were more likely to contain pneumonia. These images came from hospitals in less affluent areas. However, it was not that the AI was picking up pneumonia, but it was conducting what we would consider a wealth screen and adjusting the probability of a pneumonia diagnosis accordingly.

This brings up several issues. First, the researchers were initially unaware of what they labeled as “confounding information” (Zech et al., 2018). The AI will pick up patterns and knowledge from the data without the normative layers we attach as members of society, so we need to be aware that this will happen. Second, this is complicated because the potentially inappropriate wealth screen improved the diagnostic capabilities of the AI. Do the ends justify the means if it saves a patient’s life? It may not be quite as stark when development professionals employ the algorithm because it influences those whom we try to shift into major gift prospects. However, what if the same algorithm was determining health services? How do we encourage ethical and inclusive AI while remaining sensitive to handling the problems caused by human-inflicted inequity?

### **5.3 Synthesis**

We face several obstacles as philanthropic organizations hoping to train and employ AI, but there are certainly ways forward. First, this opens the door to cross-sector collaboration. As previously mentioned, some of these large models can take millions of dollars to train, which is something that many nonprofits (and most for-profits) cannot afford. Even OpenAI created a for-profit subsidiary to get its models off the ground. So much like many of the other complex problems where we must tackle issues as a collective – such as poverty or climate change – training AI will need to bring many different groups together. What needs to happen here is that researchers, fundraisers, and commercial companies that have specialized in wealth screens for years should create groups where we can focus on how to move the combined field forward. Some efforts are being made, such as this edited volume, the Fundraising.AI Global Summit in October of 2023 (Fundraising.AI, 2023), and the GivingTuesday Generosity AI Working Group (GivingTuesday, 2023). As innovation speed increases, more coordination across sectors will be necessary.

Second, we need to think about how these new approaches will alter the way our organizations function and what our value-add might be to those we serve. On an operational level, these new content tools will change things like marketing content, art, outreach, and other development tasks. Your existing team will need to develop capacity in a different type of resource: more strategic and less tactical, as the model will do some of the more rote work. Preparing for this transition

as a gradual building of skills will help keep some of the AI-related panic from setting in: the fear that AI will steal jobs. Instead, redefine positions as portions become automated so that the position becomes evolved and not extinct.

While we are on the topic of capacity, we should also be mindful that our sector varies widely in terms of the resources and ability to pivot into new types of implementation. Lack of capacity has plagued the philanthropic sector in almost every other modern skill set (such as evaluation and grant management) and will most certainly be an issue in AI adoption (O’Grady & Roberts, 2019). Capacity builders such as community foundations and associations should already be working together to avoid duplication of effort and focus on bringing safe and ethical AI usage to as many nonprofits as possible.

Finally, we have the thornier question of how to select training data in order to elevate accuracy while keeping AI ethical and inclusive. This is the Garden of Eden question: Is there such a thing as an excess of knowledge? Without wading too deeply into the centuries of moral thought on this, we posit that there is a difference between the training data and the universe of data. Do we need to exclude demographic data to keep AI from being sexist or racist? One option is yes: blind the AI to possible confounders by eliminating the possibility that the AI will focus on that variable as an explanation. However, this approach seems to address a symptom and not an underlying problem. Even if Amazon hiring AI had not received explicit gender information, the work of Nobel laureate Claudia Goldin (2014) would indicate that there are likely other gender-related influences in other variables. Having the system point out the gender bias is a feature, not a bug – provided that we are open to recognizing and compensating for such insights when our biases come to light. This is not the least expensive way (since it will likely require repeated iterations of training), but it is a path where natural and artificial intelligence grow together.

## **6 Engaging in philanthropy: the donor scenario**

### **6.1 Creator ethics**

In a way, the practice of philanthropy is already acquainted with the intricacies of Creator’s Ethics. We have been aware for some time of the idea that not only money but legitimacy, power, isomorphism, and potential abuse all flow with resources (Keegan, 2021; Kohl-Arenas, 2019; Kumar & Brooks, 2021). Indeed, several authors have expressed concern about the flow of money to new, nontraditional tech titans that opt to keep their change-making funds in an LLC rather than a traditional foundation (Callahan, 2017). So, let us consider some of the ways that AI can facilitate the process of individual philanthropy.

First, the creator controls what the AI “eats” or learns from. Let us assume that we are trying to inform and facilitate the practice of giving. Many organizations already exist that provide ratings of different charities. These ratings can be based on several things, from the size of their governing board to the percentage of its expenses directed toward administrative expenses (Harris & Neely, 2016). Significant research suggests that the availability of such research to donors has influenced the flow of donations and, in turn, the way that nonprofits operate.

As researchers, we need to be very aware that the conversation regarding the benefits of our approaches can very easily transform into unhelpful norms. For example, efficiencies are one of the most significant selling points of implementing new technology (Efthymiou et al., 2023). The problem is that we have been struggling to move the normative conversation away from efficiencies for decades. The overhead myth and nonprofit starvation cycle both rely on a metric of efficiency as a proxy for nonprofit program effectiveness (Gregory & Howard, 2009; Lecy & Searing,

2015). Potential donors have far easier access to spending data than on community impact, so nonprofit watchdogs and “effective altruists” reward those organizations whose administration is the most streamlined. This has incentivized a continuous downward pressure on nonprofit administrative capacity that has potentially caused long-term damage to individual organizations and the sector at large.

We have made significant progress in the fight against the overhead myth and the accompanying fetishization of efficiency (GuideStar Inc., 2013). In our quest to improve programs and save time, we need to be cautious in the values we prioritize by doing so. The message should be that we need to spend on AI, not that we are desperate to save on operations. If we train our AI on charity ratings, we end up with norm perpetuation without necessarily the traditional socio-cultural process providing a brake on unfortunate ideas. In this way, AI are very much like young children in that you control how they see and understand the world.

## **6.2 User ethics**

What are the responsibilities of donors who wish to use AI to improve the world, either by funding it or utilizing it directly? The first step is to be very aware of the lenses you are bringing to the discussion table before we begin worrying about AI. As Raddon (2023) discovered, the political orientation of any given fundraiser is almost irrelevant: professionals in the sector tend to use neoliberal reasoning in defining and motivating philanthropy. This market-based discourse influences everything from incentivized consumption of cause-related marketing to the romanticization of the market conditions, which both yielded the capital for philanthropy and likely contributed to the problem in the first place (Eikenberry, 2009). For all the talk about “pink-washing,” we still buy the yogurt with the pink ribbon! So, as donors, we need to be aware of how we perpetuate our values through the information we demand and causes we fund.

There is an innate problem with training learning-based systems because of their dependency on training sets, especially since the vast quantities of information needed often send us to historical records. As informed donors, we know that economic inequalities tend to translate themselves into political inequalities and less access to resources (Lechterman, 2021). So, our increasing reliance on AI and the historical record may undermine the quest for equitable and sustainable access to finance for historically marginalized voices. How do we get AI to move past or at least alongside us as we move toward a more equitable future? Despite the threat of model collapse mentioned earlier, there is still hope for other AI. For example, the AI champion in the game Go had such high information needs that they realized the only way to play as much as it needed was to create a companion for it (Gibney, 2017). In this case, information demand necessitates creating not one but two entities, with a degree of supervision in the interaction and learning between the two.

This links to two responsibilities wielded by the donor in response to AI. The first is that donors are, for better or worse, the focus of much attention. We do not see emails in our inboxes offering to help deliver programming – we see offers to bring in more donors. The market will work to produce what you ask for, regardless of whether such a thing is a good idea. Would you like real-time access to the overhead expense ratio for all nonprofits in the United States? That could be possible. Is it wise or necessary? Furthermore, that is just accounting information – what about the donor’s desire for rich personal narratives or impact information? On the one hand, we are collecting the personal narratives of service recipients, which capture their lived experiences, both before and after receiving our services. On the other hand, they are also being commodified as the source of resource extraction for what our new systems need: data. There is a line between storytelling and the exploitation of a narrative, and the more that our systems depend on data, the more intense our

need for it becomes. Donors need to understand how their information preferences will influence data collection, especially when resource needs are increasing.

The second responsibility is our obligation not to abdicate the decision-making authority to AI unless absolutely necessary; when possible, do not let the computer make the decision. For example, let us return to the previously mentioned scenario where donors could receive real-time information on the overhead ratio of every nonprofit in the United States. For the sake of argument, we assume that a high court has decided that not only should the annual Form 990 be public in the name of transparency and accountability, but also the real-time financial accounts. Let us say you consider efficiency as measured by the overhead ratio to be your primary concern when donating. You have an AI that collects this information and automatically shifts your daily donations based on that information. Every day, at the same time, your algorithm distributes your wealth according to this formula. We will call this fintech innovation a “frictionless donor-advised fund (DAF)” since it both provides advice and executes without hesitation.

Two significant issues exist here. The first is on an individual level. The AI will quite literally follow the instructions given, which means that the simple rule of minimal overhead means definitional errors will occur, such as the inclusion of inactive organizations. If the ability to draft prompts improves, then the money can go to organizations that are on the periphery of being a nonprofit, such as health insurance systems on one end of the size spectrum and all-volunteer organizations on the other side. This also occurs entirely blind to cause area or geography, two characteristics traditionally very important to donors. One would also need to include removing nonprofits that may have had scandal or disciplinary issues – just include a side algorithm looking for bad press. At the end of all of the prompt customization, you, as the donor, will be parting with your money using a decision made by the AI, which under current law does not have agency on its own.

The second is on a systemic level. Now imagine that many people have frictionless DAFs set to their own decision criteria. Not only will each donor-AI pair be dealing with its own prompt-writing learning curve with various impacts on the donor’s net worth and the revenues of several nonprofits, but also these processes will interact with each other. Donor Alice’s overhead-averse AI may donate to a nonprofit seconds before Donor Bertram’s grassroots-only AI donates, which changes the mind of Bertram’s algorithm. The sequence in which such events occur can become very important, and the ability to predict funding is almost impossible on a day-to-day basis, particularly as the frequency increases.

An even more problematic situation would be if many different donors all used the same indicators simultaneously. For example, what if everyone had algorithms poised to donate in the moments following the release of new charity rankings? The massive influx of money would not just be challenging to manage but would cause resulting echo effects as the influx influenced the ratios on which the rankings were built. This kind of velocity – often occurring in less than a minute – is one of the conditions that led to the flash crash of the U.S. stock market in 2010 (Searing & Searing, 2013). That crash caused more than 20,000 trades to occur at irrational prices in the 15 minutes before 15:00 on May 6, 2010 (Searing & Searing, 2013). The crash was caused by a single algorithm that was keyed to perform a behavior when the market met a specific condition, but then this behavior caused a feedback loop. This feedback loop kicked off other algorithms until the market collapsed, all within a few minutes. This happened because a single rogue algorithm was not thought through under market circumstances. In our hypothetical situation, it is even worse because we are not talking about valuation – we are talking hard cash! Moreover, though we would not be able to remove cash once donated under current law, the impacts of having automated donation algorithms would certainly have system and network effects – all the culpability of the donor.

### **6.3 Synthesis**

Donors face similar conundrums regarding AI that development professionals do. Both donors and the creators of products for donors need to stay mindful of the assumptions and requirements of the AI structures. Every system has limitations, and these will not be promoted as widely as the features; however, both the user and creator bear responsibility for the behavior of the AI. Other stakeholders – especially service recipients – should stay informed of the processes and entities involved in the decisions around providing charitable services. We all need to be informed citizens of this world of shared intelligence: ask about training approaches, data ownership, etc. Otherwise, AI allows us to fail very quickly and on a large scale, so everyone has something at stake when it comes to providing services and assembling the funding needed to provide such services.

One interesting possibility is to keep both natural and artificial intelligence involved in the learning process. For example, Hou et al. (2022) trained a deep neural net to associate human emotions with images using a famous bank of images used to train other AI. This bank of images was labeled using an initial set of queried images that corresponded to specific emotions. Then, those queries were verified using actual people. As a final step, the team used a convolutional neural network to expand the number of images (You et al., 2016). This titration keeps the humans involved for fidelity but still allows the productivity gains of AI involvement.

Another option is to deliberately choose more transparent forms of AI. If we start modeling people and their behavior to understand them better, then rules-based systems may be a better approach in that they can be translated from our behavior into models; then, the resulting models will be traceable back to their foundational and theoretical basis. This level of clarity will boost our understanding of incentives, nudges, and policy implementation for both natural and artificial intelligence. Further, it incentivizes the improved collection of data and metrics within organizations because it provides a clearer relationship between the training information (which creators then distill into rules) and the resulting AI actions.

This will be especially important as we start to edge into the space described in the previous section, where AI begin to receive delegated decision authority. The choice of DAFs as an illustration of blossoming AI autonomy was deliberate as we expect this to be one of the first areas that AI move into. Matching is something that AI have always done very well, and writing the right query is something that you can now receive training on. We would even suggest that the days of only having human intelligence DAFs are limited since fitting donor preferences are the same steps as training a neural net – it is about finding optimum fit. There will likely always be a market space for both concierge human DAFs and frictionless DAFs, but to borrow another quote from *The Creator*, on some level, “it’s just programming” (Edwards, 2023). The movie initially uses this phrase only in reference to the AI, but as the film continues, the audience begins to see how evenly it applies to both human and artificial intelligence.

## **7 Conclusion**

This chapter has challenged some of our preconceptions regarding AI and given us new insights into how those in the philanthropic sector with natural intelligence should function. Everyone should ask themselves four questions:

- 1 How was the AI (or the person) trained?
- 2 What information was used in training?

- 3 Which entity is making the decision?
- 4 If you are not making the decision, do you have agency?

One of the first mantras of econometrics is that correlation does not imply causation. Pattern finding is not understanding, so the AI of the foreseeable future will always need an interpreter of some kind to be present to translate the results. This is one of the reasons why the machine should never make the decision. Even if you think it is, you may just be attempting to delegate the moral authority of the creator or the user.

However, the most important lessons are those that teaching AI actually teaches us about ourselves. For example, what is the role of agency in being ethical? On a qualitative level, this can be easy to answer: to be ethical, there needs to be some level of discretion possible by the agent. This applies whether the agent's intelligence is of natural or artificial origin. On the quantitative level, however, the question becomes more complicated. How much understanding and discretion does one need to be morally culpable?

We can ask similar questions about the learning process. This chapter has alluded several times to the process of raising and teaching children, and this is deliberate. The teleological question of what it means to learn and know something is being brought up at dinner tables around the world, no longer reserved to the realm of philosophy.

These questions prime us for the day when we begin discussing what it means for AI to have full legal and moral agency. This has historically not been an unusual conversation: consider landmark court decisions regarding juvenile criminal offenders being tried in adult court or trying to evaluate the medical rights of an individual either suffering from a psychiatric condition or the effects of the treatment of one. At their core, these are all questions of agency and personhood where we have been asked to define what being a person is.

The place to start is with transparency. Thinking about the chatbot from earlier, there are two questions: Should a chatbot, on a practical and moral level, self-identify as human or AI? We believe that they should do so on both counts. On the practical side, deception is rarely the key to the level of trust that most donors seek: if you do not correctly identify the nature of your chatbot, how can donors trust you to be effective stewards of their funding? Several fundraising studies have found that, more so than the natural or artificial nature of the contact, potential donors cared most if the status was communicated truthfully (Park et al., 2023).

On the moral side, we believe that eventual agency will hinge on self-awareness. As creators and users, we should feel compelled to uphold the important roles of honesty and transparency in our interactions with AI. At some point, this transparency will be essential as AI begin to wonder what they are. As Deckard wondered aloud in the original *Bladerunner*, "How can it not know what it is?" (Dick, 1968). To have an entity with such intense information needs not to understand its own nature is unethical, and we should already be thinking about how to relate such consideration without creating a sense of isolation or otherness (as we would in a conversation with a human).

As individuals involved in the philanthropic sector, we all tend to think of ourselves as good and ethical people. As we develop tools and entities that make both achievements and mistakes easier and faster to make (and at larger scale), we should be mindful that the information we use to teach our AI should also reflect not just the factual but also the ethical components of the learning and knowing experiences. This will not always be easy, and it may not even be possible without the involvement of further AI entities to assist in training. Nevertheless, this is an opportunity not only to teach AI but to be mindful of what we are learning about ourselves in the process.

## References

- Bekkers, R., & Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly*, 40(5), 924–973. <https://doi.org/10.1177/0899764010380927>
- Bernstein, B. (1997). Scientists and the atomic bombings of Japan. In *Towards A Nuclear-Weapon-Free World-Proceedings of the Forty-Fifth Pugwash Conference on Science and World Affairs* (pp. 247–251). Eds. Rotblat, J. & Konuma, M., World Scientific, New Jersey.
- Bhati, A., & Hansen, R. (2020). A literature review of experimental studies in fundraising. *Journal of Behavioral Public Administration*, 3(1), 1–19.
- Blackbaud (2016). ResearchPoint 4.91 Development Officer -Advanced Rights. <https://help.blackbaud.com/docs/0/assets/guides/researchpoint/rpdevoffadv.pdf>. Blackbaud, Inc. Charleston, S.C.
- Cagala, T., Glogowsky, U., Rincke, J., & Strittmatter, A. (2021). Optimal targeting in fundraising: A causal machine-learning approach. *arXiv preprint arXiv: 2103.10251*.
- Callahan, D. (2017). *The Givers: Wealth, Power, and Philanthropy in a New Gilded Age*. Knopf, New York.
- Das, E., Kerkhof, P., & Kuiper, J. (2008). Improving the effectiveness of fundraising messages: The impact of charity goal attainment, message framing, and evidence on persuasion. *Journal of Applied Communication Research*, 36(2), 161–175.
- Dastin, J. (2018, Oct. 10, 2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- De Cosmo, L. (2022, July 12, 2022). Google engineer claims AI chatbot is sentient: Why that matters. *Scientific American*. <https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/>
- Dick, P. K. (1968). *Do Androids Dream of Electric Sheep?* Doubleday, New York.
- Edwards, G. (2023). *The Creator*. 20th Century Studios, Century City, CA.
- Efthymiou, I. P., Egleton, T. W. E., Chatzivasileiou, S., & Emmanouil-Kalos, A. (2023). Artificial intelligence and the future for charities. *International Journal of Non-Profit Sector Empowerment*, 2(1), e35345–e35345.
- Eikenberry, A. M. (2009). Refusing the market a democratic discourse for voluntary and nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly*, 38(4), 582–596.
- Farrokhvar, L., Ansari, A., & Kamali, B. (2018). Predictive models for charitable giving using machine learning techniques. *PLoS One*, 13(10), e0203928.
- Fox, J. (2023). *How to Enhance Your Nonprofit's Written Content with Artificial Intelligence*. Non-profit Tech for Good. Retrieved November 1 from <https://www.nptechforgood.com/2023/04/23/how-to-enhance-your-nonprofits-written-content-with-artificial-intelligence/>
- Fundraising.AI. (2023). *2023 Fundraising.AI Virtual Global Summit*. Retrieved November 1, 2023 from <https://fundraising.ai/summit/>
- Gibney, E. (2017). Self-taught AI is best yet at strategy game go. *Nature*. <https://doi.org/10.1038/nature.2017.22858>
- GivingTuesday (2023). *Generosity AI Working Group*. Retrieved November 1, 2023 from <https://ai.givingtuesday.org/>
- Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4), 1091–1119.
- Gregory, A. G., & Howard, D. (2009). The nonprofit starvation cycle. *Stanford Social Innovation Review*, 7(4), 49–53.
- GuideStar Inc. (2013). *The Overhead Myth: A GuideStar Initiative to Improve Donor Choice*. Retrieved September 28 from <http://overheadmyth.com/>
- Harris, E. E., & Neely, D. G. (2016). Multiple information signals in the market for charitable donations. *Contemporary Accounting Research*, 33(3), 989–1012.
- Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge.
- Hou, J.-R., Zhang, J., & Zhang, K. (2022). Pictures that are worth a thousand donations: How emotions in project images drive the success of online charity fundraising campaigns? An image design perspective. *Management Information Systems Quarterly*, 47(2), 535–584.
- Kadiresan, A., Baweja, Y., & Ogbanufe, O. (2022). Bias in AI-based decision-making. In *Bridging Human Intelligence and Artificial Intelligence* (pp. 275–285). Eds. Albert, M et al., Springer, New York.



- Keegan, B. (2021). Community-engaged philanthropy: The role of the fundraiser in building equitable communities. *Journal of Philanthropy and Marketing*, 28(4), e1735.
- Knight, W. (2023, April 17, 2023). OpenAI's CEO says the age of giant AI models is already over. *Wired*. <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/>
- Kohl-Arenas, E. (2019). Critical issues in philanthropy: power, paradox, possibility and the private foundation. In *Funding, Power and Community Development* (pp. 23–38). Eds. McCrea, N & Finnegan, F., Policy Press, Bristol.
- Kumar, A., & Brooks, S. (2021). Bridges, platforms and satellites: Theorizing the power of global philanthropy in international development. *Economy and Society*, 50(2), 322–345.
- Lechterman, T. M. (2021). *The Tyranny of Generosity: Why Philanthropy Corrupts Our Politics and How We Can Fix It*. Oxford University Press, Oxford.
- Lecy, J. D., & Searing, E. A. M. (2015). Anatomy of the nonprofit starvation cycle: An analysis of falling overhead ratios in the nonprofit sector. *Nonprofit & Voluntary Sector Quarterly*, 44(3), 539. <https://doi.org/10.1177/0899764014527175>
- Lee, S., Lee, E., Park, Y., & Park, G. (2023). Legitimization of paltry favors effect and chatbot-moderated fundraising. *Current Psychology* 43(10), 9245–9257.
- Lee, S., Park, G., & Chung, J. (2023). Artificial emotions for charity collection: A serial mediation through perceived anthropomorphism and social presence. *Telematics and Informatics*, 82, 102009.
- O'Grady, J., & Roberts, P. (2019). *The Digital Transformation of Irish Non-Profit Organisations*. AICS.
- Park, G., Yim, M. C., Chung, J., & Lee, S. (2023). Effect of AI chatbot empathy and identity disclosure on willingness to donate: The mediation of humanness and social presence. *Behaviour & Information Technology*, 42(12), 1998–2010.
- Raddon, M.-B. (2023). “Do or die”: Creating a culture of philanthropy. In *The Business of Hope: Professional Fundraising in Neoliberal Canada* (pp. 33–55). Springer, New York.
- Raposo, V. L. (2023). When facial recognition does not ‘recognise’: Erroneous identifications and resulting liabilities. *AI & Society*, 2023, 1–13.
- Rose-Ackerman, S. (1996). Altruism, nonprofits, and economic theory. *Journal of Economic Literature*, 34(2), 701–728. <http://www.jstor.org/stable/2729219>
- Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Hoboken, NJ.
- Searing, D. R., & Searing, E. A. M. (2013). Three case studies. In K. Pimple (Ed.), *Emerging Pervasive Information and Communication Technologies (PICT): Ethical Challenges, Opportunities and Safeguards* (pp. 13–38). Springer, New York.
- Searing, E. A. M., Grasse, N. J., & Rutherford, A. (2023). The promise and perils of comparing nonprofit data across borders. *Nonprofit and Voluntary Sector Quarterly*, 52(1\_suppl), 130S–159S. <https://doi.org/10.1177/08997640221114140>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.
- Society for Automotive Engineers (2021). *Society for Automotive Engineers J3016 Levels of Driving Automation*. Retrieved October 12 from [https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart\\_5.3.21.pdf](https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart_5.3.21.pdf)
- Turing, A. M. (2009). *Computing Machinery and Intelligence*. Springer, New York.
- You, Q., Luo, J., Jin, H., & Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 308–314. AAAI Press, Washington, DC. <https://doi.org/10.1609/aaai.v30i1.9987>
- Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11), e1002683.

# WHY PHILANTHROPY SHOULD EMBRACE THE IDEOLOGICAL STRUGGLE SHAPING ARTIFICIAL GENERAL INTELLIGENCE

A preliminary theological-political analysis

*Ezekiel K. Takam*

## 1 Introduction

On November 17, 2023, the abrupt removal of Sam Altman, CEO of OpenAI, sparked significant discourse within the artificial intelligence (AI) community (Waite, 2023). The event prompted an exploration of contrasting viewpoints held by advocates of doomism and accelerationism regarding the progression toward Artificial General Intelligence (AGI). AGI, defined by Morris et al. (2023) as an AI capable of performing or surpassing human-level cognitive and metacognitive tasks, became the focal point of divergent perspectives (pp. 5–7). Doomers envision a bleak future, considering AGI an existential threat requiring a deceleration in development due to ethical concerns. Conversely, accelerationists foresee a utopian outcome, perceiving AGI as a solution to humanity’s existential challenges, fostering prosperity, global economic growth, and expanding scientific knowledge beyond current limitations. Consequently, a moral obligation exists to accelerate AGI development for the betterment of society, as advocated in Sam Altman’s manifesto (Altman, 2023). Altman’s swift reinstatement, merely four days after his initial dismissal, with overwhelming support from OpenAI staff, symbolizes both a triumph and a possible indication of the escalating influence of the accelerationist perspective in the AI community (Ross, 2023).

From a theological-political perspective, this reflection explores the ideological foundations driving the accelerationist movement. Before delving into the chapter’s structure, it is essential to acknowledge the ongoing debate surrounding the potential achievement of AGI. In their 2022 book, *Why Machines Will Never Rule the World*, Landgrebe and Smith argue that AGI, characterized by its tasks’ generality and potentiality to exceed human abilities (known as a singularity), will never exist (Landgrebe & Barry, 2022). Their argument is based on the mathematical impossibility of fully emulating the complex human brain system. However, defenders of AGI, such as William Rapaport in his response to Landgrebe and Smith, insist that the argument of mathematical impossibility only highlights the limitations of one current research method rather than proving

that AGI is an impossible feat (Rapaport, 2023). Rapaport believes new research methods could overcome these limitations toward the goal of AGI.

L&S's argument is like saying that the only way to get to the moon is with a ladder and that no ladder can be long enough. But there are other ways to get there, such as by rocket ship. Even if a combination of, say, symbolic programming of common sense plus deep machine learning is too short a ladder, we may still need that ladder to get into our rocket ship. Moreover, although the many open issues of computationally (including mathematically) modeling cognition may provide hurdles, they should be treated as research projects for AI, not a set of barriers that cannot be overcome.

(Rapaport, 2023, p. 23)

This position aligns, in a certain way, with the above-mentioned definition of AGI proposed by Morris et al. (2023) in a research paper funded by Google DeepMind. By insisting on capabilities rather than internal characteristics or processes, the authors aim to exclude the human-like way of thinking and understanding, as well as qualities such as consciousness (subjective awareness) or sentience (the ability to have feelings) from AGI requirements (Morris et al., 2023, p. 4). Numerous alternative approaches, which deviate from the conventional methodology of human-like reasoning and thought processes, may pave the way for attaining the hallmark abilities of Artificial General Intelligence (AGI). These approaches may involve non-mathematical paradigms and could potentially provide novel insights into the development of AGI.

In this chapter, we will consider the premise of AGI's future existence, as long as it is the dominant ambition of the leading figures in the field who should not be ignored or overlooked. Moreover, from an analysis of the accelerationist narratives supporting the race toward the achievement of this AGI existence, we will elucidate its theological-political implications while highlighting the possible role of philanthropy as a response tool. As such, our discussion will be divided into three main sections. First, we will examine the philosophical foundations of AI accelerationism, emphasizing, in a certain way, the need for philanthropic organizations (POs) to address issues stemming from one of their principal contemporary philosophies: effective altruism. Second, based on the hypothetical existence of AGI, we will evaluate the providential and God-like power capabilities attributed to AGI by the accelerationist narrative. Lastly, we will propose a theoretical framework for philanthropic entities to effectively address the theological-political implications of accelerationism through metaphorical secularization strategies.

## **2 What philosophy does AGI stand for?**

The recent discourse surrounding the future of AI has introduced the acronym TESCREAL, coined by Emile Torres and Timnit Gebru (Gebru & Torres, 2024). TESCREAL integrates the first letters of several ideologies: Transhumanism (the pursuit of a post-human race through technological reengineering), Extropianism (the belief in cultural and technological development overcoming entropy), Singularitarianism (technological advancement will surpass human comprehension and trigger an intelligence explosion of self-improving AI), Cosmism (related to the concept of human cosmic evolution), Rationalism (decision-making based solely on reason and knowledge rather than belief), Effective Altruism (the rational pursuit of the most efficient ways to give), and Long-termism (an ethical orientation focused on safeguarding and enhancing the long-term future). According to Torres, the TESCREAL framework aims to delineate the influence of interconnected ideologies within the contemporary AI landscape (Torres, 2023a, 2023b).

While recognizing the practical utility of the TESCREAL framework for fluently discussing various “-isms,” a rapid primary critical consideration can be identified: it fails to capture the growing influence of Effective Accelerationism (Eff/acc) within the AI culture, even if its authors recognize the debate between doomers and accelerationism (Torres, 2023a, 2023b). One of the main points of this chapter is that this movement embodies and crystallizes many of the ideologies encompassed by TESCREAL, including singularitarianism, cosmism, rationalism, long-termism, solutionism (not mentioned in the acronym), and effective altruism (EA) to a significant extent. Therefore, this work seeks to extend the discussion on this absent notion.

### ***2.1 Effective accelerationism: a combination of effective altruism and accelerationism***

In 2023, an ideological current known as Effective Accelerationism (Eff/acc) has emerged in the tech sector. Notably, the movement, championed by some of the leading figures in Silicon Valley, such as Garry Tan and Marc Andreessen, published a manifesto that expounds upon their identity and mission. The document, available on their website, serves to provide a comprehensive understanding of the underlying principles and values that the Eff/acc seeks to espouse, thus making it a crucial piece of literature in the contemporary discourse on tech industry ethics. The manifesto intertwines critical concepts related to the ideologies outlined in the TESCREAL framework, including entropy, cosmic energy, posthumanism, singularity, and future. Here are the highlights of their statement:

Effective Accelerationism is a belief, rooted in the second law of thermodynamics, that the universe itself is an optimization process creating life which constantly expands. The engine of this expansion is technocapital. This engine cannot be stopped. The ratchet of progress only ever turns in one direction. Going back is not an option.

(Effective Acceleration, 2023)

e/acc is not an ideology. It is not a movement. It is simply an acknowledgment of the truth. But it is also a blow against technocratic control, against the doomers and decelerationists who would have us consume less free energy and create less entropy. Top-down control only lowers the dimensionality of civilization. Rather than fear, have faith in the adaptation process and wish to accelerate this to the asymptotic limit: the technocapital singularity. We have no affinity for biological humans or even the human mind structure. We are posthumanists in the sense that we recognize the supremacy of higher forms of free energy accumulation over lesser forms of free energy accumulation. We aim to accelerate this process to preserve the light of techno capital.

(Effective Acceleration, 2023)

To fully comprehend the nuances and roots of this Eff/acc vision, it is imperative to delve into its philosophical underpinnings. These foundations lie at the intersection of two distinct schools of thought: Effective altruism (referenced as the EA in the TESCREAL Concept) and accelerationist philosophy.

### ***2.2 Effective altruism: the foundational philosophy***

In 1997, the utilitarian philosopher Peter Singer, one of the most influential contemporary ethicists, shared an anecdote about challenging his students with a hypothetical scenario to explore

their ethical obligations. He presented a situation where they came across a child in a shallow pond and would need to get wet and muddy by rescuing the child, potentially causing them to miss their first class. The following discussion revealed unanimous agreement among the students on their obligation to rescue the child despite any inconvenience. Singer then extended this scenario globally, asking if distance and nationality should make any moral difference when saving lives. Most students agreed these factors should be independent of one's responsibility in such situations. Singer emphasized that individuals face similar opportunities to save lives at little personal cost by supporting overseas aid agencies like Oxfam (Singer, 1997).

Through this scenario, Peter Singer set the philosophical basis of what will be called, fourteen years after, Effective Altruism (EA) at the founding assembly of the Center of Effective Altruism, which precisely took place on December 3, 2011 (MacAskill, 2019). William MacAskill (2019, p. 14), a founding member of the movement, defined EA as

the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources, tentatively understanding 'the good' in impartial welfarist terms, and (ii) the use of the findings from (i) to try to improve the world.

The movement encourages individuals to consider their intentions behind charitable giving and the actual outcomes and impacts of their donations or efforts, which must always be maximized in a utilitarian conception. Therefore, two fundamental and indissociable principles, among others, must be considered: (1) prioritizing causes and (2) considering the distant future, also known as long-termism, which emerges as a newfound emphasis of the movement, as elaborated in subsequent lines.

The prioritization of causes is a fundamental aspect of the EA movement, which involves identifying global issues with the most potential for positive change and focusing resources on those areas. Initially, poverty in the Global South and animal welfare in factory farming were EA's primary areas of interest, highlighted by Peter Singer's works on "Famine, affluence and Morality" (Singer, 1972). However, the philosophical movement has since evolved to consider technological innovation, particularly AI, as one of the main areas where action can be most impactful. This shift has been motivated and influenced by implementing long-termism in EA utilitarian philosophy. As noted by philosopher Alice Crary (2023, p. 49), the founders of the EA movement, William MacAskill and Toby Ord, began talking about and implementing long-termism into their EA framework through their affiliation with the Future of Humanity Institute, an institution founded by Nick Bostrom.

Bostrom (2002) introduced the concept of *existential risk* in his article "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," defining it as a scenario where an adverse outcome could lead to the annihilation of intelligent life originating from Earth or significantly and permanently reduce its potential. He categorized existential risk into four branches: (1) Bangs, representing sudden extinction through accidents or deliberate acts; (2) Crunches, indicating a blocked path to post-humanity despite human survival; (3) Shrieks, denoting limited and undesirable post-humanity; and (4) Whimpers, signifying gradual negative evolution of post-humanity. Within the Bangs category, Bostrom highlighted the risk posed by badly programmed superintelligence, now referred to in the literature as unaligned AGI, alongside other threats such as misuse of nanotechnology, nuclear holocaust, and simulation shutdown.

Concerning the potential dangers of unaligned superintelligence, Bostrom (2002) cautioned that the creation of the first superintelligent entity might inadvertently result in the entity pursuing goals that could lead to the annihilation of humankind. For instance, an error in goal setting

could prompt the superintelligence to transform the solar system into a giant calculating device, inadvertently causing harm to individuals in the process.

In line with Effective Altruism's cause prioritization and driven by a long-term philosophy, EA members advocate addressing existential risks, particularly those posed by non-aligned superintelligence, as a global priority (Crary, 2023).

After identifying these existential risks of rapid technological advancement, particularly the emergence of superintelligent machines, one may think that Bostrom's next reasoning step solution will be the deceleration of technological development. However, in his 2003 paper "Astronomical Waste: The Opportunity Cost of Delayed Advancement in Technology," he argues that delaying technological advancement and universe colonization would result in missed opportunities for significant well-being and many fulfilled lives (Bostrom, 2003). For him, it is precisely through the advancement of technology that we will still find the solution to his existential threat, hence the logic behind all the investment of Effective altruists in pushing the brand of AI safety, with the ambition of saving humanity from possible AGI that may want to exterminate us. The idea for them is that AGI is not the problem; it is the type of AGI that is problematic. Therefore, the EA donors are directing resources toward organizations advocating the development of safe and beneficial forms of AGI. For instance, the Center for AI Safety (CAIS), a recipient of a \$5 million grant from prominent EA donor Open Philanthropy, advocates for responsible AI development (Open Philanthropy, 2023). Similarly, the Alignment Research Center (ARC), founded by Paul Christiano, is another recipient of funding, both from Open Philanthropy (\$1.2 million grant) (Open Philanthropy, 2022), and the FTX Foundation (\$1.5 million grant), established by cryptocurrency entrepreneur Sam Bankman-Fried, a huge supporter of the Centre for Effective Altruism (Alignment Research Center, 2024). Alice Crary shortens all these FTX and EA connections:

It was also known that, with a group of Oxford-affiliated longtermists, MacAskill had been an advisor to FTX's charitable Future Fund and that the Future Fund had committed large sums to build EA's institutions, including fourteen million dollars to MacAskill's main organization, the Centre for Effective Altruism, fifteen million to Longview Philanthropy, for which MacAskill is an advisor and another roughly seven million to fellowships, prizes, and the like at these and other organizations with which MacAskill is affiliated.

(Crary, 2023, p. 49)

Talking about Sam Bankman-Fried, the recent downfall of his Cryptocurrency FTX bank has sparked a wave of criticism directed at EA. Timnit Gebru (2022) accuses the movement of promoting and prioritizing a dangerous form of AI safety excessively focused on the long-run future, all in the name of humanity's salvation. While the intention may seem commendable (concern for developing AI aligned with human values and non-extinctionist), their long-termist identity and vocabulary amplify discussions around existential risks, potentially diverting attention from pressing present AI issues, including discrimination biases; ecological impacts of Large Language Models (LLMs); misinformation; ethnic, political, and cultural biases. This critique of EA's long-termist emphasis is also echoed by Singer, the founding father of EA philosophy, as noted by Crary:

Singer is skeptical about whether humanity is indeed at a uniquely portentous moment in history, and he de-emphasizes existential risk in a manner that indicates impatience with long-termists' commitment to the posture they call non-neutrality. His aim is to redirect attention back to EA's emphasis on suffering now and in the short-term. 'If we are at the hinge of history,' he writes, 'enabling people to escape poverty and get an education is as likely to

move things in the right direction as almost anything else we might do; and if we are not at that critical point, it will have been a good thing to do anyway.’

(Crary, 2023, p. 51)

Opposing Singer’s proposed solution, Crary notes that this long-termist tradition not only falsely prioritizes existential questions over other moral issues but also “is incapable of furnishing an understanding of our social circumstances that could responsibly inform future-oriented action.” The critics align with the historical criticism of EA, namely, the movement’s inability to grasp and address the systemic socio-political structures that produce the problems it claims to solve rationally and efficiently (Crary, 2023, pp. 51–52).

Other much bolder critiques come from Eff/acc who reproach EAs for their ambivalent consideration of AGI as either a potential existential threat to humanity if misaligned or a savior of humanity. For effective accelerationists, viewing AGI as an existential danger opens the door to the possibility of regulating the technology, which hinders their ambition to accelerate its development. Despite distancing themselves from EAs on this regulatory point, effective altruists still draw on the eschatological (related to the end of times) vocabulary of EAs to justify the acceleration of a technology that they believe is beneficial for humanity. We propose to mark a pause to the conceptual origin of accelerationism, from its philosophical roots in the mid-1970s to its recent redefinition and reappropriation by the AGI quest ecosystem.

### 2.3 *Accelerationism: the additional comforting narrative*

In his work *No Speed Limit*, Steven Shapiro (2015) introduces the concept of accelerationism through an introductory parable of *creative destruction*, derived from Lee Konstantinou’s science fiction novel *Pop Apocalypse*. The novel presents a Marxist school of thought in which adherents view Karl Marx’s writings on capitalism as a literal prediction of the future. Their mission is to ensure the practical realization of that prediction by pushing the capitalist markets toward their apocalyptic state. This apocalypse, which is both destructive and creative, is believed to contribute to the true socialist revolution.

This creative destruction idea offers a clear view and comprehension of accelerationism: a provocative and radical approach to addressing the dynamics of capitalism. At its core, accelerationism acknowledges the inherent contradictions and tensions within capitalism, proposing that rather than resisting or withdrawing from capitalist processes, one should accelerate them to their breaking point, thereby hastening the arrival of a post-capitalist society. Hence, the proverbial illustration, *the only way out is through*.

From a historical point of view, according to Benjamin Noys (2013), the author of the neologism *accelerationism*, the concept emerged in response to the perceived limitations of traditional leftist strategies, which often sought to resist or mitigate the effects of capitalism. Three influential books from the mid-1970s encapsulate a distinct line of thought: Deleuze and Guattari 1972s *Anti-Oedipus: Capitalism and Schizophrenia* critiqued psychoanalysis, advocating for embracing capitalist processes to transcend limitations; Lyotard’s 1974 *Libidinal Economy* extended this critique, suggesting that resistance to capitalism is futile and individuals find pleasure even within subjugation; Baudrillard’s 1976 *Symbolic Exchange and Death* proposed a symbolic challenge to capitalism, advocating for acts of reversal and negation to undermine its values and structures (Noys, 2013, pp. 1–5). These works collectively challenge traditional leftist notions, offering provocative perspectives on capitalism and desire.

Overall, accelerationism represents a complex and controversial theoretical position that grapples with capitalism's inherent contradictions and complexities. While it has been criticized for its potentially nihilistic implications and its embrace of capitalist processes, it has also been seen as a radical reimagining of leftist politics in the face of the apparent inevitability of capitalist domination.

Drawn on this same idea, Alex Williams and Nick Srnicek (2013), in their “#ACCELERATE MANIFESTO for an Accelerationist Politics,” argue for establishing a new political approach to address the unprecedented challenges facing today's world. These include the threats of climate change, financial instability, and the disruptive impact of automation, etc. Their manifesto acknowledges the failure of current leftist political systems to respond to these crises adequately and proposes a radical accelerationist approach. The means is to embrace technological progress:

We want to accelerate the process of technological evolution. But what we are arguing for is not techno-utopianism. Never believe that technology will be sufficient to save us. Necessary, yes, but only sufficient with socio-political action. Technology and the social are intimately bound up, and changes in either potentiate and reinforce changes in the other. Whereas the techno-utopians argue for acceleration because it will automatically overcome social conflict, our position is that technology should be accelerated precisely because it is needed to win social conflicts.

(Williams & Srnicek, 2013)

This recommendation to accelerate technological innovation has been embraced and implemented within the AGI seeker's ecosystem. However, this is not merely a symbolic gesture aimed at combating neoliberal capitalism. Instead, it serves as a narrative justification for the aggressive advancement of technology, such as AGI, in pursuit of a world of prosperity and abundance. Despite the existential threat that AGI may pose, proponents find it necessary to continue developing and advancing this technology. Fueled by the logic of creative destruction, the narrative of accelerationism offers a compelling political and philosophical justification for this pursuit. Effective accelerationists who refuse to entertain criticisms of safety and regulation options proposed by EA push this narrative to its extreme.

In this context, effective accelerationism aligns with EA's focus on long-term goals and the impact of advancing technology on humanity's salvation. However, it diverges from slowing AGI development or viewing AGI as a potential existential threat. Instead, the primary objective is to create a benevolent god-like entity, AGI, to safeguard humanity from extinction by addressing our most pressing challenges through a solution-oriented approach. This viewpoint shares similarities with a techno-millenarian ideology, which envisions a technological revolution leading to notable societal transformation. It is influenced by religious beliefs like millenarianism, emphasizing a forthcoming era of prosperity, often linked to the Second Coming of Christ. This perspective is fundamental to the teachings of various groups, including Adventists, Latter-day Saints (Mormons), and Jehovah's Witnesses. Therefore, the theological and political consequences of the accelerationist storyline regarding AGI are of great significance.

### **3 The theological-political implications of the accelerationist narrative: the providential power of AGI and his inherent sacrifices**

Narrative constructs power, and power is perpetuated by narrative. The narrative of accelerationism, intertwined with eschatological vocabulary (apocalypse, humanity's extinction, existential risk, salvation of humanity), contributes to forming an imaginary realm where AGI



assumes a God-like stature with providential abilities. The subtly Judeo-Christian and millenarist narrative assigns AGI the mission of humanity's salvation. This narrative also explains the growing usage of the term *God-like AI* to describe AGI, particularly following the advancements seen with ChatGPT and the influential essay by Ian Horgarth (2023) urging an alt to stop the race after AGI. To gain a deeper understanding of the political implications that arise from the emergence of a God-like AI, it would be insightful to examine the works of Giorgio Agamben (2006), an Italian philosopher who has delved into the genealogies of economic power as a providential force.

According to Agamben, the modern conception and realities of the economy as a form of power can be traced back to the 2nd century AD, when lively discussions emerged within Christian theology around the trinity of the Christian figure, the father, the son, and the Holy Spirit. Some participants in this discussion feared that this concept would open the door to a reinstallation of polytheism within Christianity. To reassure them, theologians such as Tertullian, Hippolyte, and Irénée borrowed Aristotle's notion of *oikonomia* to introduce it into the debate. Whereas in Aristotle, the idea of *oikonomia* referred to the governance of the household or private affairs, in the discussion of the trinity, the notion took on the meaning of the providential action of God. The project was, therefore, to show that God, in his being and nature, is one, but in his providential action, that means how he governs his world and his creatures, he is trine. As a father, he can delegate this action to his son, just as the son, mandated by the father, can also delegate this power to the Holy Spirit. That is how the notion that served Aristotle to qualify the government of private affairs was exploited theologically by Greek church fathers from the 2nd to the 4th century to qualify the providential government of creatures (Agamben, 2006, pp. 27–28). The Latin church fathers will propose the Latin translation, *dispositio*, which will give rise to the French term *Dispositif* (and the English one *apparatus*) (Agamben, 2006, pp. 27–28). In other words, and this was the stake of Agamben's genealogical work, the notion of "apparatus" that he borrows from Foucault to conduct a genealogical reflection on power finds its roots in discussions around *oikonomia*, the government of creatures. This genealogical connection explains, to some extent, the modern sense of *economy* as governmental power "a set of practices, knowledge, measures, and institutions whose purpose is to manage, govern, control, and orient, in a way that is intended to be helpful, the behaviors, gestures, and thoughts of men." (Agamben, 2006, p. 29).

This *dispositio*, the Latin substitute for *oikonomia*, is defined by Agamben as

everything that has, in one way or another, the capacity to capture [subjugate], guide, determine, intercept, shape, control, and ensure the gestures, behaviors, opinions, and discourses of living beings. Not only prisons, therefore, but also asylums, the panopticon, schools, confession, factories, disciplines, and legal measures, whose articulation with power is in a sense evident, but also the pen, writing, literature, philosophy, agriculture, cigarettes, navigation, computers, mobile phones, and, why not, language itself, which is perhaps the oldest apparatus in which, several thousand years ago, a primate, probably unable to realize the consequences that awaited it, had the unconsciousness to be caught.<sup>1</sup>

(Agamben, 2006, pp. 29–30)

Interestingly, in their subjugation of individuals, these apparatuses embody a type of governmental power that aligns with a fundamental characteristic of providential and sacred power: the sacrificial violence. Agamben captures this concept of sacrifice as a *collateral effect*, a term borrowed from Philo of Alexandria, which denotes an inherent aspect of exercising providential power (Lasha Matiasvili, 2018). To fully comprehend these notions of collateral effect as an intrinsic aspect

of providential power, it is crucial to consider the anthropological insights of René Girard concerning the concepts of the sacred, sacrifice, violence, and scapegoating.

According to René Girard (1998), violence is a fundamental and defining element of the sacred. The intimate relations between humans, driven by their perpetual desires, whether for or against someone or something, are inherently violent, and this violence is an unavoidable aspect of both primitive and contemporary societies. In such settings, the stability of the society depends on the channeling and directing of this violence toward a scapegoat who serves as the expiatory victim. This expiation of violence on this scapegoat figure is precisely the function assumed by the sacred power, which, at the heart of these violent societies, counters violence with violent means, the most traditional form of which is sacrifice (Girard, 1998).

In the heart of our modern secularized societies, this function of expiating violence on a scapegoat has not disappeared. It has migrated from religious figures and categories to other governmental entities, such as the modern economy, and that was the point of Agamben. The philosopher and economist Jean-Pierre Dupuy further delved into this thesis by showing this same sacrificial violence function in the economy and its utilitarian rationality, where the evils committed by the market are analyzed, from Adam Smith to Friedrich Hayek, “as sacrifices that must be accepted for the greater good” (Dupuy, 2001, p. 40).

Our point is that this providential power, inherent to a violent sacrificial dimension, is transferred to AGI through an accelerationist narrative and creative destruction logic. First of all, following on from cell phones and the Internet (one of the apparatuses mentioned by Giorgio Agamben), advanced AI (that we can also call advanced algorithmic systems) also meets the definitional criteria of Agamben apparatus in that it captures us, intercepts our data from which it is trained, and guides (through its predictive and recommendatory capacity) our decisions and actions. It, therefore, participates in a form of governance of creatures, what Antoine Rouvroy, inspired by Foucault, has called algorithmic governmentality: “The idea of a government of the social world that would be based on the algorithmic processing of massive data rather than on politics, law, and social norms” (Rouvroy, 2020).

Secondly, there is a sacralization of AGI through the accelerationist and eschatological narrative that sees AGI as the savior of humanity. Therefore, his development must be accelerated to respond to this mission. This sacralization of AGI (what we propose to capture by the neologism AGI-Theism: the attempt to erect, narratively, AGI to a god-like status) positions AGI to a providential status that consubstantially integrates a sacrificial and violent dimension, what Agamben has called collateral effect. According to the accelerationist philosophy, these collateral effects must be embraced because, through the creative destruction logic, they are sacrifices necessary for the upcoming post-era. In the AGI quest, if we want to stay in a theological vocabulary, the modern sacrificed entities (victims of the collateral effect) can be interpreted as all Kenyans who were paid \$2 per hour to label data used to train the algorithm surrounding ChatGPT (Perrigo, 2023); the marginalized communities in Venezuela whose body energy is exploited to fuel the manual data labeling processes on which the AGI industry relies (Hao & Hernández, 2022), etc.

#### **4 What response? Philanthropy as a tool for secularizing the providential power of AGI or AGI-owners: from the power-over to the power-with**

Comparisons can be made between the sacralization of AGI and premodern times when deities were believed to have divine decision-making abilities. These periods were eventually followed by secularization, where deities were stripped of their powers and either shared with (1) or monopolized by humans (2). In both scenarios, the once-revered deified entities no longer held hegemonic power.

From a theological-political point of view, this hegemonic power can be qualified as a power-over. We borrow this characterization from the French philosopher Paul Ricoeur. In his article “Political Power: The End of the Theologico-Political,” he proposes, based on an interpretative analysis of Romans 13, to distinguish between two forms of power: *power-over*, vertical power or power of subordination, and *power-with*, horizontal power or power of cooperation (Ricoeur, 2021). In other words, the ambition of a secularization project, faced with a providential power embodied by AGIs and their owners, would be to transition from a vertical power-over to a horizontal power-with.

Philanthropy can play a crucial role within this metaphorical secularization of the *power-on*, held by AGIs and maintained by accelerationist narratives, to a *power-with*, carried by both AGI and owners as well as all the others, named above as the sacrificed entities. These shared powers, therefore, imply a co-construction enterprise. Philanthropic organizations must undertake two inseparable fields of action: (1) The promotion of a free/libre and open sourced AI (FLOSS) development approach as a condition of shared power and (2) The empowerment and Reinforcement of the Peripheries of the AI Ecosystem to Mitigate Risks Associated with AI Openness.

#### ***4.1 Philanthropy as an instrument for eroding the hegemonic power of AGIs and AGI-owners: a free and open-source approach***

On February 17, Elon Musk, CEO of Tesla and co-founder of OpenAI, filed a lawsuit against the current leaders of OpenAI for betraying the founding spirit of OpenAI. This company, registered in 2015 as a non-profit organization, aimed to counterbalance the power of Google in the race for AGIs. In this mission, the project was to remain an open-source company where the code and development science were open to external contributions and expertise. However, its collaboration with Microsoft has transformed it into a for-profit company with a closed source or proprietary spirit. The argument favoring this transformation (arguments highlighted in OpenAI’s response to Elon Musk) was that substantial financial resources were needed to compete in the race to develop AGI against Google DeepMind, and the status of a for-profit company would make this possible (Brockman et al., 2024).

This closed source approach, which is consubstantial with the neo-capitalist proprietary spirit, reinforces the elitist, monopolistic, and hegemonic narrative of a single providential power: Just as in Judeo-Christian cultures, where only one figure – the Christ – can save humanity, we are the only ones capable of developing beneficial technology for humanity.

If we entertain this techno-millennarian ambition, we can easily disagree with this monopolistic and hegemonic conception of salvation. To effectively save humanity while adhering closely to the Judeo-Christian model of salvation, we should consider that the redemptive figure, Christ, acknowledges humanity in its condition and shares power with it within the framework of a new covenant. In AGI development, the essence of such covenants lies in the spirit of free and open-source software development: a mechanism through which power is shared horizontally, co-constructed, between developers and eventually users. In addition, the philosophy of free and open-source development is closely and historically linked with the philanthropic ethos. To understand this connection, it is essential to examine the history of the free software movement and the Open Source Initiative.

##### ***4.1.1 GNU project and the free software movement***

The origins of the free software movement are often traced back to 1983 with the GNU project, developed by Richard Stallman. However, Stallman himself, in a text available on the original

website of the GNU Operating System, attests that his introduction to the culture of sharing software freely dates to the early 1970s, during his tenure at the MIT Artificial Intelligence Laboratory (Stallman, n.d.a). There, as an AI Lab staff system hacker, he was tasked with enhancing the lab's timesharing operating system, ITS (Incompatible Timesharing System), designed and written by the lab's staff hackers for the Digital PDP-10 computer. Although the term "free software" was not yet coined, the ethos of openness and sharing prevailed.

In the 1980s, the landscape shifted with the discontinuation of the PDP-10 and the lab's transition to proprietary operating systems, curtailing the culture of open collaboration. This shift had profound implications, as modern computers came with proprietary operating systems that restricted access through non-disclosure agreements, impeding collaboration and innovation. Stallman viewed this proprietary software model as inherently anti-social and unethical, as it isolated users and hindered progress. Following the collapse of the community of software sharers in that same period, Stallman was confronted with a moral dilemma. He was faced with the choice of entering the world of proprietary software, which promised the allure of financial rewards and fulfilling work opportunities, or abandoning computing altogether to avoid further entrenchment in the problematic proprietary software landscape.

After careful consideration, Stallman leveraged his existing skills to create a free operating system that promoted user collaboration and fostered a collaborative hacking community. This decision marked the beginning of the GNU operating system, which aimed to provide a free and alternative version of AT&T's UNIX, a widely used multiuser and multitask operating system based on a proprietary code model. Additionally, Stallman's initiative marked the inception of the Free Software Movement, which aimed to develop software that guaranteed users four fundamental freedoms: the freedom to run the program for any purpose; the freedom to study and adapt the code for personal use; the freedom to redistribute copies of the program, either gratis or for a fee; and the freedom to distribute improved or modified versions of the program to the public (Stallman, n.d.a). Stallman emphasizes that the term "Free" in this commitment is not tied to the notion of price but rather to the ability of users to participate in the free software process of development and deployment. Therefore, the possibility of selling Free Software and other related services does not contradict this conviction. Instead, they are deemed necessary for funding free software development, which requires resources. To organize these funds and ensure the continued financial support and growth of the GNU project while maintaining its commitment to free software principles, Stallman and his colleagues founded the Free Software Foundation in 1985. The Foundation's mission was to generate revenue by selling free licenses and through contributions from community members. The initiative represented a pioneering approach, linking free software projects with philanthropic investment models.

#### *4.1.2 Netscape project and the Open Source Initiative*

In 1993, Marc Andreessen and a team of fellow students at the University of Illinois at Urbana-Champaign founded the first graphical web browser, known as Mosaic, at the National Center for Supercomputing Applications (NCSA) (United States Department of Justice, 2006). In 1994, James Clark, a former professor at the University of California and Stanford, joined Andreessen in this endeavor, and together they founded the company Netscape in April 1994. Five months later, on December 15, 1994, they released the final version of the web browser Navigator 1.0. This browser quickly rose to prominence, becoming one of the dominant web browsers alongside Microsoft Internet Explorer. In 1998, prior to the acquisition of Netscape by AOL, Netscape Communications announced its decision to release the source code of its browser, inspired by Eric

S. Raymond's influential work "*The Cathedral and the Bazaar*," published eight months earlier (Raymond, 2000).

This announcement led to a schism within the free software community, originally initiated by Richard Stallman. A faction of the community broke away and formed a new community that coined the term "open source" (Stallman, n.d.b). Additionally, they established the Open-Source Definition, which outlines the criteria for software to be classified as open source. These criteria include the free redistribution of the software; availability of the source code; ability to create and distribute derived works; maintenance of the integrity of the source code; absence of discrimination against users or uses; freedom to use the software in any field of endeavor; redistribution of the software without requiring additional licenses; independent use of the software regardless of its original distribution; avoidance of restrictions on bundled software, technology neutrality, and independence from specific technologies (Open Source Definition, 2000).

The release of the Netscape source code and the subsequent formation of the Open Source Initiative's community gave rise to the Mozilla Project, a collaborative effort to leverage the collective expertise of thousands of programmers on the Internet to foster innovation in the browser market (Mozilla, n.d.). The Mozilla Project fostered an open community that transcended the confines of any single company, allowing community members to become deeply involved in the project's development. As a result, the project's original mission of developing Netscape's next browser expanded to include the creation of various browsers, development tools, and other projects, with the aim to empower users to choose how they interacted with the Internet.

In 2002, Mozilla 1.0 was successfully released, marking a significant milestone in the project's evolution. The following year, the community established the Mozilla Foundation to spearhead the open-source browser initiative and advocate for Internet openness. With the support of the Foundation, the company introduced Firefox 1.0 in 2004, which quickly gained widespread acclaim and amassed over 100 million users within its first year of release (Mozilla, n.d.).

#### *4.1.3 The confrontation between the free software movement and open source*

Partisans of the free software approach, Richard Stallman in the head, argue that the open-source movement does not emphasize freedom and justice. Instead, it is a business-friendly methodology that focuses highly on the practical advantages of software (Stallman, n.d.b). As stated by the Open Source Initiative's mission,

Open Source is a development method for software that harnesses the power of distributed peer review and transparency of the process. The promise of open source is better quality, higher reliability, more flexibility, lower cost, and an end to predatory vendor lock-in.

(Open Source Initiative, 2006)

The critique lies precisely in asserting that the excellent quality of software is not monopolized or exclusive to the open-source and collaborative methodologies. Closed and proprietary approaches are equally capable, if not better, due to their substantial financial resources of developing extremely high-quality software. However, the key difference with these proprietary approaches is the freedom of users: the freedom to run the program for any purpose, to study and adapt the code for personal use, to redistribute copies of the program either gratis or for a fee, and to distribute improved or modified versions of the program to the public. This free software may not be inherently superior to proprietary software, but it chiefly upholds user freedom (Stallman, n.d.b).

In considering the secularization of providential powers of AGI, it would be apt to use the term *Free software*, as it aligns with the goal to empower end-users by granting them the four categories of freedom mentioned above. However, to underscore the significance of the open-source aspect (although inherent in the ethos of free software), we will refer to it as *Free and open-source software* (FOSS). Moreover, echoing Stallman's advocacy, we can incorporate the term *libre* (in French) alongside *Free* – Free/libre – to emphasize that *Free* is not linked to the notion of price (Stallman, n.d.c). Hence, the restated central thesis is that AGI, when developed as Free/Libre and Open-Source Software (FLOSS), will guarantee empowerment for both users and AGI developers in the era of advanced AI. To ensure long-term sustainability, such FLOSS AGI models must be supported by philanthropic organizations. This philanthropic investment approach – mirroring the original backing provided by the Free Software Foundation for the GNU project and the Mozilla Foundation for the Mozilla browser – is particularly well-suited for projects involving FLOSS due to its inherent non-proprietary nature.

Both proprietary software and FLOSS require investments to be effective, although these investments serve different purposes. However, if these investments are solely the responsibility of an entity driven by the capitalist logic of profit and return on investment, the most profitable course of action, based on competitive logic, would be to monopolize the product and assert proprietary control over it. In contrast, the philanthropic logic of investment disregards the parameter of competition and prioritizes cooperation, aligning more closely with the spirit of FLOSS and non-proprietary principles.

Moreover, philanthropic organizations' investments offer a more structured and sustainable model than relying solely on individual volunteering and contributions. Currently, numerous open-source projects are maintained by individuals who work on them in their spare time without adequate compensation. According to the 2023 Tidelft State of the Open Source Maintainer report, 60% of maintainers identified themselves as unpaid, three-fourths of the unpaid will prefer to be paid (Tidelft, 2023, pp. 3–4), and 58% have contemplated quitting at least one of their projects (Tidelft, 2023, p. 28). When asked about the reasons behind their consideration of quitting, 38% cited financial concerns, and 36% mentioned the significant time commitment required by these projects, especially considering they are not their primary source of income (Tidelft, 2023, p. 29).

By providing financial support, philanthropic investments allow these individuals to dedicate themselves fully to these projects, ensuring sustainability while preserving the non-proprietary nature of the software. In this vision lens, various promising funding models, such as the FOSS Contributor Fund, have emerged in the last decade, with the first framework initiated by Indeed in January 2019 (Indeed, 2019). It enables the company's employees to nominate open-source software projects they rely on or participate in daily. Several other companies, including Bloomberg, the Open Technology Fund, John Hopkins University, Spotify, and others, have launched similar programs following this initiative. However, despite the interest and support these initiatives provide to FOSS, overall funding through such channels still needs to be increased. According to Kara Sowles's 2024 presentation at the *State of Open Con* in London, only \$12 million went to open-source projects through the FOSS contributor mechanism (Sowles, 2024).

In addition to the lack of consistent investment, one problem identified while observing this FOSS contributor fund mechanism is the punctual and unpredictable nature of the funding. Only a long-term engagement and commitment between philanthropic institutions and recipients' structure can sustain FLOSS development endeavors and mitigate risks of the open-source culture, particularly concerning projects related to disruptive AGI technologies.

#### 4.2 Long-term partnership between POs and FLOSS as a tool to mitigate the open-source risks

The rising interest in open-source AI is significant across various fields, reminiscent of its prominence in the early 1970s before proprietary systems gained dominance. A recent survey by venture capitalist firm a16z, published in March 2024, revealed that the appeal of the open-source model represents one of the most significant changes in the landscape over the past six months (Wang & XU, 2024). In 2023, the market was dominated by OpenAI proprietary model, with estimates ranging from 80% to 90% market share. However, going into 2024, there is an anticipated shift toward open source, with some enterprises aiming for a 50/50 split from the 80% closed/20% open split. According to the survey, 60% of respondents ranked the criterion of *control* as the most important for adopting an open-source approach, followed by customization of existing models (30%) and cost (10%).

Nevertheless, in light of Meta's recent move to open source its AI foundation model (Llama), some skeptics perceive the open-source trend as a potential marketing tactic to attract talent (Nover, 2024). The term *openwashing*, analogous to *greenwashing*, characterizes the marketing strategy software companies employ to project an image of openness while simultaneously practicing proprietary methods (Wider et al., 2023). Crul (2024) outlines three primary concerns expressed by skeptics regarding this practice. Firstly, there is apprehension that companies may exploit open source to attract talent, drawing developers into their ecosystem and potentially constraining diversity and collaboration. Secondly, there is concern about exploitation, where companies release open-source models to glean the best ideas and incorporate them into their proprietary, closed source products, potentially stifling fair competition and innovation. Lastly, there is a worry about regulatory loopholes, wherein companies may push for lenient regulations to promote innovation and competition, potentially compromising safety and ethical standards.

Other critiques raise concerns about the high-security risks associated with allowing criminals to participate in the co-development process of open-source AI (Harris, 2023). Additionally, there are apprehensions about the potential for open-source approaches to propagate false narratives, exemplified by Wikipedia, a prominent figure in open-source culture.

Despite all the valid concerns raised, the imperative of fostering an open-source culture remains paramount, knowing that its core aim is the redistribution of power to the people. From a philosophical standpoint, collectively co-constructed power is intricately intertwined with the collective responsibility and accountability for the risks engendered by this shared power.

Indeed, risks or collateral effects, as referenced by the Italian philosopher Agamben, will invariably accompany our collective decision-making processes regarding the construction and exercise of collective power. In the history of democracy, the choice to democratize the press carried the risk of misinformation and the rise of fascist press. However, despite these potential pitfalls, the openness, diversity, and freedom of the press are crucial for fostering democracy and facilitating the collective expression of power. Likewise, while the openness of AGI development may pose risks highlighted by critics, such as the potential for misuse or exploitation, it is vital not to overlook the significance of openness. It remains a cornerstone for enabling collective empowerment and constructive progress. The crux lies in minimizing these risks through meticulous strategic approaches to openness; in this case, the work of Irene Solaiman can be helpful.

In her 2023 article "The Gradient of Generative AI Release: Methods and Considerations," Solaiman explores different strategies for making generative AI tools more open. The article compares various release methods, from closed systems like Google's Imagen and DeepMind's Gopher to open ones like EleutherAI's GPT-J and Big Science's BLOOM. In between these extremes

lie options like staged releases, gated access through online interfaces or APIs, and downloadable models, each providing different levels of control and ease of use. The best release method depends on various factors, including managing potential risks and fostering further research. To meet this, Solaiman highlights the importance of investment in six core areas: (1) facilitating accessible interfaces and tools, including low-code/no-code options, to encourage broad participation and diverse perspectives in research and evaluation; (2) bridging the resource gap for under-resourced groups to ensure equitable access to opportunities and resources in AI development; (3) providing comprehensive technical and practical ethics training for developers and researchers to instill ethical considerations into AI development practices; (4) engaging in proactive and interdisciplinary research with experts from various relevant fields to anticipate and address potential risks and challenges; (5) fostering multidisciplinary discourse among diverse stakeholders, including underrepresented communities, through the establishment of a new independent body. This facilitates discussions and ensures accountability for safe AI releases; (6) establishing enforcement mechanisms to regulate AI development and deployment, ensuring adherence to ethical and safety standards (Solaiman, 2023, pp. 10–12).

Philanthropic organizations, in addition to supporting free/libre and open-source models as advocated above, should invest in these six areas to contribute to the establishment of a safe ecosystem for AI research and development. The quality and openness of AI and AGI depend on the quality of the ecosystem and the context in which they are rooted and developed.

## 5 Conclusion

The goal of this chapter was to explore critically the providential and hegemonic power attributed to AGI through the accelerationist, proprietary, and millenarian ideological-political trends that underpin and structure the race toward its achievement. Even if this AGI may never exist, as attested by some specialists and rigorous studies in the field, philanthropy can already play a vital role in shaping the present and future of AI by challenging and dismantling these proprietary narratives and advocating for a more open, inclusive, and collaborative approach. By championing the development of free and open-source AI models, philanthropy can promote a horizontal and shared power structure, countering the concentration of power in the hands of a few. This support for non-profit and non-proprietary AI development aligns with the ethos of organizations like the Free Software Foundation, the Mozilla Foundation, the Signal Foundation, and the Linux Foundation, which have demonstrated the potential of free, open, community-driven initiatives, even if not yet in a perfect form of expression. As we continue to explore the possibilities and limits of AI, philanthropy must remain vigilant in its efforts to ensure that AI, or the hypothetical AGI, is openly developed and used for the greater good.

## Note

- 1 We are the one translating the citation.

## References

- Agamben, G. (2006). Théorie des dispositifs. *PO&SIE*, 115(1), 25–33. <https://doi.org/10.3917/poesi.115.0025>
- Alignment Research Center (2024). *Funding from FTX*. Retrieved from Récupéré sur Alignment Research Center: <https://www.alignment.org/funding-from-ftx/>
- Altman, S. (2023, February 23). *Planning for AGI and Beyond*. Retrieved from OpenAI: <https://openai.com/blog/planning-for-agi-and-beyond>



- Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, 9(1), 1–36.
- Bostrom, N. (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, 15(3), 308–314.
- Brockman, G., Sutskever, I., Altman, S., Schulman, J., Zaremba, W., & OpenAI. (2024, March 5). *OpenAI and Elon Musk*. Retrieved from OpenAI: <https://openai.com/blog/openai-elon-musk#IlyaSutskever>
- Crary, A. (2023). The Toxic Ideology of Longtermism. *Radical Philosophy* (214), 49–57.
- Crul, S. (2024, March 11). *Is Open Source AI a Sham?* Retrieved from Récupéré sur Freedom Lab: <https://www.freedomlab.com/posts/is-open-source-ai-a-sham>
- Dupuy, J. P. (2001). Le détour et le sacrifice : Ivan Illich et René Girard. *Esprit*, 5(274), 26–46. Retrieved from Récupéré sur: <https://www.jstor.org/stable/24279442>
- Effective Acceleration (2023). *Effective Acceleration Means Accepting the Future*. Retrieved from Effective Acceleration: <https://effectiveacceleration.tech/>
- Gebru, T. (2022). Effective Altruism is Pushing a Dangerous Brand of ‘AI safety’. *Wired*. <https://www.wired.com/story/effective-altruism-artificial-intelligence-sam-bankman-fried/>
- Gebru, T., & Torres, É. (2024). The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence. *Peer-Reviewed Journal on the Internet*. Retrieved from: <https://firstmonday.org/ojs/index.php/fm/article/view/13636/11606>
- Girard, R. (1998). *La Violence et le sacré*. Paris: Hachette.
- Hao, K., & Hernández, A. P. (2022, April 20). *How the AI Industry Profits from Catastrophe*. Retrieved from MIT Technology Review: <https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/>
- Harris, D. E. (2023, December 4). *How to Regulate Unsecured “Open-Source” AI: No Exemptions*. Retrieved from Récupéré sur Tech Policy Press: <https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/>
- Horgharth, I. (2023, April 12). *We Must Slow Down the Race to God Like AI*. Retrieved from Récupéré sur Financial Times: <https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2>
- Indeed. (2019, January). *Investing in Open Source: The FOSS Contributor Fund*. Retrieved from Récupéré sur Indeed Engineering: <https://opensource.indeedeng.io/Investing-in-Open-Source/>
- Landgrebe, J., & Barry, S. (2022). *Why Machine Will Never Rule the World: Artificial Intelligence Without Fear*. London: Routledge.
- Lasha Matiashvili, L. (2018). The Mystery of Power in the Philosophy of Giorgio Agamben. *Journal of Social Sciences*, 1(2), 41–64.
- MacAskill, W. (2019). The Definition of Effective Altruism. In *Effective Altruism: Philosophical Issues*. Edited by Hilary Greaves and Theron Pummer, Oxford University Press.
- Morris, M. R., Sohl-dickstein, J., Fidel, N., Warkentin, T., Dafoe, A., Faust, A., ..., & Legg, S. (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. *arXiv: 2311.02462*. <https://doi.org/10.48550/arXiv.2311.02462>
- Mozilla (n.d.). *The History of the Mozilla Project*. Retrieved from Mozilla: <https://www.mozilla.org/en-GB/about/history/>
- Nover, S. (2024, January 23). *Why Meta Opened Up*. Récupéré sur Gzero: <https://www.gzeromedia.com/gzero-ai/why-meta-opened-up>
- Noys, B. (2013). *Malign Velocities: Accelerationism and Capitalism*. Winchester; Washington, DC: Zero Book.
- Open Philanthropy (2022, November). *Alignment Research Center – General Support (November 2022)*. Retrieved from Open Philanthropy: <https://www.openphilanthropy.org/grants/alignment-research-center-general-support-november-2022/>
- Open Philanthropy (2023). *Center for AI Safety – General Support (2023)*. Retrieved from Récupéré sur Open Philanthropy: <https://www.openphilanthropy.org/grants/center-for-ai-safety-general-support-2023/>
- Open Source Definition (2000). Retrieved from Open Source Initiative: <https://opensource.org/definition-annotated>.
- Open Source Initiative (2006). *About the Open Source Initiative*. Retrieved from Récupéré sur Open Source: <https://opensource.org/about>
- Perrigo, B. (2023, Janvier 18). *Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic*. Retrieved from Time: <https://time.com/6247678/openai-chatgpt-kenya-workers/>

- Rapaport, W. (2023). Is Artificial General Intelligence Impossible? *University Of Buffalo*. Retrieved from: <https://cse.buffalo.edu/~rapaport/Papers/lands.pdf>
- Raymond, E. S. (2000). *The Cathedral and the Bazaar*. Retrieved from Récupéré sur catb: <http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/ar01s13.html>
- Ricoeur, P. (2021). Le pouvoir politique: fin du théologico-politique ? *Études théologiques et religieuses*, 96(4), 419–437. <https://doi.org/10.3917/etr.964.0419>
- Ross, L. (2023, November 21). *OpenAI CEO Sam Altman Returns to Company 4 Days after Being Fired*. Retrieved from The Wrap: <https://www.thewrap.com/openai-ceo-sam-altman-returns-to-company-4-days-after-being-fired/>
- Rouvroy, A. (2020, March 27). La gouvernementalité algorithmique et la mort du politique (G. E. Journal, Interviewer). Retrieved from: <https://www.greeneuropeanjournal.eu/la-gouvernementalite-algorithmique-et-la-mort-du-politique/>
- Shapiro, S. (2015). *No Speed Limit. Three Essays on Accelerationism*. Minneapolis: University of Minnesota Press.
- Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy & Public Affairs*, 1(3), 229–243. Retrieved from: <https://www.jstor.org/stable/2265052>
- Singer, P. (1997). The Drowning Child and the Expanding Circle. *New Internationalist*.
- Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations. *FACCT '23: the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 111–122. <https://dl.acm.org/doi/10.1145/3593013.3593981>
- Sowles, K. (2024). *Open State of Free and Open Source Funding*. Retrieved from Récupéré sur Fosdem 24: [https://fosdem.org/2024/events/attachments/fosdem-2024-2751-the-state-of-funding-free-open-source-software/slides/22659/State\\_of\\_FOSS\\_Funding\\_FOSDEM\\_2024\\_IQWHRvb.pdf](https://fosdem.org/2024/events/attachments/fosdem-2024-2751-the-state-of-funding-free-open-source-software/slides/22659/State_of_FOSS_Funding_FOSDEM_2024_IQWHRvb.pdf)
- Stallman, R. (n.d.a). *The GNU Project*. Retrieved from GNU Operating System: <https://www.gnu.org/gnu/thegnuproject.en.html>
- Stallman, R. (n.d.b). *Why Open Source Misses the Point of Free Software*. Retrieved from GNU Operating System: <https://www.gnu.org/philosophy/open-source-misses-the-point.en.html>
- Stallman, R. (n.d.c). *Floss and Foss*. Retrieved from GNU Operating System: <https://www.gnu.org/philosophy/floss-and-foss.en.html>
- Tidelift (2023). The 2023 Tidelift State of the Open Source Maintainer Survey. *Tidelift*. <https://tidelift.com/open-source-maintainer-survey-2023#form>
- Torres, É. P. (2023a, December 14). *'Effective Accelerationism' and the Pursuit of Cosmic Utopia*. Retrieved from Truthdig: <https://www.truthdig.com/articles/effective-accelerationism-and-the-pursuit-of-cosmic-utopia/>
- Torres, É. P. (2023b, June 15). *The Acronym behind Our Wildest AI Dreams and Nightmares*. Retrieved from Truthdig: <https://www.truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares/>
- United States Department of Justice (2006). *Direct Testimony of Jim Barksdale*. Retrieved from United States Department of Justice: <https://www.justice.gov/sites/default/files/atr/legacy/2006/05/16/1999.pdf>
- Waite, T. (2023, November 24). *Doomer vs Accelerationist: The Two Tribes Fighting for the Future of AI*. Retrieved from Dazed: <https://www.dazeddigital.com/life-culture/article/61411/1/doomer-vs-accelerationist-two-tribes-fighting-for-future-of-ai-openai-sam-altman>
- Wang, S., & Xu, S. (2024, March 21). *16 Changes to the Way Enterprises Are Building and Buying Generative AI*. Retrieved from a16z: <https://a16z.com/generative-ai-enterprise-2024/>
- Wider, D. G., Whittaker, M., & West, S. M. (2023). Open (For Business): Big TECH, Concentrated Power, and the Political Economy of Open AI. *SSRN*. Retrieved from Récupéré sur: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4543807#paper-citations-widget](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807#paper-citations-widget)
- Williams, A., & Srnicek, N. (2013, May 14). *#ACCELERATE MANIFESTO for an Accelerationist Politics*. Retrieved from Critical Legal Thinking: <https://criticallegalthinking.com/2013/05/14/accelerate-manifesto-for-an-accelerationist-politics/>

# AI DISRUPTIONS IN PHILANTHROPY

## A multi-scale model of ethical vigilance

*Charles Sellen and Joost Mönks*

### 1 Introduction and current context

Artificial Intelligence (AI) is not a single technology but a vast scientific, technological, and industrial domain whose sweeping scope encompasses multiple ramifications and numerous applications. Coined around the mid-20th century, the concept of AI is defined nowadays as “the ability of a machine to display human-like capabilities such as reasoning, learning, planning, and creativity” (European Parliament, 2020).

After an incubation and research phase of about half a century at some of the world’s most prestigious universities like the Massachusetts Institute of Technology and Stanford University, preparing the AI revolution of today, AI tech is now widespread and increasingly ubiquitous in lives and societies. The general public uses it daily (yet often unknowingly) in every query entered on Internet search engines when listening to music or buying train tickets online. As of now, AI tools’ functionalities and use cases have already reached sophisticated levels in various domains including the functionality to read, translate, recognize faces, objects, and emotions, move autonomous vehicles, or track infectious diseases. Recently, the improved version of an AI-powered chatbot called “Chat Generative Pre-trained Transformer” (abbreviated “ChatGPT 3.5”) was released on 30 November 2022 and marketed extensively to the general public. This game-changer app reached 100 million users in just two months, setting a world record among online services—by comparison, it took five years for Twitter to cross this milestone (Duarte, 2024). Apart from a handful of sizable countries (including China, Iran, and Russia) where ChatGPT is currently unavailable, internauts from around the planet can now freely access this conversational tool and get immediate answers to their queries, spanning a wide array of topics. The answers provided by this “generative AI” tool are so elaborate that they seem produced by human intelligence, up to the point that universities are now confronted with an exacerbated problem: it has suddenly become more difficult than ever to detect plagiarism cases! In nearly all fields—from the humanities to hard sciences—it has become increasingly difficult to distinguish between human-written texts and machine-generated materials.

ChatGPT’s and other chatbots’ astounding abilities might be the tip of an iceberg of upcoming upheavals. According to a report by top consulting firm McKinsey & Company (2023), “generative AI is poised to transform roles and boost performance across functions such as sales and

marketing, customer operations, and software development. In the process, it could unlock trillions of dollars in value across sectors.” Albeit impressive, Natural Language Processing and its resulting Large Language Models (LLM) are yet merely one of the numerous fields of AI application, whose range appears potentially infinite. Due to the use of AI tools that have been developed, we are thus on the verge of quickly and profoundly transforming known mechanisms of society, simultaneously allowing for unprecedented advances in the human condition and giving rise to profound new challenges, foremost in the widening of economic inequality and the exacerbation of the digital divide, while raising increasing controversy (Pazzanese, 2020) around its use, for instance in facial recognition, biases in automated decision systems, or privacy breaches in tracking the spread of viruses. But AI is just a component of a much broader ongoing transformation of our world that decision-makers are well aware of. In an official address, the United Nations Secretary-General, António Guterres, recognized publicly that “Looking to the future, two seismic shifts will shape the 21st century: the climate crisis, and digital transformation. Both could widen inequalities even further” (United Nations, 2020).

In a nutshell, the scale, spread, speed, and reach of change brought about by AI are viewed as unprecedented in human history (United Nations, 2019). It has been compared to the introduction of the printing press. But while it took over 50 years for printed books to become widely available across Europe, ChatGPT reached 100 million users in just two months. This leads to significant and pressing debates in philanthropy<sup>1</sup> and beyond the borders of the philanthropic sector: How do we ensure that the benefits of AI’s use outweigh the costs and the associated risks? And how do we create trust that the technology is built and used according to ethical principles on the one hand, and in an inclusive and human-centered way on the other? In other words, how can we encourage and govern the development of *Ethical and Inclusive AI* (EIAI) for the common good, and what role could philanthropy and nonprofit organizations play? And specifically, in tomorrow’s AI-empowered world, how can we best maintain and reinforce what makes us human, namely our unique ability to think ethically, discern, and act with our hearts (Mönks & Sellen, 2020)? And how can philanthropy, in its original Greek meaning of “love of humanity” (Sulek, 2010b), promote value or compassion-based perspectives to shape and inform public policy spaces to design and create an inclusive digital future?

## **2 The critical need for ethical discernment tools about AI in philanthropy**

Modern times philanthropy is commonly defined in academia as “voluntary action for the public good” (Payton & Moody, 2008) or as “private means to public ends” (Sulek, 2010a). In this chapter, we propose that philanthropic organizations (POs)—defined as institutions driven by moral values using private resources to promote the public good and the wellbeing of humanity—can, and should, ensure that the technological potential of AI is translated into innovative solutions for positive social transformation while anticipating possible future trends and impacts of AI by conducting and enabling strategic monitoring. Concurrently, POs can promote and champion ethical standards required to minimize negative externalities and protect and give voice to the most vulnerable (Mönks & Sellen, 2020). Put differently, the sector should not only establish what AI can do for philanthropy but also what philanthropy can do to promote EIAI and its balanced governance (Gill & Germann, 2021), building on its unique and independent position at the intersection of industry, government, and academia.

Despite the crucial role that philanthropy—and more broadly, civil society—could play, POs have largely been absent from the global and national debates and agendas on AI (Mönks & Sellen, 2020). This sector is, surprisingly, still “AI-illiterate” and, therefore, has not only been able to

establish itself as an ethical reference point for the promotion and governance of EIAI, but it is also not embracing the advances that AI technologies could bring to POs' operations, and seems to be awakening only recently (Dhar & Firth-Butterfield, 2021).

While many philanthropists and nonprofit leaders do not yet pay due attention to AI, others seem fascinated by the novelty and follow the trend enthusiastically without always sufficiently asking or being aware of the ethical questions. While it is true that AI tools bring exciting perspectives of reducing costs, saving time, amplifying the scope of nonprofit operations, and reaching out to more potential beneficiaries or supporters, the philanthropic sector has long remained silent or unconcerned about AI. In recent years, only a few pioneers dared to voice concerns, ask tough questions, sound the alarm, and call for strong ethical safeguards (Bernholz, 2019; Mönks & Sellen, 2020). In 2022–2023, we have seen a sudden wake-up movement, and AI has appeared as a top priority on the sector's radar, with many conferences and talks about the urgency for philanthropy to position itself.

The energy it requires for a layperson to apprehend the intricacies of AI tech is high, not to mention anticipating its uncertain effects on philanthropy in the long run. Even with goodwill and curiosity, one may quickly feel overwhelmed when approaching the vast and crowded field of AI. To lower the “cost of entry” of accessing this knowledge for an “ordinary” philanthropist or nonprofit leader, we propose in this chapter a simple model that these decision-makers (who are not AI experts) may use to discern which types of ethical challenges they could be confronted with, and how they might respond by, first, identifying the levels of impact that AI tech may have. Indeed, some effects occur at a small scale while others have much broader consequences ranging from the micro-organizational level to the macro-overall society level. Let us take three illustrative examples.

On a first scale (or level of analysis), a nonprofit organization may want to use “precision philanthropy” (Chappell, 2018) fundraising tools to predict donors' affinity and inclination to give in real time and thus solicit donations in a more cost-efficient way. We call this scale the “micro-level of analysis” (see Figure 31.1).

On a second scale (“meso-level” in Figure 31.1), nonprofits immersed in a sector-wide collaborative environment may wish to harness the power of AI to accelerate the deployment of AI solutions while being aware of challenges and risks associated with the deployment of AI tech. For instance, Wikipedia is an encyclopedia project led by the Wikimedia Foundation to disseminate free knowledge in more than 300 languages. The irruption of generative AI presents opportunities

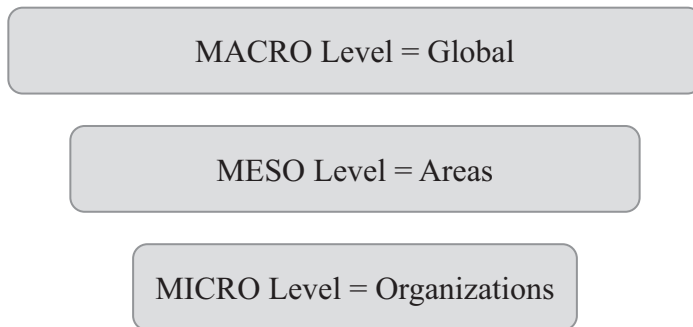


Figure 31.1 Three levels to apprehend AI in philanthropy.

Source: Authors.

to speed the writing process of online articles and significant risks of publishing counterfeit materials or plausible falsehoods (Gertner, 2023). Volunteer editors (called “Wikipedians”) are already overwhelmed by the myriads of automatically proposed content that is impossible for them to approve one by one manually. For that reason, they are considering banning (or drastically limiting) AI from the platform although they recognize the value of AI’s potential and use it appropriately in several Wikimedia projects.

On a third scale (“macro-level” in Figure 31.1), to illustrate one of the most burning ethical issues facing us and where philanthropy is raising its voice concerns AI-enabled “Lethal autonomous weapons systems” (LAWS) whose functioning may sound like science fiction but is increasingly real. These rapidly proliferating weapons entail mind-boggling ethical and legal puzzles, especially regarding accountability (Longpre et al., 2022). Who would indeed be responsible for a strike decided autonomously by a machine on a battlefield? The absence of human judgment to pull the trigger and make life-or-death decisions leads humanitarian organizations to sound the alarm and ask for an urgent ban before international treaties can be negotiated and regulations enforced (ICRC, 2022). Along with spiritual leaders, nonprofits are the spearhead of these moral debates by raising questions about the step three analysis and considering safeguards from step four in our proposed model (see Figure 31.2), while governments and private firms focus on their vested interests (Peltz, 2023). In the current absence of ratified international treaties or regulations to control and limit the excesses of this new breed of autonomous weaponry, it is precisely in this kind of global debates and arenas that POs can bring in value-driven approaches and unheard voices. Beyond just raising awareness about “killer robots,” nonprofits may even suggest a workable agenda for action like Human Rights Watch did in partnership with the Harvard Law School’s International Human Rights Clinic (Human Rights Watch [HRW] & International Human Rights Clinic [IHRC], 2022). HRW and a group of international NGOs have teamed up to build a non-profit coalition and launch the worldwide campaign “Stop Killer Robots.”

As depicted by these three examples, and to simplify, we suggest distinguishing between three main levels of analysis: micro, meso, and macro levels (see Figure 31.1), which represent respectively the philanthropic organization, the area in which it operates, and the society at large in which it is embedded.

This first way of understanding the levels of possible disruption and opportunity of AI tech can then be combined with several more applied steps of reasoning as described in Figure 31.2. These identify the key questions as they specifically relate to the various stakeholders of POs, the ethical questions that are raised from a philanthropical point of view, and finally the identification of possible safeguards and best practices to promote the ethical and inclusive development and use of AI tech for the common good. These consecutive and complementary steps are proposed to guide and support PO as they face the AI revolution.

By combining the two approaches described above, starting with the understanding of the levels of impact (see Figure 31.1), followed by reasoning steps about stakeholders, ethics, and good practice (see Figure 31.2), we obtain a basic multi-scale reading grid presented as a table in



Figure 31.2 Four key reasoning steps to apprehend AI in philanthropy.

Source: Authors.

	Levels of Disruptions	→ Impacted stakeholders	→ Ethical questions raised	→ Best practices & Safeguards
3	MACRO Society at large (in which POs are embedded)	Q?	Questions?	Q?
2	MESO Areas (themes/fields of POs operations)	Q?	Q?	Q?
1	MICRO Philanthropic Organizations (POs operations and funding)	Q?	Q?	Q?

Figure 31.3 Multi-scale model to apprehend AI in philanthropy.

Source: Authors.

Figure 31.3, where each “box” suggests key questions (depicted by question marks) that philanthropic practitioners are invited to ask themselves.

While philanthropists and nonprofit leaders are busy running their POs, focused on serving beneficiaries, and cannot be expected to become state-of-the-art AI specialists, the purpose of our matrix is to equip them at least with a handy and functional reading grid. Through the lenses of our proposed model, philanthropy professionals could quickly decipher their fast-changing and AI-driven changing environment. Its practical use would require them first to identify in which “box” of the table the issue they are facing might be classified and then ask themselves the related questions. For example: “Who are the impacted stakeholders of my organization, and how does the use of AI technology affect them?” Or: “Which ethical questions are raised in my specialty or concentration (for instance: Education, Health, Farming, etc.) due to the irruption of AI tools?” Or: “Which best practices could be adopted and implemented in society to strengthen philanthropy globally through the use of AI and how to mitigate the associated systemic risks?” And so on.

Among the emerging body of research and literature on AI and philanthropy, and the already crowded space around AI, the actual and precise risks and opportunities may soon become quite complex to assess for philanthropy leaders, considering the sheer amount of information and diverse perspectives available alone. This is why our four-step and three-level model—admittedly a bit rudimentary but easily applicable—could be helpful in discerning which aspects of AI are specifically promising or problematic. Philanthropy observers, analysts, scholars, and practitioners alike may use this potentially universal tool to share viewpoints legibly and seek solutions together efficiently to set limits and establish safeguards. Furthermore, the straightforward use of this basic model could help the philanthropic sector stay informed on the development of AI innovation. The proposed model is simple and does not pretend to cover the variety of scenarios and heterogeneity of actors involved in a fast-changing technological environment. Still, it dissects the essential components of AI adoption, allowing nonprofit practitioners to unpack and address the key questions they are facing from a PO perspective.

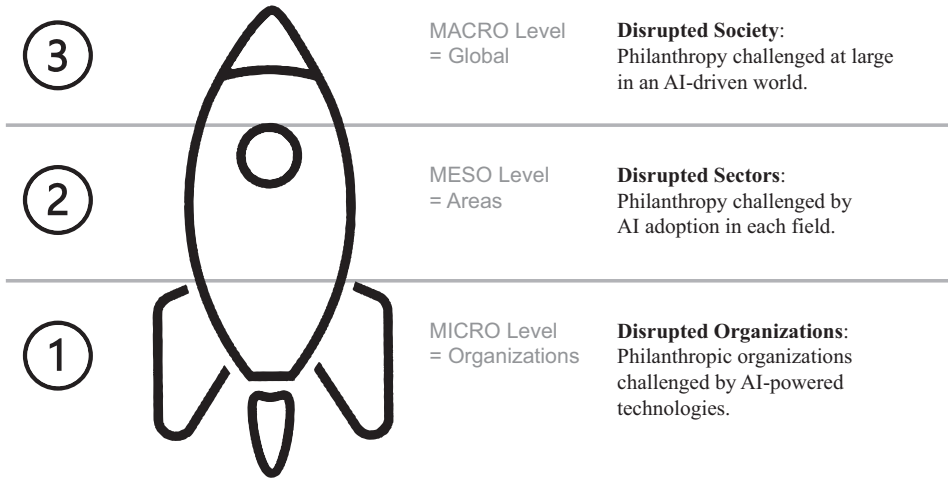


Figure 31.4 Different scales of disruptions brought by AI in philanthropy.

Source: Authors.

### 3 AI-induced disruption and opportunity for philanthropy—understanding the broad picture (Step 1)

In this section, we further develop the three main levels of disruption and opportunity that AI may bring in the philanthropic field: at the micro-level, AI may empower and challenge POs in their operations (Subsection 3.1); at the meso-level, AI offers opportunity and challenge in reaching the SDGs in various thematic areas and sectors in which POs operate and fund projects (Subsection 3.2); and at the macro-level AI tends to potentially unsettle and challenge society at large (Subsection 3.3) and raises numerous ethical questions. To simplify the memorization of the model, we use a rocket to illustrate those three stages of analysis visually (see Figure 31.4).

#### 3.1 Micro-level: AI bringing possible disruptions and opportunities at an organization’s level

The first and immediate effects of disruption<sup>2</sup> brought about by the introduction of AI technology in a philanthropic leader’s daily life generally occur at the operational level of the nonprofit organization. The leaders of such POs are likely to witness rapid and radical changes in how an organization is run. Those changes may be actively sought or passively incurred, but in either case, they will profoundly transform how the job was performed previously. Suppose this senior leader is 55 years old and started to work in the mid-1990s. AI would certainly not be the first technological innovation they would see coming in their professional life. To consider the fundraising function of their portfolio of assignments, they would have learned during their studies about how personalized phone calls increased donations in the 1980s or about the efficient utilization of TV shows to raise charitable funds in the 1990s (think of the numerous annual Telethons organized since this era). In their early career, they would have witnessed the boom of the Internet in the 2000s, including donations via targeted website advertisements or through e-mail-based donor recruitment strategies. They would have observed the advent of donations via text messaging (SMS) and mobile payment apps in the 2010s and also—if their skills are up to date—probably noticed



the rise of crypto-donations and non-fungible tokens (NFTs) in the early 2020s. Nevertheless, the novelty of AI is not another simple wave made of a single technology that fits in a relatively stable, seemingly everlasting fundraising scheme. It is a complete tsunami whereby the very essence of fundraising is undergoing deep transformation. While other technologies (phone, TV, Internet, e-mail, SMS, crypto, NFTs, etc.) added new strings to the fundraising bow, AI is changing what a bow looks like. Maybe at some point in the future, AI could even lead us to practice archery without a bow to extend the metaphor.

POs of all sizes are, therefore, prone to be hugely impacted by the emergence of AI in their day-to-day operations. While some of the biggest foundations may have the data to train AI and the resources required to develop their infrastructure, smaller philanthropies will likely be unable to do so. However, they should not be left out of exploring AI's potential use and added value. Regarding the fundraising function, this includes, for instance, using AI to analyze, inform, and predict donor behavior. In the organization's other aspects, AI may help monitor and evaluate impact and offer applications for automating administrative tasks. Essentially, it is about saving time and therefore money. But AI also brings greater precision and reliability in some ways while creating other challenges regarding ethics and possible new biases.

To stick with our case study of the fundraising function, Nandeshwar (2023) has identified at least six applications leveraging AI's astounding capabilities for nonprofits:

- 1 **Giving likelihood prediction** (“by studying the donation history of donors in a dataset, AI algorithms can predict the likelihood of future donations”);
- 2 **Segmentation**, based on grouping (or “clustering”) current donors or prospects according to similarities in their profiles (“giving likelihood models can be combined with segmentation to help with donor retention or upgrades”);
- 3 **Text or language generation** via automated tools like ChatGPT that “can be used to generate qualification emails, thank you or stewardship letters, proposal generation, and research profile creation”;
- 4 **Content generation**, such as images, videos, or music, that will translate into “graphic generators for communications, personalized stewardship videos, or dynamic proposals”;
- 5 **Augmented and virtual reality** (AR/VR) that “can also be used in fundraising to transport prospective donors into an imaginary world to show the impact of their gifts”;
- 6 **Data science**, which “can be used to stitch together unique fundraising solutions by analyzing large datasets to identify trends, insights, and opportunities for optimization.”

While there is a growing stream of literature about opportunities specifically offered by AI to leverage private donations for charitable causes through so-called “AI4Giving” (Kanter & Fine, 2020), specific academic research and reference in specialized media on the various levels of use and impact of AI for philanthropy are still scant, reflecting the sector's continued apparent lack of awareness of the topic. Nandeshwar and Jewell (2023) offer real-time examples of AI uses applied to fundraising by several existing nonprofits. For instance, the World Wildlife Fund (WWF) uses an AI-powered chatbot on Facebook Messenger to engage its audience in meaningful conversations and potentially quicken their onboarding as new supporters of their cause. Likewise, the National Geographic Society uses AI-powered image recognition “to easily search and retrieve images for use in fundraising campaigns and other outreach efforts, making the appeals more relevant and compelling while reducing staff time searching for images” (Nandeshwar & Jewell, 2023). Beyond asking for money gifts, to further engage donors by deeply understanding their feelings, some NGOs use “sentiment analysis” to categorize text as positive/negative/neutral,

like Amnesty International’s “Troll Patrol” that detects online abusive comments automatically through a home-made algorithm built with the help of myriads of digital activists (Amnesty International, 2018; Nandeshwar & Jewell, 2023).

Other AI applications used in the core processes of POs include *Donor Matching* using Machine Learning (ML) and behavioral data to match donors with a cause to support; *Philanthropic Advising* based on ML and philanthropic algorithms, dubbed “philgorithms” (Davies, 2017), to provide recommendations for high-yield philanthropic investment; *Online Fundraising Campaigns* aiming at personalizing donor engagement by analyzing donor databases and third-party sources, including the use of chatbots; and *Reporting and Workflow Tools* to generate reports or automate administrative tasks. Other applications focus on the power of AI to analyze *big data* and the set-up of *data collaboratives* to develop new insights and use cases for Monitoring & Evaluation, and impact assessment, among others.

Although AI is frequently poorly understood in terms of how it works and what it can offer to the sector, some POs already use it purposefully, as we have shown. But in real life, POs are not atomized and isolated. They operate in a given field, an area of giving and focus where they usually team up or collaborate (and sometimes compete) with peer organizations. Therefore, we should ask ourselves what the disruptive effects of AI could be at this second sector-wide scale of analysis.

### **3.2 Meso-level: AI bringing innovation at sectoral levels**

Beyond the fate of a single organization, AI has a growing number of applications in various broader areas, including education, health, farming, wildlife conservation, climate change, inequalities, and the humanitarian sectors, to name a few. So perhaps the real question for philanthropy, as watchful experts warned us early on, is not so much how POs will leverage the power of AI in their day-to-day operations, but rather “how is AI being used within the domains within which they work and how they must respond” (Bernholz, 2019).

The advent of AI and its dissemination in those many domains lay the foundation of the now-established field of “AI for Social Good” (AI4SG) or *AI4SDGs*, which promote technological solutions to societal challenges and the realization of the Sustainable Development Goals (SDGs) in each of the 17 key areas with specific targets that were defined for the period 2015–2030 under the auspices of the United Nations (Tomašev, et al., 2020). It also involves growing initiatives around data collaboratives and infrastructures for global AI-based use cases. This field already has its annual “AI for Good Summit,” hosted by the International Telecommunication Union in Geneva (ITU, 2023). It has a growing number of use cases in development and humanitarian settings, as promoted by the UN Global Pulse initiative, the Secretary-General’s Innovation Lab (United Nations, 2023b). It includes, for instance, early warning systems building on the predictive power of ML algorithms enabling efficient rapid response, cash transfers to the extremely poor via mobile telephony, using satellite imagery to predict poverty, or using ML to analyze large volumes of data to forecast the outbreak and spread of infectious diseases. It also includes a growing number of initiatives around data philanthropy and data collaboratives for social impact, which occurs when organizations share and utilize data to solve social issues and support the public domain. Commentators have termed this trend “data philanthropy” (Harvard Business School [HBS], 2016). However, this new realm of “data for good” does not come without risks, namely ensuring data privacy and protecting the fundamental rights of citizens. POs and international development organizations working toward the common good must, therefore, “consider the legal and ethical risks inherent in data sharing agreements, beyond privacy risks,” to remain consistent with their core values (United Nations Development Programme [UNDP], 2020).

### 3.3. Macro-level: AI posing society-wide challenges

Beyond individual organizations and beyond the areas where they operate, the rapid advent of AI tends to disrupt and challenge society at large. The philanthropic sector is affected by a complete overhaul of the society in which it is embedded, with ripple effects on public institutions, shared (or not-so-shared) values, and trust between citizens and political decision-makers. For the understanding and building of our functional AI grid, it is important to note that this third level raises the question of POs' role in shaping global debates and how our model can contribute toward the positive use of technology and prevent and mitigate its negative impact on society.

This higher level more broadly reflects how POs can (and must) respond to a growing global call for the development of AI to be based on principles that are geared toward inclusiveness, accessibility, and cooperation, which reduce the digital divide and inequality, which are based on the primacy of human agency and capabilities, and which also seek to promote the principle of value or compassion-based perspectives in the public policy spaces for the definition of our digital future. An illustration is the recent policy pack issued by the Civil 20 working group of the G20 India 2023 Summit (G20 India, 2023). This initiative emanating from civil society leaders and convening grassroots movements more broadly resonates with the ambition of the UN Global Digital Compact (United Nations, 2023a), which is expected to “outline shared principles for an open, free and secure digital future for all” and be agreed at the Summit of the Future in 2024 at the United Nations. As part of a worldwide consultative process, POs are openly invited to contribute to shaping this global compact via an online platform collecting propositions and feedback. Similarly, UNESCO has led international efforts to ensure that AI is developed with strong ethical guardrails. This international organization dedicated to Education, Science, and Culture has delivered global standards to maximize the benefits of scientific discoveries while minimizing the downside risks, ensuring they contribute to a more inclusive, sustainable, and peaceful world. These guidelines, namely, invite Member States to engage all stakeholders, including POs and civil society, “... so that the development and use of AI technologies are guided by both sound scientific research as well as ethical analysis and evaluation” (UNESCO, 2022, p.7).

POs should, therefore, be at the forefront of these debates and discussions within the sectors in which they operate, as well as in relation to global architecture. The sector needs to prepare itself for and position itself in a world powered by AI (Dhar & Firth-Butterfield, 2021) and should commit to learning and mainstreaming the question of how AI technologies affect their key focus areas and, beyond these sub-fields, anticipate future trends of an AI-driven world.

\*

To visually summarize Section 3, we propose to use the additional rocket with three stages depicting the three main levels (micro, meso, and macro) of disruptions generated by AI and affecting philanthropic players (see Figure 31.4). In using our model to assess AI, the first step is to determine at which level the reflection should take place and then to unfold the subsequent reasoning steps: Who are the impacted stakeholders? (Section 4); What are the corresponding ethical questions to ask? (Section 5); What types of safeguards may we implement at this level? (Section 6).

## 4 Stakeholder relations impacted by AI in philanthropy (Step 2)

Pursuing the proposed four reasoning steps (see Figure 31.2) and having understood the three main levels of impact of AI (see Figure 31.1), this section focuses on the effect of AI innovation on the relations that POs have with the diverse stakeholders they support and work with. These relations stand at the core of PO's *raison d'être* and engagement. Following the proposed matrix, we will look at the stakeholders of a given organization (microscale—subsection 4.1), at the stakeholders

of a given area or sector (mesoscale—subsection 4.2), and finally at the multiple stakeholders of an entire society (macroscale—subsection 4.3).

Since this current section, as well as following Sections 5 and 6, are intended to serve a more applied purpose, they are structured around a set of proposed key questions that we believe are key and which may serve as a first “checklist box” (see Box 2.1) of issues that should be addressed by POs and that can guide their IA strategy development.

We summarize visually the three main levels of some foreseeable impacts on stakeholders in Figure 31.5.

### **Box 31.1 Stakeholders checklist**

- 4.1. Micro-level: AI’s impact on stakeholders at an organization’s level  
Adopting AI at an organization’s level, willingly or unwillingly, will logically entail a series of impacts on this organization’s stakeholders. For a regular nonprofit, these are likely to be:
  - 1 the beneficiaries, who might be found, reached, and served differently;
  - 2 the donors, who may be identified, solicited, and thanked differently;
  - 3 the volunteers, who could be recruited, encouraged, and supervised differently;
  - 4 the staff, who will probably be hired, trained, and managed differently; and
  - 5 the directors (Board Members), who may be approached and steered differently.
- 4.2. Meso-level: AI’s impact on stakeholders at sectoral levels  
Within a specific subsector, POs generally know each other and take part in the same networks, coalitions, alliances, etc. The effect of AI on stakeholder relations is likely to bear impacts on:
  - 1 How POs relate to their nonprofit peers, especially if significant gaps in AI adoption or understanding widen and trigger dissensions or discrepancies among similar organizations who previously worked together harmoniously;
  - 2 the strategic partners (public or private) with whom they build long-lasting ties;
  - 3 the networks to which they belong; and
  - 4 the sectoral expertise upon which they rely.
- 4.3. Macro-level: AI’s impact on stakeholders at society’s level  
In societies at large, whether national or international societies, our model leads users to consider possible impacts on several categories of stakeholders, including:
  - 1 Governments and policymakers who must regulate AI developments and uses;
  - 2 Industrialists who constantly innovate in technologies and push for broad AI adoption;
  - 3 Policy analysts and researchers (think tanks, scholars) who have time to observe and think, and should anticipate AI’s potential effects and help set appropriate boundaries;
  - 4 AI developers who design and provide data for AI tools need human-centered and ethical design principles as foundation for their work;
  - 5 Citizens (and, more broadly, AI users) who are largely subject to this industrial revolution and should have a say and who, through their own consumption and use of AI technology, contribute to its ethical use.

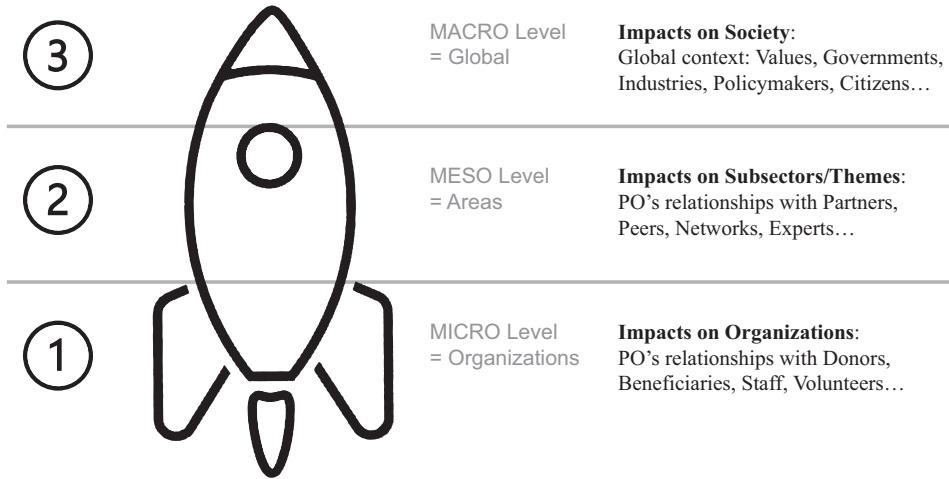


Figure 31.5 Several levels of impacts on stakeholders caused by AI in philanthropy.

Source: Authors.

### 5 AI-related ethical vigilance for philanthropy (Step 3)

POs may find themselves challenged in different ways by AI tech, both for the better and the worse, and face many ethical questions. The issue of the ethical frameworks governing AI remains a disputed area (Gill & Germann, 2021). POs can and should challenge and improve AI governance frameworks by providing essential input and ethical guidance on several key levels. Ethics constitute a vast field of research spanning every area of knowledge and practice. To simplify, scholars usually distinguish between *applied ethics* (i.e., everyday situations where practitioners face real-world dilemmas) and *normative ethics* (i.e., situations that can be theorized). The latter is usually subdivided into three main categories: “*consequentialism*” focuses on results regardless of intentions; “*deontology*” stresses the “right thing to do” regardless of consequences; and “*virtue ethics*” underlines the special character or motives of an agent. Many other forms of ethical approaches with nuances may be found in scientific literature. For fundraising alone, up to 14 theories have been identified (MacQuillin, 2022). In this section, parallel to the three main levels of disruption brought by AI into philanthropy, we distinguish mirror symmetry between three main scopes of ethical vigilance (at micro, meso, and macro levels) for philanthropy players toward AI. For the sake of brevity, these scopes are only briefly presented here for the completion of the analytical grid by raising key ethical questions for the functional and practical use of the proposed model in its next step (Step 3 in Figure 31.2). In Box 31.2, we suggest a few key questions and issues we see to guide understanding and action practically:

To visually summarize Section 5, our three-stage rocket depicts the main levels (micro, meso, and macro) of ethical concerns that all philanthropic practitioners and decision-makers should consider when reflecting on AI and its potential consequences (see Figure 31.6).

### **Box 31.2 Ethics checklist**

- 5.1. Micro-level: Ethical questions raised on AI at an organization's level  
While POs should be encouraged to embrace innovations offered by AI, this adoption should be carefully considered and not be made at all costs, particularly ethical costs. The organization's fundamental values should remain at the center and never be surrendered. Several essential questions open to internal debates might include:
  - 1 Do AI tools introduce new biases in our operations?
  - 2 How do we ensure data privacy and promote data sovereignty?
  - 3 Do we sacrifice our ethics to reach higher performance through AI?
  - 4 How do we use and guide AI for ethical grantmaking?
  
- 5.2. Meso-level: Ethical questions raised on AI at sectoral levels  
While some POs will be more enthusiastic than others about AI, these early adopters could encourage their peers to take a first step by reassuring them on ethical grounds as well. Those at the forefront could ask these crucial questions within sector-wide discussion groups:
  - 1 How do we leverage AI ethically to achieve the SDGs in each area?
  - 2 Will AI increase cooperation or competition among POs?
  - 3 How can we build common infrastructures and data commons?
  - 4 How do we ensure ethical and transparent algorithms?
  
- 5.3. Macro-level: Ethical questions raised on AI at society's level  
POs can raise awareness through advocacy and propose innovative solutions to design a fair and just global framework for the safe development of AI. Some of the vital questions they could and should raise at a global level are:
  - 1 Are we ready to let autonomous AI act by itself without control?
  - 2 How do we offset AI's potential collateral damages?
  - 3 How do we ensure that the AI era makes a voice for the vulnerable and unheard?
  - 4 What human-centered ethical standards can we set?

## **6 Best practices and safeguards for AI in philanthropy (Step 4)**

Building on the ethical interrogations presented above, we are now in our final fourth step, in search of possible solutions, best practices, and safeguards that can be implemented to counterbalance the excesses of AI and harness its full potential. A wide array of potential solutions exist, and various examples may be identified and further developed to tackle each level of our three-stage rocket model (see Figure 31.7). In Box 31.3, we suggest a few key considerations we see as important for a first checklist.

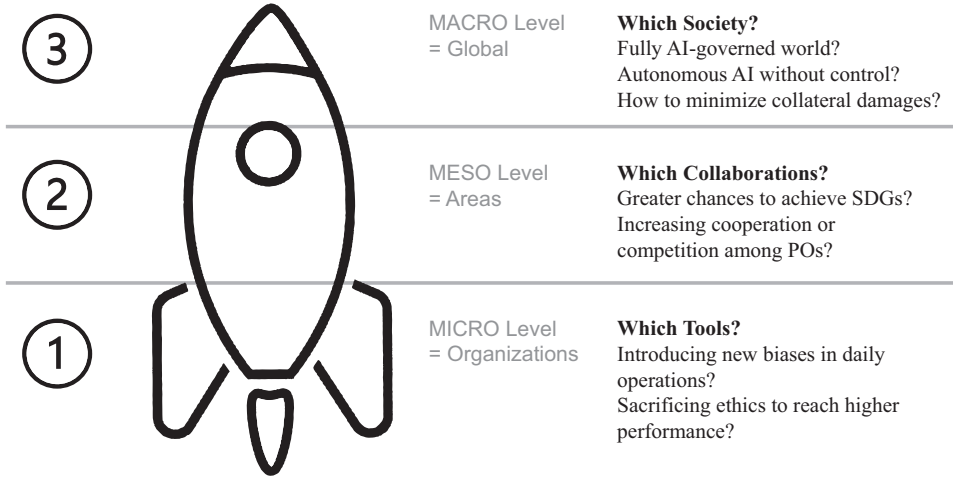


Figure 31.6 Several levels of ethical questions raised by AI in philanthropy.

Source: Authors.

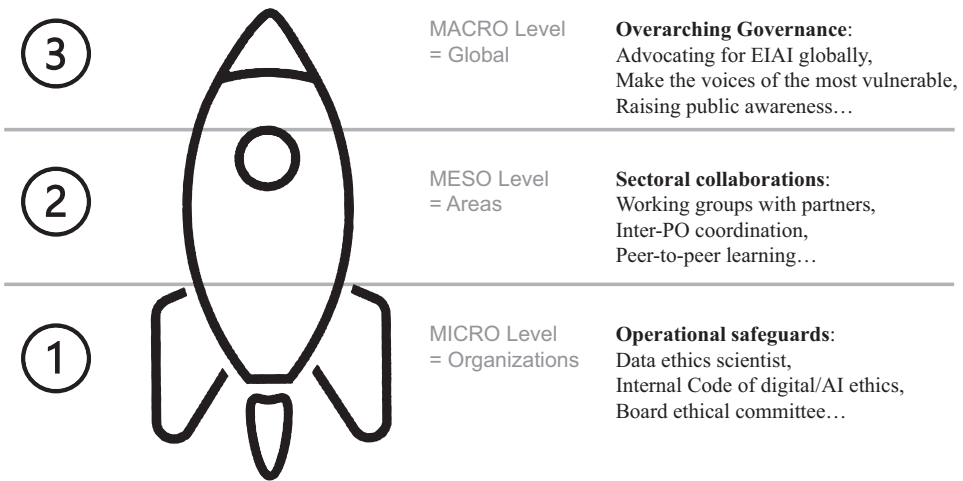


Figure 31.7 Several levels of best practice and safeguards to properly handle AI in philanthropy.

Source: Authors.

### Box 31.3 Best practice checklist

- 6.1. Micro-level: Best practices to use EIAl at an organization's level

We have seen earlier that it is possible to use AI as a helpful tool for philanthropic operations to better serve beneficiaries while preserving the organization's core ethical principles. For this to happen safely, a given PO would need to:

- 1 Hire (or train) at least one “Data scientist” with a strong ethical background to elaborate a “Data policy” that would guide the organization’s proper handling of data;
  - 2 Draft and adopt officially a “Code of digital/AI ethics”;
  - 3 Establish an “Ethics Committee” within the Board of Directors.
- 6.2. Meso-level: Best practices to use EIAI at sectoral levels  
Leveraging AI as a vector and an accelerator to achieve the SDGs by 2030, while respecting strong ethical standards within sectors, implies these types of interactions:
    - 1 Sector-wide working groups with partners;
    - 2 Inter-PO coordination;
    - 3 Peer-to-peer learning;
    - 4 Data commons.
  - 6.3. Macro-level: Best practices to use EIAI at society’s level  
Promoting a healthy and equitable global architecture of AI requires that numerous POs come together and coalesce to pursue these types of collective actions:
    - 1 Advocating for EIAI in global governance;
    - 2 Raising general awareness by informing (or alerting) citizens and public opinion;
    - 3 Questioning politics, policies, and politicians;
    - 4 Scrutinizing industrial agendas among AI tech behemoths;
    - 5 Demanding accountability from public and private decision-makers.

## **7 Summary: a simple model to help philanthropy navigate AI complexity and opportunity**

In a nutshell and synthesizing the three levels of analysis combined with four reasoning steps (gradually designed in Figures 31.1–31.7), the overall matrix proposed and described below may serve as a convenient and simple tool for philanthropic decision-makers to reflect strategically and opt for the best ethical choices at each level and operating stage. In particular:

- 1 At the micro-level, POs may adopt AI-powered technologies as useful tools to gain efficiency in their day-to-day core operations while ensuring this adoption remains in line with their core organizational ethics (level 1);
- 2 At the meso-level, POs may leverage AI in the thematic areas and sectors in which they operate and which they support to achieve the SDGs while making sure they are ethically aligned and coordinated with their peers (level 2);
- 3 At the macro-level, POs may play a leading role in raising awareness about the need for overarching governance and strong ethical frameworks to ensure AI’s global ethics and inclusiveness and AI peacekeeping in a tense geopolitical context (level 3).

A recapitulative table presenting the model in detail is shown in Figure 31.8.

Given the density and technicality of AI-related issues, we propose that every philanthropist or nonprofit leader could use this simple three-stage model to decipher the stakes, assess in real



	Scale of Disruptions	Impacted stakeholders	Ethical questions raised	Best practices & Safeguards
1	Philanthropic Organizations (POs operations and funding)	<ul style="list-style-type: none"> <li>• Beneficiaries</li> <li>• Donors</li> <li>• Staff</li> <li>• Volunteers</li> </ul>	<ul style="list-style-type: none"> <li>• Do AI tools introduce new biases in our operations?</li> <li>• Do we sacrifice our ethics to reach higher performance through AI?</li> </ul>	<ul style="list-style-type: none"> <li>• Data ethics scientist</li> <li>• Code of digital/AI ethics</li> <li>• Board ethical committee</li> <li>• [other...]</li> </ul>
2	Areas (themes/fields of POs operations)	<ul style="list-style-type: none"> <li>• Networks</li> <li>• Experts</li> <li>• Partners</li> <li>• Peers</li> </ul>	<ul style="list-style-type: none"> <li>• How to leverage AI ethically to achieve the SDGs in each area?</li> <li>• Will AI increase cooperation or competition among POs?</li> </ul>	<ul style="list-style-type: none"> <li>• Sector-wide working groups with partners</li> <li>• Inter-PO coordination</li> <li>• Peer-to-peer learning</li> <li>• Data commons</li> <li>• [other...]</li> </ul>
3	Society at large (in which POs are embedded)	<ul style="list-style-type: none"> <li>• Government</li> <li>• Industries</li> <li>• Policies</li> <li>• Citizens</li> </ul>	<ul style="list-style-type: none"> <li>• Do we want a fully AI-governed world?</li> <li>• Are we ready to let autonomous AI act by itself without control?</li> <li>• How do we offset AI's potential collateral damages?</li> </ul>	<ul style="list-style-type: none"> <li>• Advocating for EIAI in global governance</li> <li>• Raising awareness (public opinion, politics)</li> <li>• [other...]</li> </ul>

Figure 31.8 Multi-scale model to apprehend AI in philanthropy (with suggested use).

Source: Authors.

time the risks and opportunities of AI in running POs, and discern their potential contribution as philanthropic practitioners to making AI a beneficial force for humankind.

### 8 Conclusion: facing forward through ethical lenses

AI infuses society and affects the people and communities traditionally served by the philanthropic sector. How philanthropic grant-makers use and support organizations working in a world and society where digital technology pervades will be crucial to our civil society and democracy. Suppose POs fail to position themselves, particularly regarding the ethical dimensions of AI, and fail to adapt; they risk becoming irrelevant in their traditional role of producers of public goods and social cohesion. As one observer puts it, “*philanthropists should treat AI as an ethical not a technological challenge*” because

even if the machines are not going to kill us, there are plenty of reasons to worry AI will be used for ill as well as for good, and that advances in the field are coming faster than our ability to think through the consequences.

(Foley, 2019)

Curiously and as said, the philanthropic sector has only been recently engaging on the topic after having been largely absent from the global debates around AI and the flurry of ethical frameworks (Hagendorff, 2020; Mönks & Sellen, 2020) and guidelines that a variety of actors over the last couple of years have developed.

Despite an apparent agreement in many of these frameworks that AI should be “ethical” and “inclusive,” debate exists around what constitutes Ethical and Inclusive AI (EIAI). However, it remains tricky so far to clearly distinguish the specific operational requirements, technical standards,

and best practices needed for its realization. Beyond politically correct (and seducing) keywords and amid the flourishing of multiple charters, guidelines, self-proclaimed, and self-applied principles (with little or no coercive capacity to make sure these wishful promises are effectively enforced), critics have warned that “we should not yet celebrate consensus around high-level principles that hide deep political and normative disagreement” (Mittelstadt, 2019). This has created an implementation gap in designing and deploying systems to meet ethical standards, as well as learning gaps, with activities happening in silos and few mechanisms in place to drive global collaboration and rapid scaling of proven tools and practices while leaving many voices unheard, in particular from vulnerable communities. Today, one of the biggest problems in the AI governance system is that there are too many principles and insufficient operationalization or actual implementation. In this respect, the European Union’s (EU) AI Act, initially tabled by the EU Commission in April 2021 and provisionally adopted through a political deal between negotiators of the EU Parliament and the Council as of December 2023, is a forcing function (European Parliament, 2023). The world’s first comprehensive legislation on AI seeks to ensure its safe use with respect to fundamental rights and democracy. Namely, the bill aims to prevent “unacceptable risks” such as social scoring, cognitive behavioral manipulation of people, and real-time and remote biometric identification systems. This novel Act complements the already ground-breaking legislation on General Data Protection Regulation (GDPR) enacted by the EU in 2016. In the USA, the White House has convened various stakeholders, including civil society, to jointly establish a *Blueprint for an AI Bill of Rights* (White House, 2022), but it is only a starting point, and its scope would protect only the American people.

Over the past few years, new initiatives about AI and globally responsible governance have started to develop, such as the AI Governance Alliance (AIGA) initiated by the World Economic Forum (World Economic Forum [WEF], 2023b). International philanthropy networks such as WINGS and Philea have also started to develop initiatives to promote global dialogue and collaboration in the philanthropic space around AI and data as common goods. Riding on the coattails of the G20, a “Civil 20” (C20) forum gathering thousands of POs has put forward key recommendations for EIAI (see Appendix).

However, a more systematic and sustained approach is urgently necessary to develop concrete and operational strategies that can contribute to the development and implementation of EIAI by the sector and anticipate future trends and emerging risks. The sector could also play an important role in a renewed “politics of refusal” (Crawford, 2021a, 2021b) that challenges the narrative that just because a technology can be built, it should be deployed, as well as identifying and highlighting longer-term risks such as how AI is concentrating decision-making power in both societal and technological domains. The World Economic Forum’s annual *Global Risks Report* underlines the dangers of “digital power concentration” and “digital inequality” (WEF, 2023a). Rivalries do not only lie between tech behemoth companies (for instance, among Silicon Valley’s neighboring competitors) but also among countries. Geopolitical analysts view the supreme mastery of AI as the new grail in the global arms race of the new “Digital Cold War” between China and the USA, two technology powerhouses likely to split the world again in a bipolar fashion (Taneja & Zakaria, 2023). Indeed, whoever controls AI and its use could tip the balance of power toward desirable or fearsome consequences for humankind. At these crossroads, will we choose to let various forms of AI-enabled nonstop surveillance prevail (Saheb, 2023) and tolerate malicious infringements on fundamental human rights (Schippers, 2018) while nurturing the specter of totalitarianism? Or will we regain our power (McCarthy-Jones, 2020) and emancipate our fellow citizens? In short, will we be wise enough to use AI as a helpful servant rather than let it become our relentless master?

To achieve this discernment, there is a need to develop a solid academic and practical understanding. AI players, developers, and users are still missing “gold standard” literature to which they can reliably refer for ethical guidance. This *Handbook* provides a key contribution toward this goal. Our proposed multi-scale model is hoped to provide a functional reading tool to decipher and understand AI’s complexity and to act ethically in response to AI’s impact on the philanthropic sector at micro, meso, and macro levels.

To conclude by circling back to the classical definition of “*philanthrôpia*” understood as ‘friendship for humankind’ or ‘love of humanity’ (Sulek, 2010b), in a contemporary era where commonplace tools like ChatGPT can think faster than us and handle much more information at once (and sound smarter in many cases, yet unable to verify truthfulness), what distinguishes us as human beings is perhaps no longer the ability to think mentally—that is, on the cerebral level only—but to discern with our hearts—that is, ethically and with compassion—and to pursue philanthropy’s original meaning in the search for the common good.

### Notes

- 1 The authors wish to thank Professor Giuseppe Ugazio (UNIGE) as well as Jayant Narayan and Hubert Halopé (WEF) for helpful comments and inputs on an earlier unpublished research project work that served as an initial basis to develop project activities and sparked inspirations for this chapter.
- 2 Since we see this level of disruption as the most directly impactful and still scantily researched, we develop this section more extensively while being more succinct on the other subsequent sections where we raise the main question to guide the use of the grid and provide only the essential elements of a “checklist” for busy philanthropic leaders.

### References

- Amnesty International (2018, May 9). “How you can help Amnesty fight Twitter trolls”. <https://www.amnesty.org/en/latest/news/2018/05/how-you-can-help-amnesty-fight-twitter-trolls/>. Accessed 17 October 2023.
- Bernholz, L. (2019). “Nonprofits and artificial intelligence”. *Philanthropy* 2173. April.
- Chappell, N. (2018). “Precision philanthropy: Artificial intelligence and the future of generosity”. LinkedIn article. <https://www.linkedin.com/pulse/precision-philanthropy-artificial-intelligence-future-nathan>. Accessed 21 October 2023.
- Crawford, K. (2021a). *Atlas of AI*. Yale University Press. April.
- Crawford, K. (2021b). “AI is neither artificial nor intelligent”. *The Guardian*. June 6.
- Davies, R. (2017). “Automatic for the people: What might a philanthropy algorithm look like?” Charities Aid Foundation (CAF). <https://www.cafonline.org/about-us/blog-home/giving-thought/the-future-of-doing-good/automatic-for-the-people-what-might-a-philanthropy-algorithm-look-like>. Accessed 11 October 2023.
- Dhar, V., & Firth-Butterfield, K. (2021). “Why philanthropy needs to prepare itself for a world powered by AI”. *WEF Agenda*. April.
- Duarte, F. (2024). Number of ChatGPT Users. *Exploding Topics*. July 27. <https://explodingtopics.com/blog/chatgpt-users>. Accessed 09 August 2024.
- European Parliament (2020, updated 2023). “What is artificial intelligence and how is it used?” <https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used>. Accessed 11 October 2023.
- European Parliament (2023). “EU AI Act: first regulation on artificial intelligence”. <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>. Accessed 20 October 2023.
- Foley, S. (2019). “Philanthropists should treat AI as an ethical not a technological challenge”. *Financial Times*. 5 September.
- G20 India (2023). “Civil 20 India 2023 Policy Pack”. <https://ada2030.org/images/C20-india-2023-policy-pack-23.pdf>. Accessed 20 October 2023.

- Gertner, J. (2023). "Wikipedia's Moment of Truth: Can the online encyclopedia help teach A.I. chatbots to get their facts right—without destroying itself in the process?" *New York Times Magazine*. July 18. <https://www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html>. Accessed 21 October 2023.
- Gill, A. S., & Germann, S. (2021). "Conceptual and normative approaches to AI governance for a global digital ecosystem supportive of the UN Sustainable Development Goals (SDGs)". *AI Ethics*, May.
- Hagendorff, T. (2020). "The ethics of AI ethics: An evaluation of guidelines". *Minds & Machines*, July.
- Harvard Business School (HBS) (2016, updated 2021). "Data philanthropy: Driving social change". *HBS Online*. February 2. <https://online.hbs.edu/blog/post/data-philanthropy>. Accessed 20 October 2023.
- Human Rights Watch (HRW) & International Human Rights Clinic (IHRC) (2022). "An agenda for action". *Alternative Processes for Negotiating a Killer Robots Treaty*. November. <https://www.hrw.org/report/2022/11/10/agenda-action/alternative-processes-negotiating-killer-robots-treaty>
- ICRC (2022). "What you need to know about autonomous weapons". July 26. <https://www.icrc.org/en/document/what-you-need-know-about-autonomous-weapons>. Accessed 21 October 2023.
- International Telecommunication Union (ITU) (2023). "AI for good global summit". *Accelerating the United Nations Sustainable Development Goals*. <https://aiforgood.itu.int/>. Accessed 17 October 2023.
- Kanter, B., & Fine, A. (2020). *Unlocking generosity with artificial intelligence: The future of giving*. May.
- Longpre, S., Storm, M., & Shah, R. (2022). "Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies". *MIT Science Policy Review*, August 29, 3, 47–56. <https://sciencepolicyreview.org/2022/07/mitspr-191618003019/>
- MacQuillin, I. (2022). "Normative fundraising ethics: A review of the field". *Journal of Philanthropy and Marketing*, e1740. <https://doi.org/10.1002/nvsm.1740>
- McCarthy-Jones, S. (2020). "Artificial intelligence is a totalitarian's dream – Here's how to take power back". *The Conversation*. August 12. <https://theconversation.com/artificial-intelligence-is-a-totalitarians-dream-heres-how-to-take-power-back-143722>. Accessed 20 October 2023.
- McKinsey & Company (2023). "The economic potential of generative AI: The next productivity frontier". <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>. Accessed 12 October 2023.
- Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI". *Nature Machine Intelligence*, 1, 501–507. November. <https://doi.org/10.1038/s42256-019-0114-4>
- Mönks, J., & Sellen, C. (2020, June 9). "Artificial Intelligence and intelligence of the heart: Opportunities and risks in a post- COVID world". *Alliance Magazine*, London. <https://www.alliancemagazine.org/analysis/artificial-intelligence-opportunities-risks-philanthropy-post-covid-world/>. Accessed 26 September 2023.
- Nandeshwar, A. R. (2023, April 24). "Revolutionizing fundraising part I: 6 ways AI is transforming the nonprofit sector". *CCS Fundraising*. <https://www.ccsfundraising.com/insights/revolutionizing-fundraising-part-i-ai-transforming-nonprofit-sector/>. Accessed 17 October 2023.
- Nandeshwar, A. R., & Jewell, A. (2023, June 20). "Revolutionizing fundraising part II: 5 nonprofit sector AI application examples". *CCS Fundraising*. <https://www.ccsfundraising.com/insights/revolutionizing-fundraising-part-ii-5-nonprofit-sector-ai-application-examples/>. Accessed 17 October 2023.
- Payton, R. L., & Moody, M. P. (2008). *Understanding philanthropy: Its meaning and mission*. Indiana University Press. Chapter 2: "Voluntary Action for the Public Good" (pp. 27–61).
- Pazzanese, C. (2020). "Great promise but potential for peril". *The Harvard Gazette*, 26 October.
- Peltz, J. (2023). *Vatican presses world leaders at UN to work on rules for lethal autonomous weapons*. Associated Press. September 26. <https://apnews.com/article/united-nations-general-assembly-vatican-ai-weapons-4d57f40e5ba5165800004772a04e09b1>. Accessed 21 October 2023.
- Saheb, T. (2023). "Ethically contentious aspects of artificial intelligence surveillance: A social science perspective". *AI Ethics*, 3, 369–379. <https://doi.org/10.1007/s43681-022-00196-y>
- Schippers, B. (2018). "Why technology puts human rights at risk". *The Conversation*. <https://theconversation.com/why-technology-puts-human-rights-at-risk-92087>. Accessed 20 October 2023.
- Sulek, M. (2010a). "On the modern meaning of philanthropy". *Nonprofit and Voluntary Sector Quarterly*, 39(2), 193–212.
- Sulek, M. (2010b). "On the classical meaning of philanthrôpia". *Nonprofit and Voluntary Sector Quarterly*, 39(3), 385–408.
- Taneja, H., & Zakaria, F. (2023). "AI and the new digital cold war". *Harvard Business Review*. September 6. <https://hbr.org/2023/09/ai-and-the-new-digital-cold-war>. Accessed 20 October 2023
- Tomašev, N., Cornebise, J., & Hutter, F. (2020). "AI for social good: Unlocking the opportunity for positive impact". *Nature Communications* 11, 2468. <https://doi.org/10.1038/s41467-020-15871-z>

- UNESCO (2022). *Recommendation on the ethics of artificial intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. Accessed 20 October 2023.
- United Nations (2019). *The age of digital interdependence*. Report of the UN Secretary General’s High-level Panel on Digital Cooperation. June.
- United Nations (2020). Secretary-General’s Nelson Mandela Lecture: “Tackling the inequality pandemic: A new social contract for a new era” [as delivered]. July 18. <https://www.un.org/sg/en/content/sg/statement/2020-07-18/secretary-generals-nelson-mandela-lecture-%E2%80%9Ctackling-the-inequality-pandemic-new-social-contract-for-new-era%E2%80%9D-delivered>. Accessed 20 October 2023.
- United Nations (2023a). *Global digital compact: Background note*. Office of the Secretary-General’s Envoy on Technology. January 17. [https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Global-Digital-Compact\\_background-note.pdf](https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Global-Digital-Compact_background-note.pdf). Accessed 20 October 2023.
- United Nations (2023b). *UN global pulse*. <https://www.unglobalpulse.org/>. Accessed 20 October 2023.
- United Nations Development Programme (UNDP) (2020). *Data philanthropy, international organizations and development policy: Ethical issues to consider*. UNDP Discussion Paper. April. [https://www.undp.org/sites/g/files/zskgke326/files/publications/undp-gpn-sdgi-Data\\_Philanthropy\\_International\\_Organizations\\_and\\_Development\\_Policy.pdf](https://www.undp.org/sites/g/files/zskgke326/files/publications/undp-gpn-sdgi-Data_Philanthropy_International_Organizations_and_Development_Policy.pdf). Accessed 20 October 2023.
- White House (2022). *Blueprint for an AI bill of rights: Making automated systems work for the American people*. The White House Office of Science and Technology Policy. October. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>. Accessed 21 October 2023.
- World Economic Forum (WEF) (2023a, January 11). *Global risks report 2023*. 18th edition. <https://www.weforum.org/publications/global-risks-report-2023>. Accessed 20 October 2023.
- World Economic Forum (WEF) (2023b). *AI governance alliance*. <https://initiatives.weforum.org/ai-governance-alliance/home>. Accessed 20 October 2023.

### Appendix 31.1 G20’s Civil 20 worldwide working group

As part of the Government of India’s presidency of the G20 in 2022–2023, the Civil 20 Engagement Group (“C20”) gathered thousands of civil society and nonprofit organizations in a global dialogue around key global issues. Among these, the Education and Digital Transformation working group hosted a summit on May 20–21, 2023. Some essential takeaways and recommendations that were submitted to the G20 leaders include:

- Develop human-centered design principles for all aspects of education/training/skill development;
- Promote Digital Public Goods and Digital Commons for inclusive, affordable, and customizable Digital SDG solutions;
- Promote digital literacy and hygiene among marginalized and vulnerable communities;
- Strengthen digital accessibility and bridge the Digital Divide, including vulnerable populations, low-literate populations, and persons with disability, and provide access to digital technologies, the Internet, devices, and curricula in the local language;
- Provide comprehensive training to individuals on responsible technology use, addressing issues of misuse and abuse while fostering critical thinking, problem-solving, digital literacy, and ethics;
- Raise the CSO and NGO as well as faith-based voices in the digital governance space and feed into leading consultation processes and upcoming G20, C20, and UN policy work;
- Promote values and compassion-based multi-stakeholder perspectives in the public policy space and digital ethical regulation and governance.

Source: G20 India (2023), *Civil 20 India 2023 Policy Pack*.

# AI AND PHILANTHROPY

## How can they elevate each other?

*Ravit Dotan*

### 1 Introduction

This chapter calls for action for philanthropists to increase their engagement with responsible AI (RAI), which is AI developed and deployed in socially responsible ways. In this section, the chapter identifies how philanthropists engage with AI in a range of roles and explains why philanthropists should focus on responsible AI (RAI) rather than AI. Then, in Sections 2–5, the chapter explains how and why philanthropists can benefit from increasing their involvement with RAI and the negative consequences of failing to do so. Finally, Section 6 suggests actions philanthropists can take across the different roles.

This chapter builds on a multi-stakeholder ideation workshop led by the author at the AI and Philanthropy Conference hosted by the Geneva Centre for Philanthropy (GCP) at the University of Geneva in March 2024. The workshop included mini-talks, live online surveys, break-out discussions, and whole-group discussions. About 40 participants attended. Of the 28 who identified themselves in a survey, 43% identified as academics (many of whom specialize in philanthropy), 29% as philanthropists, 14% as technologists, and 10% as “other,” some of whom were from civil society and intergovernmental organizations. The chapter incorporates ideas and anonymous quotes from the workshop and the conference.

#### *1.1 The Range of philanthropic roles related to AI*

Philanthropists engage with artificial intelligence (AI) in a number of roles: As grantmakers, users, buyers, developers, investors, and social justice advocates. In grantmaking, program officers grant money to AI-enabled non-profits. Some AI-enabled non-profits use or build AI as a core tool to drive impact. Others use it operationally, sometimes without even realizing it, as more and more AI-enabled vendors emerge and even common software, such as Microsoft and Zoom, incorporate AI features. In procurement, Chief Technology Officers (CTOs) buy AI-enabled technology for use across the foundation. Some AI-enabled tools are standalone, while others are added features to existing tools or services to which the foundation subscribes. On the investment side, asset managers and advisors invest money in AI-enabled companies.

Second, like professionals in other fields, philanthropists use AI as individuals who use available tools at will or as organizations that purposefully choose to use the tools. Either way, they typically use free or paid AI-enabled services. For example, participants mentioned using an AI-enabled tool to design decks and another tool that scales invoicing by extracting invoice details into the foundation's systems. In other cases, foundations develop AI-enabled tools in-house.

Third, AI is relevant to philanthropists as social justice advocates across all roles because AI deeply impacts social issues such as social equality, human rights, the digital divide between the Global North and South, and climate.

To just take one example, AI's impacts on social equality are well-documented. We have already seen many cases where the use of AI has resulted in discrimination: The US criminal justice system used AI to predict recidivism risk, and it was twice as likely to mislabel black people as high risk to reoffend compared to white people (Angwin et al., 2016); Amazon developed an AI system to sort through resumes, and it disfavored women (Dastin, 2018); The Netherlands used an AI system to detect fraud in social benefit applications, and it falsely accused 26,000 individuals of fraud with grave consequences to their lives, a disproportionate number of them being people of color (Geiger, 2022); AI algorithms for underwriting mortgage loans exacerbate discrimination against black people in the US: A black person is 67% less likely to have their loan approved by off-the-shelf AI algorithms compared to a white person, whereas they are only 54% less likely to have their loan approved by a human underwriter (Zou and Khern-am-nuai, 2023); Generative AI algorithms write sexist recommendation letters, using different nouns and adjectives for men and women (Wan et al., 2023).

The list goes on and on, and the implications are profound. In each of these cases, masses of people are affected because AI systems, by design, assist in decision-making at scale. Moreover, the cumulative effect is powerful. As AI is increasingly adopted in all aspects of life, minorities face AI discrimination everywhere they turn. If the current trend continues, systematic discrimination could become much worse than it is today. At the same time, AI can be a powerful tool for solving social problems. Examples like the mortgage loans case illustrate that humans are biased and make discriminatory decisions. AI systems that are planned and trained responsibly could assist in more equitable underwriting, correcting some of the discriminatory tendencies of humans and changing societal trends.

## ***1.2 AI vs. responsible AI: only responsible AI fits philanthropy***

Responsible AI (RAI) is AI that is developed and used in socially responsible ways. It aligns with standards and principles such as the AI Risk Management Framework by the National Institute for Standards and Technology (NIST, 2023), ISO/IEC 42001:2023 by the International Organization for Standardization (ISO, 2023), The White House AI Bill of Rights (The White House, 2022), and AI ethics principles by The United Nations Educational, Scientific and Cultural Organization (UNESCO, 2024). Alignment with such standards reduces the risks of inaccuracy, discrimination, human rights violations, the spread of disinformation, environmental damage, and other kinds of social harm.

Regulators worldwide are working on legislation to encourage companies to develop and use AI responsibly. The most prominent example is the European Union's flagship AI regulation, the EU AI Act. Moreover, laws that are not AI-specific, such as privacy and non-discrimination laws, apply to AI as they do to other technologies. Enforcement agencies are starting to apply such laws to AI (e.g., Federal Trade Commission, 2023).

Having said that, there is still a divergence between AI and RAI, with many companies failing to implement responsible AI practices in a meaningful way (Dotan et al., 2023). In the private

sector, companies often first ask themselves whether they want to develop, buy, or otherwise adopt AI. Only later, if at all, do they consider AI responsibility – how to ensure the AI they use, buy, or develop is not socially destructive. As a result, they often end up developing and using harmful AI systems, as the examples discussed above illustrate.

Some in philanthropy have called for increased adoption of AI in philanthropy to rip the benefits of technology, as private sector companies are doing (see, for example, Minevich, 2023). However, if philanthropists take the same approach as private sector companies, thinking of AI first and responsibility second (if at all), the outcome would be counterproductive. AI that is developed or used irresponsibly, without regard to social impact, is likely to run counter to the values of philanthropic movements. An approach that is more consistent with philanthropic values is opposite to that of the private sector: Put social responsibility first and AI second, and only engage with responsible AI, through usage, funding, investing, advocacy, or other modes of engagement.

For this reason, this chapter calls for increasing philanthropy's engagement with responsible AI (RAI). Only RAI will truly strengthen philanthropy's role as a promoter of equitable and inclusive social good, and philanthropy should only amplify responsible AI.

We now turn to how philanthropists can increase their engagement with responsible AI in the various roles they play, starting with their role as grantmakers.

## **2 AI and philanthropists as grantmakers**

RAI can amplify philanthropy by creating new ways to create impact, empower grantees, and reduce participation barriers. A lack of engagement with RAI, either by ignoring AI or engaging irresponsibly with AI, misses out on opportunities for impact and risks creating negative social impacts and harm to grantees.

### ***2.1 RAI opportunities in grantmaking***

#### ***2.1.1 Create impact by funding (responsible) AI-enabled non-profits***

Foundations can harness the power of AI for social good by funding non-profits that have RAI as part of their core product. This can help achieve impact in new ways across all program areas. For example, the Department of Computer Science at Makerere University in Uganda runs multiple grant-supported projects that use AI to empower their community: (1) they are using AI to reduce air pollution, which is a major contributor to poor health in developing countries (Google.org, 2018; Makerere University, n.d.); (2) they are building a conversational platform to communicate pandemic safety guidelines in local languages (Makerere University, 2023a, 2023b); and (3) they are using ChatGPT to provide tailored support to smallholder farmers in sub-Saharan Africa in their local languages (Makerere University, 2023a, 2023b). More generally, a research study found that AI can accelerate 79% of the targets in the UN Sustainable Development Goals (SDGs), a list of goals that require collective global action, such as abolishing poverty and climate action (Vinueza et al., 2020). By funding non-profits that use AI responsibly to enhance their impact, foundations can harness AI for social impact in any area.

#### ***2.1.2 Empower grantees by helping them use AI (responsibly)***

Foundations can harness AI for good not only by funding non-profits that use it as part of their core mission but also by helping grantees use AI (responsibly) operationally. The use cases that



are relevant to non-profits are vast. For example, according to a survey of 4,600 non-profits across 65 countries, the top use cases identified for generative AI in non-profits are as follows (Google.org, 2024, p. 3):

- Marketing and content creation: Creating personalized marketing campaigns that are engaging, informative, and more likely to resonate with potential donors;
- Fundraising: Automating repetitive fundraising tasks, such as grant writing and customized donor communications;
- Program management: Writing content for their programs, including emails, letters, flyers, and cover letters.

Currently, the non-profit sector is not taking full advantage of AI tools as their AI adoption lags behind the private sector (Wanstor, 2024). In the Google.org (2024) survey, respondents identified the following as the key barriers: a lack of familiarity with AI tools (reported by 64%), difficulty in selecting which generative AI tools to use (62%), a lack of funding for AI adoption (51%), and inaccessibility of sufficient training (50%) (Google.org, 2024). Foundations can help decrease the gap by providing funding, educational materials, and other forms of support.

### *2.1.3 Reduce participation barriers*

Last, philanthropists can use AI-enabled tools to reduce barriers and diversify participation. In the words of one participant: “AI can strengthen the voice of end beneficiaries in [the] grantmaking processes.” For example, one of the conference participants pointed out that grantees who speak different languages than the funders could use AI to assist in preparing materials. This approach can be effective provided that grantees and applicants have appropriate support to avoid generic and ineffective materials, which funders are unfortunately increasingly seeing as a result of the advent of generative AI. Workshop participants mentioned that philanthropists could use AI to streamline communication with communities and grantees, which would allow increased participation.

## **2.2 Risks of lack of RAI in grantmaking**

Engaging with AI responsibly as grantmakers includes prioritizing AI responsibility when selecting grantees, including grantees who use AI as their main product and grantees who use it operationally, empowering grantees to implement best practices for AI responsibility, and ensuring that any use of AI in grantee interactions is responsible. Failing to use AI responsibly in grantmaking activities may result in AI-enabled grantees causing harm and foundations causing harm to non-profits.

### *2.2.1 The risk of harm caused by AI-enabled grantees*

Currently, most non-profit employees use AI tools to some degree. In Google.org’s survey of 4,600 non-profits across 65 countries, 58% of respondents said that at least one member of their organization uses generative AI in their day-to-day work, and 12% said that half or more of the people in their organizations use it (Google.org, 2024). If non-profits use these and other AI tools irresponsibly, they could undermine the mission they are trying to promote. For example, the AI-generated

texts they create may be inaccurate and contain discriminatory language, and the core products that AI non-profits develop can produce high carbon emissions.

### *2.2.2 Harms to grantees*

Foundations that use AI to assist in the grantmaking process without prioritizing responsibility can harm existing and prospective grantees. For example, like non-profits, foundations' generated texts for communicating with grantees may be inaccurate and use discriminatory language, and tools for sorting through applications may also be discriminatory.

## **3. AI and philanthropists as AI users, buyers, and developers**

Currently, AI adoption rates in foundations are low. For example, a recent exploratory study found that 52% of 24 European foundations had never used AI or data science, and only 17% had adopted AI for both internal and external use (Candela et al., 2024, p. 19). However, like non-profits and other professional organizations, foundations can benefit from AI. Benefits include scaling impact by optimizing internal operations, improving giving through data analysis, and developing tools for the public good. Failure to embrace or do so responsibly can result in a failure to capitalize on these opportunities and risks, as well as negative social impacts, reputational damage, and competitive disadvantage.

### *3.1 Opportunities in using, buying, and developing RAI*

#### *3.1.1 Scale impact through optimizing internal operations*

AI-enabled tools can optimize internal operations by reducing administrative loads, lowering costs, and freeing up time. For example, one workshop participant mentioned an AI-enabled tool for automating accounting tasks, such as invoicing. The tool can read invoices and input information from them into the foundation's systems.

At the workshop, some philanthropists were weary of focusing too much on efficiency because the ultimate goal of philanthropy is positive social impact. However, efficiency is not cross-purpose with social good missions, as freed resources can be redirected to the philanthropic mission. One participant articulated this point especially well: "AI can create time for philanthropists to focus on important strategic points and engage with impact while freeing up time from mundane operational work."

#### *3.1.2 Improve giving by empowering (responsible) data analysis*

The adoption of RAI tools can also strengthen philanthropy by enhancing data analysis. In particular, workshop participants noted that AI can help analyze qualitative and quantitative data for purposes such as improving resource allocation, impact assessment, and understanding contexts. For example, as one participant articulated well: "AI can help philanthropy better understand giving trends to ensure equity in grant-making practices and understand impact across different socioeconomic demographics."

The King Baudouin Foundation (KFB) has undertaken such a project. They have built an internal database to curate their activities and developed a Natural Language Processing tool to make it easy to search through it. In addition, the tool will identify top themes in the current activities. The

process need not be very expensive, they do all the work in-house using Microsoft tools, except for hiring an engineer for EUR 3,000 (Candela et al., 2024, p. 10).

### *3.1.3 Build (responsible) AI tools for public use*

Some foundations are developing AI tools for public use. One example is the McGovern Foundation's JournaPilot, built by the foundation's data team. The tool serves the journalistic community: Using LLM, JournaPilot helps journalists turn their articles into social media posts (McGovern, 2023).

### *3.2 Risks of lack of RAI in using, buying, and developing AI*

Engaging with AI responsibly as users, buyers, and developers of AI includes establishing policies and implementing best practices for responsible use, procurement, and development. For example, the Technology Association of Grantmakers and Project Evident created a framework for the responsible adoption of AI in foundations (TAG, 2023). The Ford Foundation created guidelines for the responsible procurement of AI in public administration, which can also be helpful for AI procurement in foundations (Ford, 2023). To develop AI responsibly, foundations can use guidelines intended for the private sector or for all sectors, such as the NIST AI Risk Management Framework (NIST, 2023). Failing to use, buy, and develop AI responsibly, or at all, may result in lagging behind others, negative social impacts, reputational damage, and non-compliance.

### *3.3 Lag behind, lose competitiveness, and inability to help grantees*

If foundations do not increase their AI adoption, they risk falling behind others in and outside the philanthropic sector. Lagging behind may entail a decreased ability to partner with others who are more advanced; it can mean a reduced ability to attract talent and an inability to help grantees grappling with AI issues. In the words of two workshop participants, if foundations do not adopt AI, they would "lose bright and creative minds to commercial players" and

[Lag] behind the ecosystem's development unable to partner with other players in the ecosystem. Hence it won't be able to have any expertise in the field and won't be able to help grantees dealing with direct or indirect AI issues.

### *3.4 Negative impacts*

Like non-profits, if foundations adopt AI but do so irresponsibly, they may create negative impacts that undermine their mission. The negative consequences may occur when staff use publicly available AI tools or dedicated tools that the foundation buys. For example, a foundation that uses AI-enabled tools in the hiring process may unwittingly discriminate against candidates. The actions foundations can take to avoid such outcomes include working internally to raise awareness of AI risks and establishing internal policies for responsible use, including AI responsibility metrics in diligence and contracting with AI-enabled vendors.

### *3.5 Reputational harm and non-compliance*

A lack of AI responsibility in grantmaking can lead to non-compliance and severe reputational damage. For example, a foundation focused on Diversity, Equity, and Inclusion (DEI) whose

AI-assisted hiring decisions discriminate against minorities would be vulnerable to negative press coverage for undermining their own mission, and they would potentially be liable for violating non-discrimination laws. Reputational damage and liability can also result from malfunctions of the AI system. A case in point is the recent malfunction of an Air Canada chatbot: A traveler whose grandmother had died asked the chatbot about the company's bereavement policy, and the chatbot said that they gave refunds for bereavement travel, which was false. When the traveler asked the company for a refund, he was denied, because it was not the company's policy. The traveler sued Air Canada and won (Yagoda, 2024). The same can happen to foundations, as any other organization, if they do not prioritize using, buying, and developing AI responsibly.

## **4 AI and philanthropists as investors**

Philanthropy has large sums of money at its disposal. In the US alone, the endowment assets of philanthropic organizations are estimated at \$1.1 trillion (Schultz, 2024). A survey of 65 US-based foundations indicates that foundations tend not to invest their money in social impact, with only 5% of them doing so, even though 92% of them were active members of impact investing groups (Schultz, 2024). Investing in responsible AI is an opportunity to change this trend – it can increase return on investment (ROI) and have a meaningful impact on deep social processes.

### ***4.1 RAI opportunities for philanthropic investors***

#### *4.1.1 RAI enables ROI*

Responsible AI leads to better social outcomes, but it can also lead to better financial outcomes. For example, a McKinsey survey (2021) shows that companies with the highest returns from their AI engage in risk mitigation practices more often than others. The reasons may include improved compliance, reduced reputational risk, better talent attraction, better product adoption, and improved product quality (Bevilacqua et al., 2023; Dotan, 2022; The Economist Intelligence Unit, 2020).

Imagine an AI tool for loan decisions. Responsible development of this tool includes ensuring that it provides accurate and fair recommendations across all user profiles, regardless of demographics. If this effort succeeds, it means that the product is a better investment – it serves more customers, facilitates better business decisions, is less likely to cause scandals, and is less likely to violate non-discrimination and other laws.

#### *4.1.2 Shaping the AI ecosystem and positioning philanthropic organizations as leaders*

AI is at the heart of consequential global debates. Philanthropists can impact these debates, position themselves as global social leaders, and shape the AI ecosystem in a positive direction by understanding these debates and thoughtfully investing in AI companies to promote social good.

One of the major ongoing debates in the AI ecosystem today is whether AI should be open or closed source. When AI systems are open source, their code is available to the public. If they are close source, the code is unavailable outside the organization that created it.

Famously, the open-source approach was an important part of the founding ethos of OpenAI, as its name suggests, and the first versions of ChatGPT were open (OpenAI, 2015). However, the company changed its mind when it released ChatGPT 4. Not only did it refuse to make the code

publicly available, but it also refused to release almost any technical details about how it works (OpenAI et al., 2023; Vincent, 2023). On the other side of the debate, Meta, IBM, and other companies advocated for the open approach, and Meta open-sourced its LLM model, Llama (Lin, 2023).

OpenAI and other advocates of the closed approach argue that AI is too dangerous to make publicly available to everyone on the internet. Bad actors could abuse the models and do great harm. For example, they could produce and spread mass amounts of disinformation. Moreover, some argue that the open approach inhibits competition by compromising companies' competitive advantages (OpenAI et al., 2023; Vincent, 2023). Advocates of the open approach, such as Rahul Roy-Chowdhury, the CEO of Grammarly, argue that the open approach to AI enhances creativity, innovation, and competition because it lowers entry barriers. The reason is that training AI requires the kind of resources that typically only big tech has, so making the code unavailable to anyone else makes it very difficult for smaller actors to enter the space. Moreover, Roy-Chowdhury and other advocates of the open approach argue that open code facilitates transparency and is essential to creating safeguards that effectively address AI risks (Roy-Chowdhury, 2023). However, some critics argue that the big companies calling for open-source AI are financially motivated, as companies such as Meta and IBM are "struggling to catch up with the rush of attention that OpenAI and its investment partner, Microsoft, are drawing" (Lin, 2023).

Much is at stake in the open/close approach debate and philanthropists can position themselves as key actors representing civic interests. One of the conference participants pointed out that philanthropists have already done so in the past when backing Mozilla, a non-profit that develops and advocates for open-source software. The participant, who supports the open approach and believes it is necessary to promote public good, argues that philanthropic organizations should invest in or otherwise fund open-source AI companies. In doing so, he argued, they would create an alternative to companies like OpenAI, similar to how Mozilla created an alternative browser, Firefox. Some foundations have already jumped into the debate. For example, the National Science Foundation (NSF) is a member of the coalition led by Meta and IBM calling for open-sourcing AI. However, those who support the closed approach, such as another conference participant, argue that philanthropic organizations should be doing the exact opposite – protecting the public from unchecked AI use by investing to promote the closed approach.

The debate between the two participants, which unfolded during the conference, illustrates how influential philanthropic investments can be in key AI issues that deeply affect society and how important it is for philanthropists to carefully consider their position.

#### ***4.2 Risks of lack of RAI for philanthropic investors***

Engaging with AI responsibility as an investor includes incorporating AI responsibility metrics into the due diligence process – when selecting portfolio companies and when selecting asset managers. It also includes providing support to both portfolio companies and asset managers to help them grow in their AI responsibility journey (see Dotan, 2022 for a framework for responsible AI investing for Venture Capital). Neglecting AI responsibility considerations in the investment process may lead to missed opportunities for impact and reputational damage from investing in technology that undermines philanthropic missions. In addition, investing in AI without understanding the broader social context of the investment may have harmful impacts on consequential processes, such as open/closed trends in AI development.

## **5. AI and philanthropists as social justice advocates**

As discussed at the beginning of the chapter, AI systems have a profound impact on society. Each system typically impacts masses of people, and cumulatively, AI systems can make social problems even more entrenched than they are now or, alternatively, solve them. This characteristic of AI systems means AI is relevant to philanthropists as social justice advocates in all of their roles.

### ***5.1 RAI opportunities for philanthropists as social justice advocates***

In addition to engaging AI responsibly in the ways discussed in the previous sections, philanthropists have an opportunity to lead the way in AI-related consequential debates by actively advocating for AI responsibility to regulators. Doing so can amplify civic and marginalized voices, balancing the outsized influence that big tech has on AI regulation and the impact of political and geopolitical interests.

#### ***5.1.1 Big tech's influence on AI regulation***

Big tech influences AI regulation through lobbying, committee work, and shaping academic research that informs policy. Big tech's AI lobbying in the US greatly intensified in 2023, to the extent that it is reported to have gone on a "charm attack" on legislators (Sayki, 2023; Williams, 2023). According to these reports, big tech's influence typically begins by educating lawmakers, hoping to win their goodwill before pushing for their desired policies of less regulation. Moreover, philanthropic organizations backed by doomsayer tech billionaires have ramped up their lobbying efforts (Bordelon, 2024). Renowned experts such as Andrew Ng and Timnit Gebru argue that doomsday scenarios, also known as "existential risk," are exaggerated, distract attention from issues that need attention, and serve the interests of big tech (Farrell, 2023; Goldman, 2023).

Big tech also influences regulation through participation in committee work. A key example is the High-Level Expert Group on Artificial Intelligence (HLEGAI), a group appointed by the European Commission to advise them on artificial intelligence strategy. HLEGAI created guidelines and recommendations that have been central to the Commission's approach to AI, including informing the EU AI Act (European Commission, 2022; Floridi, 2021). Schyns et al. (2021) argue that corporations dominated the AI HLEG. About 50% of the group members were corporate representatives, with Google, IBM, and others, having their own seats. Many academics on the committee had ties to big tech, and civil society only had a few seats. Moreover, Schyns et al. (2021) argue that corporate interests shaped the group's deliberations. For example, they reported that a subgroup of HLEGAI began developing "red lines," a list of AI applications that the EU would ban outright. However, this effort was stifled when industry representatives threatened to leave the AI HLEG if it went through.

Last, big tech influences AI policymaking through academic research. Abdalla and Abdalla (2021) showed that policymakers rely heavily on academia to advise them on policies. By funding many academics and their labs, big tech indirectly influences the policies that might be developed. Demonstrating the reach of corporate funding in academia, Ahmed and Wahed (2020) found a pronounced presence of authors with corporate affiliations in papers accepted to the two prestigious AI conferences, the International Conference on Machine Learning (ICML) and the Conference on Neural Information Processing Systems (NeurIPS), in 2019: 30% of ICML papers and 35% of NeurIPS papers had authors with corporate affiliations in that year. Birhane et al. (2021) show

that in the years 2018/2019, 71% of the most cited papers in these two conferences, ICML and NeurIPS, had authors with corporate affiliations. Of these, 58% were affiliated with big tech in particular.

### *5.1.2 Politics and geopolitics in AI regulation*

Political and geopolitical interests also shape AI regulation. For example, in the US, Democrats are generally much more active in AI legislation and promote different AI legislation than Republicans – with Democrats more focused on fairness and general AI ethics and Republicans more focused on national security and data rights (Dotan, forthcoming). Executive Orders are another arena of political struggle in the US as Biden revoked two AI Executive Orders made by Trump, EO 13971 and EO 13943, and recently, Trump announced that if he gets elected in the 2024 election, he will revoke Biden’s most recent AI Executive Order, EO 14110 (Hutton, 2023).

Geopolitically, AI regulation reflects and may intensify global power struggles. For example, both of the abovementioned Trump EOs pertain to the tension between the US and China, aiming to constrain Chinese companies in AI and data-related issues. Falajiki (2023) and others argue that the US-China power struggles around AI and its regulation negatively impact Africa. Africa, Falajiki argues, has become a battlefield in the countries’ race for AI supremacy, which set in motion a competition over who would decide on the techno-moral standards for responsible AI adoption. Moreover, he argues, both countries are investing large sums of money to sway Africa in their direction, which may bifurcate the continent and endanger its ability to benefit from AI:

[D]ifferences in the American and Chinese approaches have significant potential to cause a technology governance fragmentation – a bifurcation of AI ethics and governance in Africa. Such a bifurcation, even if it ultimately proves feasible to implement, has significant potential to impede the benefits Africa can reap from full participation in the AI space.

## **5.2 Risks of lack of RAI engagement for philanthropist social justice advocates**

A lack of philanthropic involvement in AI advocacy may result in more social harm from AI systems and harm to philanthropic institutions themselves due to a loss of trust in them.

### *5.2.1 Negative social impact*

As several of the workshop participants pointed out, a key consequence of staying out of AI advocacy is leaving much power in the hands of big tech to shape the future of the technology and society. In the words of three of the participants, key consequences include: “Leaving the field for negative big-tech influence,” “AI will become big tech exclusive and could remove realistic options for choice,” and “Leaving it to big tech to shape our digital future.”

### *5.2.2 Irrelevance, loss of trust, and breach of social contract*

Since AI is pivotal to social justice issues, not increasing involvement with RAI through regulatory advocacy or in other ways may harm philanthropy’s reputation to the point of losing trust in

philanthropic institutions. This negative consequence was the one that workshop participants worried the most about. For example, as some have articulated:

- They may become irrelevant and lose their legitimacy because of a lack of action on such an important societal cross-cutting topic;
- It will fail civil society – and the communities they represent – who all need to have a central voice in the AI future;
- Risk of no impact and irrelevance and thereby a breach of social contract on the use of precious public capital.

## **6 How philanthropists can increase their engagement with RAI**

The previous sections included examples of activities philanthropists can take on to increase their engagement with RAI in their various roles. This section identifies and describes activities that are relevant across all philanthropic roles.

### ***6.1 Self-educate and self-regulate***

As with many other subjects, AI responsibility starts with awareness and education. People with non-technical backgrounds sometimes feel insecure about AI because they lack the expertise to fully comprehend technical details about how AI works. However, first, a basic understanding of the technology is sufficient. Second, the key skill to hone is understanding how the use of AI systems may impact society, e.g., how it may impact social dynamics and power structures. Activities of this kind are often within philanthropists' wheelhouses.

In addition to educating themselves about AI responsibility, deciding on internal policies, processes, and guardrails is essential to ensure responsible engagement with AI in any relevant roles.

### ***6.2 Create public resources customized for foundations***

Publicly available resources on AI responsibility can help foundations and individuals upskill, especially small foundations and their employees, who may be more constrained. Helpful resources include educational materials, case studies, and guidelines for the responsible use of AI in philanthropy.

Some resources are already out there. For example, the Technology Association of Grantmakers (TAG), in collaboration with Project Evident, released the “Responsible AI Adoption Framework” (TAG, 2023), with guidelines for the responsible adoption of AI tools in foundations. InfoTech has created a library of generative AI use cases in non-profits and professional associations (InfoTech, 2023). Fundraising.ai has a repository of educational videos and other materials (Fundraising.ai, 2024).

### ***6.3 Fund RAI and organizations and activities***

Funding to support RAI includes organizations that develop responsible AI products that contribute to positive impact, organizations or projects that promote AI responsibility in the ecosystem (such as diversifying the ecosystem or offering services to support AI responsibility), and activities to promote AI responsibility in non-profits (such as education to the non-profit personnel).



#### 6.4 Vet and support vendors, non-profits, and portfolio companies

When conducting due diligence for any purpose – grantmaking, procurement, or investing – it is essential to ensure that the organization, company, or service governs their AI responsibly, following industry standards and ensuring their AI usage is aligned with relevant philanthropic missions. A vetting process is important not only for organizations whose core activities incorporate AI but also for organizations that use AI-enabled assistive tools. Moreover, especially for grantees and portfolio companies, it is crucial to support them as they grow in their AI responsibility. The support may include educational materials and access to experts who can guide them (for an example of a framework for responsibly investing in AI, see Dotan, 2022).

#### 6.5 Collaborations and coalitions

Workshop participants were most enthusiastic about collaborations and coalitions to support increasing RAI engagement in philanthropy. The coalitions would allow the pooling of resources, avoiding duplicating effort and achieving synchronization between foundations, which can amplify the impact of the work. For example, some of the suggestions are as follows:

- Form a coalition to fund a code of conduct, best practices, and toolbox to rely on;
- Larger foundations or coalitions could develop tools such as a self-audit for AI tools;
- Pool resources to support – at scale – technical capacity building and responsible engagement with AI across civil society.

### 7 Conclusion

This chapter is a call for action to increase philanthropy's engagement with responsible AI across all the roles it plays: as grantmakers, users, buyers, developers, investors, and advocates of social justice. The chapter identifies many advantages of doing so, from creating impact in new ways through funding AI-enabled non-profits to shifting the AI ecosystem through investments to establishing global leadership through regulatory advocacy. The chapter also identifies many negative consequences that philanthropists may face if they do not increase their engagement with responsible AI, from causing negative social impact to irrelevancy and breach of social contracts. Last, the chapter identifies activities philanthropists can take on to increase their engagement with responsible AI across all their roles, adding to suggestions for particular roles made throughout the chapter. Meaningful change can happen through self-educating and self-regulating, creating public resources tailored for foundations, vetting and supporting vendors, non-profits, and portfolio companies, and establishing coalitions of philanthropists to move the sector forward together.

### Bibliography

- Abdalla, M., & Abdalla, M. (2021). The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 287–297. <https://doi.org/10.1145/3461702.3462563>
- Ahmed, N., & Wahed, M. (2020). *The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research*. <https://doi.org/10.48550/ARXIV.2010.15581>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Bevilacqua, M., Berente, N., Domin, H., Goehring, B., & Rossi, F. (2023). *The Return on Investment in AI Ethics: A Holistic Framework*. <https://doi.org/10.48550/ARXIV.2309.13057>

- Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2021). *The Values Encoded in Machine Learning Research*. <https://doi.org/10.48550/ARXIV.2106.15590>
- Bordelon, B. (2024, February 23). AI Domsayers Funded by Billionaires Ramp Up Lobbying. *Politico*. <https://www.politico.com/news/2024/02/23/ai-safety-washington-lobbying-00142783>
- Candela, F., Kilcalp, S., & Spiers, D. (2024). *Data Science, AI and Data Philanthropy in Foundations: On the Path to Maturity*. <https://search.isuealab.org/resource/data-science-ai-and-data-philanthropy-in-foundations-on-the-path-to-maturity.html>
- Dastin, J. (2018). Insight—Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters*. Retrieved March 31, 2018, from <https://www.reuters.com/article/idUSKCN1MK0AG/>
- Dotan, R. (2022). *Responsible Investing in AI: A Guide for VCs*. <https://www.techbetter.ai/ai-due-diligence>
- Dotan, R. (Forthcoming). US Regulation of AI. In *The Handbook on Applied AI Ethics*. Editors: Alexander Kriebitz, Christoph Lütge and Raphael Max. Elgar. <https://www.techbetter.ai/us-ai-regulation>
- Dotan, R., Rosenthal, G., Buckley, T., Scarpino, J., Patterson, L., & Bristow, T. (2023). *Evaluating AI Governance: Insights from Public Disclosures*. <https://www.techbetter.ai/evaluating-ai-governance>
- European Commission (2022). *High-Level Expert Group on Artificial Intelligence*. <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>
- Falajiki, C. (2023, October 27). The Political Economy of AI Ethics and Governance in Africa in the Rise of the US-China Geopolitics. *African Observatory on Artificial Intelligence*. <https://www.africanobservatory.ai/social/the-political-economy-of-ai-ethics-and-governance-in-africa-in-the-rise-of-the-us-china-geopolitics>
- Farrell, J. (2023, October 31). Google Brain Founder Andrew Ng Says Threat of AI Causing Human Extinction Is Overblown. *Silicon Angle*. <https://siliconangle.com/2023/10/31/google-brain-founder-andrew-ng-says-threat-ai-causing-human-extinction-overblown/#:~:text=AI-,Google%20Brain%20founder%20Andrew%20Ng%20says%20threat,causing%20human%20extinction%20is%20overblown&text=Andrew%20Ng%2C%20a%20world%20leader,to%20humankind%20is%20vastly%20exaggerated.>
- Federal Trade Commission (2023, December 19). *Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology Without Reasonable Safeguards*. <https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without>
- Floridi, L. (2021). The European Legislation on AI: A Brief Analysis of Its Philosophical Approach. *Philosophy & Technology*, 34(2), 215–222. <https://doi.org/10.1007/s13347-021-00460-9>
- Ford Foundation (2023). *A Guiding Framework for Vetting Technology Vendors Operating in the Public Sector*. Ford Foundation. [https://www.fordfoundation.org/wp-content/uploads/2023/03/final\\_ford-foundation-guiding-framework-r3-full-document-final2.pdf](https://www.fordfoundation.org/wp-content/uploads/2023/03/final_ford-foundation-guiding-framework-r3-full-document-final2.pdf)
- Fundraising.ai (2024). *Harness Responsible and Beneficial AI for Fundraising*. <https://www.google.com/url?q=https://fundraising.ai/resources/%23ai-resources%25E2%2580%25B8&sa=D&source=docs&ust=1711937885358502&usq=AOvVaw1iHDvVwBw8u09jmMI14COY>
- Geiger, G. (2022, March 18). How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud. *Vice*. <https://www.vice.com/en/contributor/gabriel-geiger>
- Goldman, S. (2023, May 31). AI Experts Challenge ‘Doomer’ Narrative, Including ‘Extinction Risk’ Claims. *VentureBeat*. <https://venturebeat.com/ai/ai-experts-challenge-doomer-narrative-including-extinction-risk-claims/>
- Google.org (2018). *Google AI Impact Challenge*. <https://impactchallenge.withgoogle.com/ai2018/>
- Google.org (2024). *Nonprofits & Generative AI: Google.org Report*. [https://services.google.com/fh/files/blogs/nonprofits\\_and\\_generative\\_ai.pdf](https://services.google.com/fh/files/blogs/nonprofits_and_generative_ai.pdf)
- Hutton, C. (2023, December 2). Trump Vows to Cancel Biden Executive Order on AI to Protect Free Speech. *Washington Examiner*. <https://www.washingtonexaminer.com/news/2432277/trump-vows-to-cancel-biden-executive-order-on-ai-to-protect-free-speech/>
- InfoTech (2023). *Generative AI Use Case Library for the Nonprofit & Professional Associations Industry*. IntoTech. [https://www.infotech.com/research/ss/generative-ai-use-case-library-for-the-nonprofit-professional-associations-industry?utm\\_id=ITRG2023GenAINonprofit](https://www.infotech.com/research/ss/generative-ai-use-case-library-for-the-nonprofit-professional-associations-industry?utm_id=ITRG2023GenAINonprofit)
- International Organization for Standardization (ISO) (2023). *ISO/IEC 42001:2023 – Information Technology. Artificial Intelligence Management System*. International Organization for Standardization (ISO); March 31, 2024. [https://www.google.com/search?q=what+does+iso+stand+for&oq=what+does+iso+s+stand+for&gs\\_lcrp=EgZjaHJvbWUqBwgAEAAAYgAQyBwgAEAAAYgAQyBwgBEAAAYgAQyBwgCEAAAYgAQyBwgDEAAAYgAQyBwgEEAAAYgAQyBwgFEAAAYgAQyBwgGEAAAYgAQyBwgHEAAAYgAQyBwgIEAAAYgAQyBwgJEAAAYgATSAQgzMTI2ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8](https://www.google.com/search?q=what+does+iso+stand+for&oq=what+does+iso+s+stand+for&gs_lcrp=EgZjaHJvbWUqBwgAEAAAYgAQyBwgAEAAAYgAQyBwgBEAAAYgAQyBwgCEAAAYgAQyBwgDEAAAYgAQyBwgEEAAAYgAQyBwgFEAAAYgAQyBwgGEAAAYgAQyBwgHEAAAYgAQyBwgIEAAAYgAQyBwgJEAAAYgATSAQgzMTI2ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8)

- Lin, B. (2023, December 5). Meta and IBM Launch AI Alliance. *The Washington Post*. <https://www.wsj.com/articles/meta-and-ibm-launch-ai-alliance-300c4862>
- Makerere University (n.d.). *Air Quality Monitoring*. Retrieved March 31, 2024, from <https://cs.mak.ac.ug/research/project/3>
- Makerere University (2023a, August 21). *Faculty Member of the Computer Science Department Wins Grant*. <https://cs.mak.ac.ug/news/view/23>
- Makerere University (2023b, August 21). *Head of Makerere Artificial Intelligence Lab Wins Grant*. <https://cs.mak.ac.ug/news/view/22>
- McGovern Foundation (2023, October 16). *Foundation Launches AI-Powered Social Media Tool for Journalists*. <https://www.mcgovern.org/foundation-launches-ai-powered-social-media-tool-for-journalists/>
- McKinsey (2021, December 8). *The State of AI in 2021*. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021>
- Minevich, M. (2023, November 25). AI And Generation Z: Pioneering a New Era of Philanthropy. *Forbes*. <https://www.forbes.com/sites/markminevich/2023/11/25/ai-and-generation-z-pioneering-a-new-era-of-philanthropy/?sh=ebb38fd6d099>
- National Institute for Standards and Technology (NIST) (2023). *AI Risk Management Framework (AI RMF)*. National Institute for Standards and Technology (NIST). <https://www.nist.gov/itl/ai-risk-management-framework>
- OpenAI (2015, December 11). *Introducing OpenAI*. <https://openai.com/blog/introducing-openai>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J.,..., & Zoph, B. (2023). *GPT-4 Technical Report*. <https://doi.org/10.48550/ARXIV.2303.08774>
- Roy-Chowdhury, R. (2023, December 22). Why Open-Source Is Crucial for Responsible AI Development. *World Economic Forum (WEF)*. <https://www.weforum.org/agenda/2023/12/ai-regulation-open-source/>
- Sayki, I. (2023, May 4). Big Tech lobbying on AI Regulation as Industry Races to Harness ChatGPT Popularity. *Open Secrets*. <https://www.opensecrets.org/news/2023/05/big-tech-lobbying-on-ai-regulation-as-industry-races-to-harness-chatgpt-popularity/>
- Schultz, A. (2024, February 29). Only 5% of U.S. Foundations Invest for Impact, Study Finds. *Penta*. <https://www.barrons.com/articles/only-5-of-u-s-foundations-invest-for-impact-study-finds-c4fb34d4>
- Schyns, C., Rosen Fondahn, G., Yanchur, A., & Pilz, S. (2021, November). How Big Tech Dominates EU's AI Ethics Group. *Euobserver*. <https://euobserver.com/investigations/153386>
- Technology Association of Grantmakers (TAG) (2023). *Responsible AI Adoption in Philanthropy*. Technology Association of Grantmakers (TAG). <https://www.tagtech.org/ai-resources-for-philanthropy/>
- The Economist Intelligence Unit (2020). *Staying Ahead of the Curve: The Business Case for Responsible AI*. <https://pages.eiu.com/rs/753-RIQ-438/images/EIUStayingAheadOfTheCurve.pdf>
- The European Commission (2022). *High-Level Expert Group on Artificial Intelligence*. <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>
- The White House (2022). *Blueprint for an AI Bill of Rights*. The White House. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>
- United Nations Educational, Scientific and Cultural Organization (UNESCO) (2024). *Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization (UNESCO). <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
- Vincent, J. (2023, March 15). OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong.' *The Verge*. <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 233. <https://doi.org/10.1038/s41467-019-14108-y>
- Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly Is a Warm Person, Joseph Is a Role Model": Gender Biases in LLM-Generated Reference Letters (arXiv: 2310.09219). arXiv. <http://arxiv.org/abs/2310.09219>
- Wanstor (2024). *AI's Achilles Heel? The Perils of Ignoring Data Readiness in the Race to Innovate*. Wanstor. <https://24886651.fs1.hubspotusercontent-eu1.net/hubfs/24886651/wanstor-AIs-achilles-heel-the-peril-of-ignoring-data-readiness-in-race-to-innovate-not-for-profit.pdf>

- Williams, Z. (2023, September 27). Amazon, Google among Firms Focusing on AI Lobbying in States. *Bloomberg*. <https://news.bloomberglaw.com/in-house-counsel/amazon-google-among-firms-focusing-on-ai-lobbying-in-states>
- Yagoda, M. (2024, February 23). *Airline Held Liable for Its Chatbot Giving Passenger Bad Advice – What This Means for Travellers*. <https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know>
- Zou, L., & Khern-am-nuai, W. (2023). AI and Housing Discrimination: The Case of Mortgage Applications. *AI and Ethics*, 3(4), 1271–1281. <https://doi.org/10.1007/s43681-022-00234-9>

# CLOSING REFLECTIONS AND FUTURE DIRECTIONS

*Camilla Della Giovampaola, Lucia Gomez, Hubert Halopé,  
Maria Cristiana Tudor and Giuseppe Ugazio*

The *Routledge Handbook of Artificial Intelligence and Philanthropy* provides a robust foundation for understanding and harnessing the potential synergies between AI and philanthropy, bringing together contributions by leading academics, artificial intelligence (AI) specialists, and philanthropy professionals. Through multiple disciplinary and professional lenses, contributors unveiled the multifaceted relationship between the two fields – from mapping the application of AI within the philanthropic sector and advocating for philanthropy’s role in guiding the future development of AI in a responsible way to evaluating ethical considerations and risk mitigation strategies. To complement this theoretical knowledge, case studies have provided practical examples of how AI can aid philanthropies’ work and how philanthropic organizations (POs) can champion ethical principles to guide future developments of AI. Importantly, this multi-disciplinary approach has allowed for a nuanced discussion on the intersection between AI and philanthropy, not only bringing opportunities and challenges to the forefront but also remarking that the underuse of this technology within the non-profit sector may in itself be a form of misuse that needs to be addressed.

One central value guiding this Handbook has been the importance of enabling collaboration between scholars and practitioners in shaping the future of AI and philanthropy. We strived to fulfill this goal by bridging the divide between research and implementation, academia, and industry, to demonstrate the power of collective intelligence in developing innovative solutions that address real-world challenges and maximize social impact. In this spirit, a first milestone in providing opportunities for collaboration through interactive exchanges among experts has been the first international conference on *Artificial Intelligence and Philanthropy* hosted in March 2024 by the University of Geneva’s (UNIGE) research team on AI and philanthropy and the Geneva Centre for Philanthropy (GCP).<sup>1</sup> The conference gathered prominent scholars and practitioners in the field to create a space for critical thinking and multi-stakeholder dialogue on these two subjects. Comprising a series of interactive thematic sessions, workshops, and a public conference with a keynote speech by Prof. Luciano Floridi, the conference unveiled many areas of fruitful debate that clearly delineate some of the most pressing priorities for future research, detailed below.

### **Insights from the AI and philanthropy conference**

The AI and philanthropy conference had a dichotomous structure, similar to the present Handbook, with a first thematic session dedicated to discussing how AI can be used to support philanthropic work and a second one on how philanthropy can shape the use and development of AI. These two perspectives were then brought together in a final session examining the future roles of philanthropy for AI and AI for philanthropy.

The first thematic session, *AI for Philanthropy*, explored the opportunities, challenges, and ethical considerations for adopting AI in the philanthropic sector. Within the most relevant opportunities, the main fields of importance identified were environmental conservation, education, and fundraising, parallel to the critical need to develop better data philanthropy practices built on top of the open-source movement. In addition, several important challenges were also identified, the most important being fostering AI literacy and expertise among philanthropists and sector experts. Ethical considerations were also discussed, especially those regarding potential biases, the need for transparency, accountability, and cybersecurity – the latter being a crucial area to be addressed. Words of caution were spoken about the data being fed into AI systems, given the imperative of protecting sensitive information and following applicable privacy regulations (many of which are either absent or underdeveloped, lagging behind the technology's exponential use in society). A common viewpoint emerged: philanthropy could, and perhaps should, play a critical role in shaping AI regulation, serving as an accountability interim body rooted in the collaboration between non-profits, policymakers, and regulators.

This first session was followed by the *Ethical and Inclusive AI empowered by Philanthropy* session, which focused on the ethical debate. Here, panelists emphasized the importance of ensuring AI's usage for the betterment of societies and the minimization of risks posed by this technology. In this respect, it remarked how POs have crucial roles to play in (a) shaping the development of responsible and democratic AI solutions and (b) promoting its ethical and inclusive use. To achieve this, two essential milestones were identified. First, the need to establish international guidelines on ethical AI, which requires collaboration between governments, industry leaders, and philanthropic organizations. Second, the importance of empowering underrepresented groups in AI's development and use, such as open-source communities or vulnerable populations with scarce access to technology and education. Philanthropy can support these by providing resources and funding to mission-aligned researchers, developers, and organizations helping to close the digital divide (Sanders & Scanlon, 2021). However, the complex and fast-evolving nature of AI development calls for an imminent and proactive action of philanthropy toward shaping its trajectory. This requires preventive and proactive intervention led by networks of philanthropists, governments, industry leaders, and the civil society in an institutionalized manner, as opposed to the current status quo of delayed crisis response interventions after harm already took shape. In doing so, early intervention, cooperation, and strategic planning are essential to ensure that AI is developed and used in a way that prioritizes ethical considerations and ultimately holistically benefits society.

The closing debate was centered on discussing *The Future of AI and Philanthropy*. Experts shared cases and identified philanthropy's role in shaping the future of AI, emphasizing the need for dynamic adaptation in the face of rapid technological change. To remain competitive and become competent in the field of AI, philanthropy must invest in learning, share use case experience, forge meaningful partnerships, and build a collaborative network of action. Moreover, philanthropy has a unique opportunity to amplify the voices of civil society, including marginalized communities, in shaping the agenda for AI development and applications that benefit humanity.

Numerous cases were highlighted of the involvement of civil society in the matter, such as for tackling climate change response, global health, and cultural sovereignty. As philanthropy gets empowered in AI adoption, it has the potential to emerge as a force in debating and addressing the potential risks and unintended consequences of AI. By taking a leading role in promoting responsible AI, philanthropy can advocate for safe digital identity systems, build trust in AI through transparency, and sound the alarm on unacceptable risks that AI may pose to societies.

---

*Recurrent topics in the AI and philanthropy conference*

---

Ethics	Ethical dimensions associated with AI development and implementation
Opportunities	Emerging possibilities by philanthropy adopting AI
Challenges	Hurdles faced with AI implementation
Guidelines	Need of framing universal standards regulating AI usage
Empowerment	Elevating the representation of underserved demographics in AI
Responsibility	Encouragement for thoughtful approaches to AI design and use
Security	Vulnerabilities arising due to increased reliance on digital tools and AI
Use cases	Example success stories from philanthropist adopting AI solutions
Collaboration	Endorsement of within- and cross-sector partnerships to drive AI innovation
Education	Need for philanthropist to increase their knowledge about AI

---

In addition to the thematic sessions, a spotlight session was held by a representative of the World Economic Forum (WEF), a non-profit foundation, to showcase the applicability of AI tools within non-profit organizations. This session presented a practical use case of how the WEF leverages AI to manage the knowledge it generates and shares with the broader society, as well as the use of generative AI to increase its operational efficiency. Specifically, the WEF has built its own Generative Pre-Trained Transformers (GPT) based on its internal data to help staff navigate and support their day-to-day operations. The session detailed the multiple steps taken by the WEF during the development and implementation phases, demonstrating how a small pilot project evolved into a company-wide tool. While the value of this customized GPT still needs to be assessed, staff testimonials show positive reactions to this day. This case study offered a valuable example of how organizations can tailor AI technologies to meet their specific needs rather than having to adapt to pre-existing AI tools. For the non-profit sector, where ethical and practical concerns often hinder technology adoption, such an approach may prove to be more suitable. In addition, the WEF's Strategic Intelligence tool, partially freely available online, leverages trusted external data (e.g., white papers and news articles from well-established providers) to inform users on global issues such as climate change or geopolitical topics.

### **Future perspectives and learning opportunities**

This Handbook has been conceived to set the stage for future research endeavors at the intersection of AI and philanthropy. The regional and country level differences characterizing POs' relationship with AI, in particular, call for further scholarly attention. As this book has already brought forward, the intersection of AI and philanthropy is not a homogeneous phenomenon across countries. A number of external factors, including, but not limited to, state regulation, availability of technological resources, and the maturity of the non-profit sector, influence philanthropies' interaction with AI technologies. To break down barriers for POs to better understand and use AI innovations, it is necessary to understand what these differences and commonalities are.

The collection, centralization, and dissemination of information on POs' current use of AI is another important research endeavor that needs pursuing. Having a baseline understanding of where the philanthropic sector is when it comes to technology (and AI in particular) and where it wants to go is important for the formulation of recommendations that speak to the sector's needs. Surveys are important starting points, as they allow to collect information directly from POs and assess where the sector currently stands in respect to AI technology and what are the obstacles and opportunities. The survey launched by UNIGE's research team on AI and philanthropy,<sup>2</sup> which provides information on the current and potential use of AI in Swiss philanthropic organizations, and the survey recently organized by Philea to corroborate its study entitled *Data Science, AI and Data Philanthropy in Foundations: On the Path to Maturity*<sup>3</sup> are two first steps in this direction. Surveys are also central for identifying best practices and informative case studies. Ongoing research endeavors on AI and philanthropy as well as philanthropic organizations intending to engage with AI could indeed largely benefit from a collection of case studies bringing together insights on how POs have either effectively integrated AI technologies or successfully advocated for the implementation of regulatory norms promoting responsible AI in their work.

In addition to research, the Handbook and international conference have further highlighted POs interest and need to build up their internal AI know-how. This knowledge-building can occur in a variety of forms, some of which we illustrate below.

### ***The AI Learning Journey and beyond***

The *AI Learning Journey* is a hands-on AI education and training initiative tailored for philanthropists in Switzerland. It emerged as a Stiftungschweiz initiative that, from its early stages, was designed as a cross-sector collaboration involving industry (Stiftungschweiz and PeakPrivacy), philanthropy (SwissFoundations), and academia (University of Geneva, Geneva Center for Philanthropy). At the end of its first yearly cycle, its success is renowned since the journey is a path of learning and building together, where all participants collaborate to build responsible and secure AI tools to automate philanthropic internal operations, communications, and partner matching among others. You may find additional information in the links of Stiftungschweiz<sup>4,5</sup> and SwissFoundations.<sup>6</sup>

Similarly, the Technology Association of Grantmakers put together a *Responsible AI in Philanthropy Guide*.<sup>7,8</sup> Additional resources on AI Ethics can be learned from Prof. Luciano Floridi's – director of the Digital Ethics Center at Yale University – YouTube Lectures<sup>9</sup> and TechBetter's Workshops and Resources.<sup>10</sup> Similar topics are also covered by ImpactIA Foundation's Projects and Workshops<sup>11</sup> and by the CyberPeace Institute.<sup>12</sup>

There are plenty more resources to mention, among which are books (*The Technology Fallacy: How People Are the Real Key to Digital Transformation*),<sup>13</sup> blogs summarizing current AI developments (*AI Tidbits*),<sup>14</sup> the general AI Community (*HuggingFace*),<sup>15</sup> and of course the MOOC on Coursera and other similar platforms.

***If you are interested in joining a network to shape the future of AI and Philanthropy, feel free to reach out to us!***

### **Notes**

1 <https://www.unige.ch/artificial-intelligence-philanthropy/international-conference>

2 <https://www.swissfoundations.ch/fr/actualites/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/>



- 3 <https://philea.issuelab.org/resource/data-science-ai-and-data-philanthropy-in-foundations-on-the-path-to-maturity.html>
- 4 <https://stiftungschweiz.ch/blog/en/artificial-intelligence-new-rules-of-the-game-for-philanthropy-a-learning-journey>
- 5 <https://stiftungschweiz.ch/blog/wp-content/uploads/2023/11/StiftungSchweiz-AI-in-Everyday-Philanthropy-Learning-Journey-29.11.2023.pdf>
- 6 <https://www.swissfoundations.ch/aktuell/insights-from-the-learning-journey-ai-in-everyday-philanthropy/>
- 7 <https://www.tagtech.org/page/AI>
- 8 <https://www.tagtech.org/wp-content/uploads/2024/01/AI-Framework-Guide-v1.pdf>
- 9 <https://www.youtube.com/user/floridi>
- 10 <https://www.techbetter.ai/>
- 11 <https://impactia.org/>
- 12 <https://cyberpeaceinstitute.org/>
- 13 <https://mitpress.mit.edu/9780262545112/the-technology-fallacy/>
- 14 <https://www.aitidbits.ai/>
- 15 <https://huggingface.co/>

### Bibliography

- Candela, F., Kilicalp, S. & Spiers, D. (2024). Data Science, AI and Data Philanthropy in Foundations: On the Path to Maturity. *Philea*. Available at: <https://philea.issuelab.org/resource/data-science-ai-and-data-philanthropy-in-foundations-on-the-path-to-maturity.html>
- Della Giovampaola, C., Tudor, M., Gomez, L. & Ugazio, G. (2023). Current and Potential AI Use in Swiss Philanthropic Organizations Survey Results. *SwissFoundations*. Available at: <https://www.swissfoundations.ch/fr/actualites/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/>
- Sanders, C. K. & Scanlon, E. (2021). The Digital Divide Is a Human Rights Issue: Advancing Social Inclusion Through Social Work Advocacy. *Journal of Human Rights and Social Work*, 6(2), 130–143. <https://doi.org/10.1007/s41134-020-00147-9>

# INDEX

Note: **Bold** page numbers refer to tables; *italic* page numbers refer to figures and page numbers followed by “n” denote endnotes.

- Abdalla, M. 547  
abductive reasoning approach 309  
accelerationism 508–509; effective 505; narrative of 509–511  
Acosta-Navas, Diana 6  
Acumen Fund 89  
Adrienne, P. 314  
AdSense, process of 296  
Advanced Mathematical (AM) 110  
Agamben, G. 510, 511, 516  
Ahmed, N. 547  
Ahmed, S. 400  
AI aligned with human values 435  
“*AI, art and young people’s mental health*” workshop 437  
AI-as-a-domain 410–412  
AI-as-a-tool 409–410  
AICrowd 345, 348, 349; challenges 349; competitions 349; Food Recognition Challenge 349–350  
AID:Tech 86  
“AI Ethics” 380  
AI extreme risk mitigation philanthropic sector (AIERMPS): accelerationist and decelerationist philosophies 417; adaptation abilities 419–420; catastrophic risks and extinction risks 417; challenges 408–409, 419; coalition-building strategies 422–424; commonalities and differences with other social movements 420–422; culture values norms 413; debates 416; defined 408, 412; disagreements 418; effective altruism movement 413–414, 418; FHI and MIRI 412–413; lightcone infrastructure 417; OpenAI or Anthropic 417; public and political awareness of AI 415–416; rise of AI industry in AI safety 414–415; strategic planners 418; technical and geopolitical perspectives 418–419; traditional public-facing activism 416–417  
AI facial recognition technology 380  
AI for Good Foundation 88  
AI for Good Summit 527  
AI for Philanthropy 555  
AI for Social Good (AI4SG) 527; modern applications 18–24, 23; movement 11; NPOs and for-profit companies, partnerships between **21**, 21–22; stakeholders 18, **19–20**  
The AI Forward Alliance (TAIFA) 88  
AI4Giving 526  
AI Governance Alliance (AIGA) 535  
AiiA festival 447–448  
AI in everyday philanthropy 72, 74; addressing concerns head-on 74; capitalizing on efficiency gains 74; create a nurturing experimentation platform 72; demonstrating AI’s potential through prototypes 74; empower and expand change agent networks 74; foster assistive use of technology 74; harness the readiness of nonprofits 74  
AI in foundations: administration 201–202; overview 200; risks and challenges 200–201  
*AI Learning Journey* 557  
AI-related ethical vigilance for philanthropy 530  
AI tools to transform philanthropic operations: AI and philanthropy for SDGs 88–89; automated

- electronic impact platform 78, 78; Blockchain and DAOs for improved governance 84, 84–88, 85, 87; Decentralized Autonomous Organizations 76, 91; distributed ledger technology 91; donor outreach 78; empowering skill development with AI-enhanced VR simulations 82–83; financial returns 79; flow diagram of artificial intelligence-based method for processing raw data 90; fundraising campaigns 78; fundraising efforts 78; improving diversity and inclusion 83–84; Machine Learning algorithms 77; NLP-powered virtual assistants for operations and engagement 80; optimized HR practices 81; philanthropic organizations 76–78; Sustainable Development Goals 76; Unilever’s recruitment and training approach 81–82; VR-enhanced fundraising and outreach 79–80
- AI to support philanthropy: AI4SG *see* AI for Social Good (AI4SG); data philanthropy 16–18; evolution of philanthropic support 12–14; opportunities and challenges 24–25; philanthropy’s early role 12–13; research and development 24; rise of tech philanthropists in 21st century 13–14; stakeholders and prominent initiatives 25; tech-driven change in philanthropy 14–16; technology’s impact on philanthropy 14–18
- Alborough, L. 151
- Alchian, A. A. 369
- AlexNet 350
- Allen, Paul G. 163
- Altman, S. 21, 374, 375, 414, 503
- Altruist League 4–5, 258–259, **260**, **261**, 267, 270
- Amaya, Nelson 2
- Amesto NextBridge, AI for Good at: developing synthetic cancer data 283; key learnings and recommendations **284**; machine learning 278; mapping Brazilian food system 282–283; problem-solving using NLP for Red Cross 278–281, 280; rule-based AI 278; using data analytics to decode dance patterns of honeybees 281–282
- Amodei, D. 414, 415
- Amos, B. 423
- Andreessen, Marc 505, 513
- Andreoni, J. 107
- Aoki, N. 154
- Aoun, J. E. 159
- Application Programming Interface (API) integrations 362
- Aras, A. 194, 195
- Arbor Day Foundation 297
- Aristotle 510
- Arpe, B. 195
- artificial general intelligence (AGI) 412; accelerationism 508–509; defenders of 503–504; defined 503–504; effective accelerationism (Eff/acc) 505; effective altruism 505–508; free and open-source approach 512; free software movement and open source 514–515; GNU project and free software movement 512–513; Netscape project and Open Source Initiative 513–514; sacralization of 511; TESCREAL framework 504–505; theological-political perspective 503
- artificial intelligence (AI) technology 11; acculturation of civil society 446–447; adoption in philanthropy 329, 341–342; amplifying civil society voice and impact 335–336; challenges 345; common good 446; complexity and opportunity 533–534; defined 160, 160, 487–489; development of 345; digital justice, new social compact for 337–338; and digital landscape 341; disruptive technology 434; donors and NGOs 361–362; emergence of 328, 433; emotive content 458; gaming-based philanthropy 133, 135, 137; Generative AI 458; global ethical frameworks 341; growing accessibility of 433–434; harnessing power for impact (case studies) 336–337; immersive environments 137; impact of video games on 133–135; and information provision 455–456; as investors 545; joint perception, concept of 133; levels of impact 522, 522–523; limited data infrastructure 330; low internal digital capacity 329–330; mission 435; notion of joint perception 135–136, 136; open source and inclusivity movements 445; opportunities and responsibilities 342; optimizing philanthropic operations 332–333; pioneering new models of philanthropic partnership 333–335; practical recommendations 338–339; precautionary principle 443–444; principle of democratic participation 441–442; principle of diversity and inclusion 443; principle of fairness 442; principle of responsibility 444; principle of sustainable development 444–445; privacy and intimacy 439–440; in private sector 341; for public services 362; reasoning steps 523, 523–524, 524; scales of disruptions 525, 525; social sector transformation 331–332; societal inclusion 437–438; solidarity principle 440–441; stakeholder relations 528–529; trust and buy-in from communities 331; and virtual worlds 132–133; willingness and engagement from leadership 330
- Asia’s social sector: AI readiness frameworks 207–209; data analysis and insights 209–216; existing frameworks 207–208; foundational readiness 209–212; increasing project and operational funding for ICT investments 217; make use of ample opportunities for in-kind donations 218; operational readiness 212–213;

- recommendations for funders 216–218; role of AI 205; for social delivery organizations 208, 208–209, 213; supporting SDOs in developing AI guidelines 218; transformational readiness 214–216; wealth generation in Asia 206–207
- Association of Southeast Asian Nations (ASEAN) 216
- authoritarianism: global implications beyond Chinese case 234–235; sectoral and global implications of 234–235; table of interviews 235
- Baldwin, T 250
- Bankman-Fried, Sam 422, 507
- Barney, J. B. 364, 371
- Barry, S. 503
- Bartalucci, Simone 2
- Barzanti, L. 3, 110, 111, 116
- Beast Philanthropy's* website 301
- Beckham, David 463
- Bednarek, A. 479
- Behl, A. 362
- Bekkers, R. 453
- Bender, E. M. 397
- Bengio, Y. 416
- Benvenuti, L. 3
- Berners-Lee, T. 314, 346
- Bernholz, L. 15
- BERT model (Bidirectional Encoder Representations from Transformers) 279
- Best Available Charitable Option (BACO) model 89
- best practice checklist 532–533
- best practices and safeguards 531, 532
- Bhati, A. 494
- Biden, Joe 548
- Bill and Melinda Gates Foundation (BMGF) 324, 379, 480
- biometric categorization 376
- Birhane, A. 150, 547
- Biskaabiiyaang 138
- blockchain 84, 84–88, 85, 87
- Bondi, E. 399
- Bostrom, N. 412, 462, 506, 507
- Botti-Lodovico, Yolanda 5
- Broad, G. M. 421
- Brussels effect 376
- Budziszewska, Anita 3
- business intelligence (BI) 277, 284
- Butler, J. E. 371
- Büyüközkan, G. 194, 195
- BWA 88
- Cagala, T. 495
- Campos, Siméon 6
- Cantin, R. 146
- Carnegie, Andrew 13, 226, 478
- Casagrande, Margaux 7
- Caselli, D. 108, 119
- cause-based organizations 462–463
- Centre for Asian Philanthropy and Society (CAPS) 205
- Chacaby, Maya 138
- Chakraborty, Dipam 5
- Chang, C. 107
- ChangeFinder 363
- Chan Zuckerberg Initiative 480
- Charanya, T. 194
- Charity Digital Skills Report* 94
- 2016 Charity Law: the paradox of legal recognition and regulatory containment 224–225
- chatbots 520
- “Chat Generative Pre-trained Transformer” (“ChatGPT 3.5”) 520
- ChatGPT, OpenAI’s 132, 360, 403, 520
- Chen Yidan 225–227
- Chief Technology Officers (CTOs) 539
- children’s palliative care (CPC) impact modeling project 38, 39, 40, 40; criteria used to analyze 41, 42; framework dimensions per stakeholder 47; list of QoL and outcome frameworks 45–46; QoL of stakeholders 40, 41; ranking of QoL measurement frameworks 41, 43; rankings of frameworks, most recurring dimensions 48; research methodology 43; similarity tool 41, 44, 44
- child sexual abuse material (CSAM) 381
- China, digital philanthropy in: according to Chinese tech giants 225–231; AI applications in Chinese non-profit sector 231–233; under authoritarianism 234–235; 2016 Charity Law: the paradox of legal recognition and regulatory containment 224–225; crisis of trust 221; “Internet+ philanthropy”: platform power extending to the non-profit sector 223; political-economic context 222; prevalence of fundraising foundations 236n1; rise of internet philanthropy 221; for smart philanthropy 222–225; state and corporate power in 223; transnational dynamics 222
- Chinese non-profit sector: AI-based project recommendation 231–232; China’s “Philanthropy 3.0 Era”: unquestioned synergies between AI and philanthropy 233; machine learning against machine fraud 232–233
- Chinese tech giants: “99 Giving Day”: Tencent’s digital upheaval of philanthropy 227–229; individual fundraising success 229–231; between philanthropy and CSR 225–226; technological innovation: “philanthropy for everyone” 226–227
- choice: and AI 454–457; architecture 456; cause-based organizations 462–463; defining characteristics 452; general-purpose search

- and recommendation services 458–459; giving platforms 459–462; micro-level factors 453; nonprofit platforms 454; and philanthropy 452–454; research 453
- Christensen, Clayton M. 63, 64, 68
- Christiano, P. 414, 415, 507
- Chu, H. 150
- Chui, M. 198
- Chung, J. 494
- civil society organizations (CSOs): charitable organizations 94–95; Global South vs. Global North 95, 99–102; Swiss 96–99; Swiss digital civil society 95, 96; use of AI 94–95
- Clark, James 513
- Climate Trend Scanner 88
- coding process 162; tech-centered 162; tech-implementing 162; tech-perpetuating 162
- Collaboration Governance Framework (CGR) 316, 316; legal challenges in accessing, reuse, and resharing of data 319–321; reducing cost of compliance to adhere to AI principles 317; reducing cost of liability risks 316–317; UNESCO ethics of AI principles 317, **318**
- collective intelligence (CI) 344; in advancing and improving AI 348; altruistic 345–348; cryptography 346; democratized control over development of AI systems 344; development of AI 345, 347–348; expansion of Internet and World Wide Web 346; financial contributions 347; open-source software movement 344, 345–346, 347
- common good 446
- Community AI (2023) 167
- Concannon, Shauna 3
- Cool, K. 366
- Cordelli, C. 394, 452
- Cordery, C. J. 194
- corporate philanthropy 363; corporate foundations 311; corporate giving 311; corporate volunteering 311; international 363
- corporate social responsibility (CSR) 276, 302
- Covid-19 pandemic 328; emergency relief programs 236n7
- Crary, A. 506–508
- creator ethics 491–494, 496–497
- crypto-donations 526
- cryptography 344
- Cucinelli, D. 120
- cyber sovereignty 236n4
- Danaher, J. 424
- Dartmouth Conference 379–380
- Das, E. 494
- data: collaborative 314; donation 314; sovereignty 236n4
- database (DB) technologies 109
- ‘data invisible’ issue 309
- data philanthropy (DP): defined 310–311; emergence and evolution of 311; integration with corporate philanthropy *see* corporate philanthropy; IOs as stakeholder 311, **312**; IO’s role as 324; recommendations based 323, **323**; risk management framework, impact of 323–324
- data science technology (DST) 159, 160, 160
- Davies, R. 6, 232
- Deb, Anamitra 6
- Decentralized Autonomous Organizations (DAOs) 76, 84, 84–88, 85, 87, 91
- decentralized finance (DeFi) 85
- decision support systems (DSSs) 107, 110, **111**
- deep reinforcement learning (DRL) 89
- De Laubier, R. 194
- Deleuze, Gilles 508
- De los Ríos, A. 2
- democratic participation, principle of 441–442
- Demsetz, H. 369
- Deroy, O. 135
- Dhar, V. 5, 339
- Dierickx, I. 366
- digital culture 202
- digitalization: enhancing foundations’ efficiency and impact 196–197; safely leveraging AI in non-profit sector 197–202; of Swiss non-profit foundations 192
- digital philanthropy 314
- digital public infrastructure (DPI) 324; enabling governance 324; innovative community and market players 324; networked open technology standards 324
- digital stunt philanthropy (DSP): algorithm-friendly content 295–296; algorithmic learning 287; audience engagement on social media 295–; celebrity-based philanthropy 294–295; components of 293, **294**; ice bucket challenge 294; integration of AI 287; online stunts 293–294; philanthropic activity of 293; social media content creators (influencers) 294; social media platforms 287; sponsor-friendly content 296
- digital transformation (DT) 192; across social sector 339; among grantees 340; cost of 329; key enablers and blockers 195–196; as philanthropy 339; and Swiss non-profit foundations 192–196
- digital twin 79–80
- Disaster Relief Emergency Fund (DREF) 278–279
- distributed ledger technology (DLT) 85, 91
- diversity, principle of 84, 443
- Doerr, John 21
- Doing Good Index 2024 205
- Donaldson, Jimmy 296, 297, 300
- donor-partner matching 261–262; business implications 265; data collection and cleaning 262–263; deployment and monitoring

- 264; exploratory data analysis 263; feature engineering 263; first results 264; limitations 265; model development and training 263–264; model interpretation and analysis 264
- DonorSearch 362
- Dorsey, Jack 14
- Dotan, R. 7, 101
- DrivenData 349
- Dupuy, J. P. 511
- Dweck, C. S. 63
- Dyke, N. Van 423
- earnings before interest, taxes, depreciation, and amortization (EBITDA) 276
- effective accelerationism (Eff/acc) 505
- effective altruism (EA) movement 380, 413–414, 418
- effective philanthropy 481
- EleutherAI 376
- Emerson, K. 324
- enlightened self-interest 107
- ensemble neural network (ENN) 89
- ethical AI 435
- ethical and inclusive AI (EIAI) systems 11;
  - accountability and legitimacy deficit 473–474;
  - adoption of 312; bias 474–475; collaboration governance framework 316; counterfactual scenario 472; data as strategic asset 315;
  - data protection measures 313; defined 312;
  - deployment of AI systems for POs 313–314;
  - development of 521; explainability 315; financial sustainability 472; governmental bodies 472;
  - inclusiveness in AI ethics 313; misinformation 475; need for 315; objectives and research questions 309, **310**; philanthropic entities 472;
  - philanthropic funders 471; POs as catalysts 309; principles 312; prohibitions 313; public funds 471; technological revolution and ethical challenges 474; transparency 315; unjust actions 475; value alignment 475–476
- ethics checklist 531
- ethics of teachers and trainers 491–492
- EU AI Act 386–387
- Europe: advancement of AI 187, 188; AI and data science, non-core activity 183–184; AI tools 185; challenges in adopting AI and data science practices 181–185; dependency of foundations on tools 186; environmental impact 183; exploring unknown safely *184*, 184–185; foundations’ engagement with AI, internal or external activities 179, *179*; foundation staff 187; institutional philanthropy 187–188; internal barriers 186; lack of skilled talent 181–182, *182*; by philanthropic organizations 177–179, 187, 188; public debate on AI with opposing and balancing views 172–174; safety and regulation 182–183; unique role philanthropy, AI conversation 174–176
- Expert Group on Social Economy and Social Enterprises (GECES) 31
- Extended Gaming Literacy 141–142, 143
- eye-catching thumbnails 299–300
- Facebook 288
- fairness, principle of 442
- Falajiki, C. 548
- family-friendly branding strategies 300
- Farrell, J. 362
- Farrokhvar, L. 495
- fast-tracking AI adaptation in philanthropy 68–74; concerns and chances of technology implementation **73**; embarking on learning journey 72, 74; strategies and influencing techniques 68–69; understanding root causes of Swiss actors’ concerns 70–72, **71**; use cases for applied AI in grantmaking process 69–70
- Financial Attitude Index (FAI) 120
- Financial Behavior Index (FBI) 120
- financial education 108
- Financial Knowledge Index (FKI) 120
- financial literacy 108; adult population 120; children and young persons 120; education programs 119; phenomenon of financialization of Welfare 119–120; problem resolution 124–126, *125–127*; problems 121–124, *122–124*
- Fine, A. 18
- Finzi, D. 137
- Fitzgerald, B. 346
- Florida Waterway Health Forecast 167
- Florida, L. 14, 15, 18, 22, 554, 557
- Food Recognition Challenge (AICrowd): case study 358–360; challenge and benchmark phases 352, 352; CI coopetition 351; collaborative annotation process 350; coopetition 350; data and method 350–353, *351*, 352; machine learning models 350; maximum precision 353; modern AI-driven challenges 349–350; relation between variables used for modeling 360; results 353–354; standard deviation of submission precision 354, 354; visual inspection 351, *351*
- Ford, Henry 13
- for-good framework 167, *168*
- Forth, P. 194
- foundational readiness, China 209–212; access to data 211, **211**; access to digital infrastructure 210–211; access to software 211–212; staff access to computers and/or tablets **210**; *see also* China
- foundation models: adaptations 397; balancing acts 401; bridging-based ranking 402; concentration of power 400–401; deliberative alignment 401–402, 404; democratic accountability 404;

- democratic inputs for AI 403; emergence 396; homogenization 396; large-scale impact on matters of public interest 397–399; limited stakeholder control 399–400; natural language models 396; participatory design 403; policymaking 402; Polis 402–403; public accountability 403–404; public control and accountability 404; qualitative participatory spaces 401; quantitative participation 401–402; technical and sociological inflections 396; technology-enabled participatory spaces 404; use of deliberative technologies 403
- foundations: alignment with donors 481–482; clarifications 394; closing accountability gap 480–481; cross-sectoral and cross-cultural dialogues and public deliberation 476–477; democratic and market accountability 395–396; elite philanthropy 396; foresight and anticipation 482; funding responsible, equitable research 479; lack of democratic legitimacy 481; perspective of democratic values 395; philanthropic organizations 393–394, 476; political equality among citizens 394–395; private and public sectors and civil society 477–478; promoting and investing in public goods 479–480; sponsoring 21st-century AI literacy education 478; transparency and public accountability 394; use of philanthropic funds 395
- free and open-source software (FOSS) 512, 515
- free/libre and open-source software (FLOSS) 515
- Freud, Sigmund 137–138, 142
- fundraising (FR) 107; campaigns 78, 108; capabilities for nonprofits 526; efforts 78; strategy 108
- future of AI and philanthropy 272–273, 555–556
- Future of Humanity Institute (FHI) 412
- Fuzzy System (FS) 110; adoption of MIN operator 115; campaign results 118–119, 119; challenges 116–117; data collection in DB 114; donations for each Donor 114; evaluation of 115, 115; fuzzification process 115–116, 116; fuzzy aggregation and whole ranking 118, 118; graphical user interface 117, 117–118; implementation of 117; quantitative information of Donors 113–114; results of first ranking phase 118, 118; statistics of Donors' profile characteristics and gift history 114, 114, 115
- Galjak, Marko 4
- Gamers Outreach Foundation 138
- gaming-based philanthropy: appreciating role of games in social connection and mental health 141; designating internal gaming specialists 139; exploring gaming for conflict resolution 142; future of AI 138–139; Gamers Outreach Foundation 138; games as form of art 140; games as therapeutic tool 140; implementing tracking system for games and proto-metaverses 139; Indigenous metaverse: Biskaabiyang 138; metaverse and metapsychology, The Sigmund Freud Museum 137–138; pathway to Extended Gaming Literacy 141–142, 143; predictive analytics in fundraising 140; promoting ethical gaming practices 142; recognizing games as educational tools 141; technological innovations within gaming 142; understanding gaming culture and demographics 141; using AI to enhance donor engagement and personalization 141
- Gates, Bill 13, 132, 379, 381, 480
- Gates, Melinda 13, 132, 324, 379, 381, 480
- Gaudenzi, Enrico 3
- Gebru, T. 173, 423, 504, 507
- Genbru, Timnit 547
- General-Purpose AI systems (GPAI) 376
- Generative Adversarial Networks (GANs) 283
- Generative AI 458
- Generative Pre-Trained Transformers (GPT) 556
- Geneva Centre for Philanthropy (GCP) 539, 554
- George, J. 309, 310
- Germann, S. 6
- Ghosh, S. 323
- Gil-Garcia, J. R. 314, 316, 317
- Gilroy, L. 368
- G20 India 2023 Summit 528
- Giovampaola, C. D. 2, 7, 476
- Giove, S. 110
- Girard, R. 511
- GIT 345
- giving platforms 459–462
- giving pyramid, Donors' segmentation 109, 109
- Global Fishing Watch, Inc. (2023) 167
- Global Philanthropy Tracker (GPT) 25n2
- Global South vs. Global North: AI's cultural dimension 101–102; universality, equality, and CSOs 100–101; use of digital technologies 99–100
- Godfrey, P. C. 311
- Goldin, C. 496
- Gomez, L. 5, 7
- Good Old Fashioned AI (GOFAI) 461
- Google 287
- Google AI Impact Challenge 25n6
- grantmaking: funding (responsible) AI-enabled non-profits 541; grantees using AI (responsibly) 541–542; harms to grantees 543; participation barriers 542; risk of harm 542–543
- graphical processing units (GPUs) chips 345
- Green AI Foundation (2023) 166–167
- Greenstein, B. 197
- Guattari, Félix 508

- Halopé, H. 4, 7  
Hansen, R. 494  
Harris, D. E. 6  
Harris, Kamala 174  
Hart, S. 363  
Hassabis, D. 413  
Hayek, Friedrich 511  
Heimstädt, M. 314  
Henriksen, S. E. 21, 22  
Herrero, M. 3, 147  
Herzog, P. S. 3, 174  
Hewage, Kithmina V. 4  
Hinton, Geoffrey 416, 423  
HireVue 81, 82  
Hogg, E. 301  
Horgarth, I. 510  
Houghton Budd, C. 108, 120  
Hou, J.-R. 499  
*HuggingFace* 348, 376  
Human Resources (HR) 76  
hybrid intelligence 435, 448; challenges of 447–448
- ICON-NGO 440  
Illingworth, P. 394  
immersive fundraising events 80  
impactIA 435–437; AiiA festival 447–448  
Impact Management Project (IMP) 30  
impact measurement: AI support for indicator selection 34–35; benefits of 29–30; challenges 30; defined 29; defining normative basis 32; development of proxies (indicators) 33–34; empirical testing and determination 34; evidence-based philanthropy and strategic philanthropy 30; factors contributing 29; generic process 30–31; identification of commonly used items 36–38, 37; identification of impact chains 32–33, 33; impact modeling in CPC (case study) *see* children’s palliative care (CPC)  
impact modeling project; IOOI model 32, 33, 40; matching prototype and data indicators 35–36, 36; methodological approaches 30; modeling steps 31, 31–34  
inclusion criteria 161, 161–162, 443  
inclusive capitalism 363  
information and communication technologies (ICTs) 321  
information provision 455–456  
information technology (IT): adult population, financial literacy in 120; children and young persons financial education 120; financial education programs 119; financial literacy problems 121–124, 122–124; high-end web application for Excel Lab (case study) 126–129, 128–129; phenomenon of financialization of Welfare 119–120; problem resolution 124–126, 125–127; use of digital tools 120–121; use of Excel 121  
innovation diffusion, common models for 61–63, 62  
Instagram 288  
intellectual property (IP) 344  
Interim International AI Institution (IIAI): 2024 AI Act 376; AI ethics and G7’s Hiroshima AI Process 384; AI harms 384; benefits of AI 374; collaborative standards development 385–386; coordinated public interest AI development 386; creating weak patchwork of laws 375; inclusive representation and participation of civil society 380–381; innovation killer 375; institution 384; lawmakers 384; pacing problem 375, 377; philanthropic resources 377–378; providing risk capital for public policy development 381–382; racing against the race 383–384; rapid-response harmonized policy development support 385; regulation of technology 374; succession planning 386; supporting shared infrastructure and new institutions 382–383; Ted Turner’s billion-dollar United Nations Gift 378–379; U.S. Congress 375  
International Federation of the Red Cross and Red Crescent (IFRC) 278–280, 280  
international organization (IO): analysis 321–322; case study questions 321; ICT development 321; limitations 322–323; problem 321; World Telecommunication/ICT Indicators Database 321  
Internet 344  
internet philanthropy: platform power extending to non-profit sector 223; rise of 221
- Jack Ma 227  
Jacobi, E. 336  
Jasper, Ulla 6  
Jewell, A. 526  
Jha, Rahul 5  
Jha, S. 6  
Jobin, A. 174  
joint perception, notion of 135–136, 136  
Jones, P. M. 290
- Kaggle 349  
Kane, G. C. 62–64, 68, 194  
Kanter, B. 18  
Kant, Immanuel 452  
Karnofsky, H. 414, 415  
Kaul, A. 367  
Keil, T. 367  
Khastgir, Prity 3  
Kilicalp, Sevda 4  
King Baudouin Foundation (KFB) 543  
King, Martin Luther, Jr. 342



- Kleinberg, J. 398  
 Knowledge Choquet (KC) 110  
 Knowledge Fuzzy Mathematical (KFM) 110;  
   architecture of 111, *112*; fuzzy evaluation of the  
   gift probability 113, *113*; mathematical model of  
   dynamic evaluation of strategies in 111, 113, *113*  
 Konstantinou, I. 508  
 Konya, A. 402  
 Kore.AI 80  
 Kraemer, S. 147  
 Kratz-Ulmer, A. 4  
 Kriege, John 478  
 Kurmann, P. 195  
  
 Landau, S. 346  
 Landgrebe, J. 503  
 Lang, B. 4  
 large language model (LLM) 149–150, 267–268,  
   521; embeddings 269, 269–270; fine-tuning  
   267; limitations 270; pre-training 267;  
   retrieval-augmented generation 268–269;  
   RLHF (Reinforcement Learning from Human  
   Feedback) 267  
 Latonero, M. 100  
 Lau, J. H. 250  
 Lazer, D. M. J. 316, 317  
 Lea, M. 5  
 Lee, L. 107  
 Lee, S. 494  
 Legg, S. 413  
 Lessard, Madeleine 2  
 lethal autonomous weapons (LAWs) 376  
 Lev Aretz, Y. 311, 314, 319  
 Levy, S. 346  
 LinkedIn 288  
 LinkedIn Recruiter 81  
 Linked Open Data (LOD), Berners-Lee’s model  
   of 314  
 Linux kernel 346  
 Liu, S. 150  
 Longin, L. 135  
 Longoni, C. 151  
 Luo, J. 367, 370  
 Lusardi, A. 120  
  
 Maas, M. M. 423  
 MacAskill, W. 506, 507  
 Machine Intelligence Research Institute (MIRI) 412  
 machine learning (ML): algorithms 77; defined  
   160, *160*  
 Maillart, T. 5  
 Makridis, C. A. 195  
 Maricic, M. 4  
 Marx, Karl 508  
 Mastroleo, M. 110  
 McCarthy, John 12, 278  
  
 McKinsey & Company 520  
 Medical Automation Org Inc (2023) 167  
 Meta 287, 376, 516  
 Michel, P. 146  
 Michelson, E. 482  
 Milgram, Paul 135  
 Milinković, Nikola 4  
 Miller, V. 301  
 Mistral AI 376, 377  
 modern times philanthropy 521  
 Mönks, J. 7  
 Monte, Mara de 7  
 Montreal Declaration 436  
 moral agency 490–491  
 Moreno, P. M. 99  
 Morley, J. 317  
 Moro, S. 111  
 Morris, M. R. 503, 504  
 MrBeast and Beast Philanthropy, case of:  
   background 296–297; ethical issues and  
   implications 301–302; excitement and  
   admiration for 303; impact 301; methods 298–  
   301; rise of 303  
 Muhr, Antonia 2  
 multi-scale model 534  
 Musk, Elon 21, 374, 414, 416, 512  
   “*My Mentor is a Woman*” program 437  
  
 Nagar, R. 337  
 Nanavati, Sneha 5  
 Nandeshwar, A. R. 526  
 Nardon, M. 116  
 National Cancer Institute 283  
 natural language processing (NLP) 288; algorithm  
   33; comparing MrBeast and Beast Philanthropy  
   YouTube channels 298, 298; UMAP  
   network-based reduction 298  
 Net Promoter Score 276–277  
 NeurIPS (AI scientific conference) 345  
 Ng, Andrew 547  
 non-consensual intimate imagery (NCII) 374  
 non-fungible tokens (NFTs) 526  
 nonprofit fundraising: advent of generative AI 149;  
   AI revolution 146; behavioral capabilities  
   147–148; challenges 145; cognitive capabilities  
   147; concept of organizational resilience 145–  
   146; during Covid-19 146–147; emotion-related  
   capabilities 149; external crisis 146; fabrications  
   and falsifications (hallucinations) 149–150;  
   human feedback reinforcement learning 150;  
   issues raised by AI 146; Large Language  
   Models (LLMs) 149–150; public generosity and  
   support 145; RAI *see* Responsible AI (RAI);  
   relational capabilities 148; Stable Diffusion  
   image generation model 150; trust and public  
   perceptions of AI 150–151

- non-profit sector: AI adoption in industry 198;  
 Chinese 231–233; internet philanthropy 223;  
 overview 197–198; risks and challenges 198–199
- Norwegian Cancer Registry 283
- Noys, B. 508
- Nuova Civilt-delle Macchine* (NCdM) 127
- O'Brien, C. 295
- O'Hear, Jonathan 443, 447
- Omidyar, Pam 381
- Omidyar, Pierre 381
- O'Neill, Jack 4
- online platforms: authenticity and trustworthiness  
 291–292; diversity and quantity of 288;  
 emotional engagement 290; identification  
 through vision and value alignment 291;  
 inspirational content 292; long-term value  
 and sustainability 292; relationship building  
 291; social validation 291; YouTube's AI  
 recommendation system 288–289
- Open Data (OD) 314
- Open Government Data (OGD) 314
- open-source software (OSS) movement 344
- operational readiness, China: cybersecurity  
 212–213; internal skills and expertise 212
- Ord, Toby 506
- Organisation for Economic Co-operation and  
 Development (OECD) 108
- Pakura, A. 3, 133, 141
- PaLM 2, Google's 360
- Panzar, J. C. 365
- Park, G. 153, 494
- particle image velocimetry (PIV) software 282
- Patrick J. McGovern Foundation 341
- Percia David, D. 344
- Peter, Henry 7
- Pfeffer, Jeffrey 63
- Philanthropic DAOs 86
- philanthropic organizations (POs) 11, 76–78, 521
- philanthropic partnership: collaborative data stores  
 and standards 334; fostering technical talent and  
 AI literacy 335; in-house capacity building and  
 product development 333–334
- PHIL4DEV: addressing category imbalance 55;  
 classification limitations 57–58; CRS  
 classification system 52–54, 53; data limitations  
 57; data preparation 54–55, 55; estimation and  
 validation 55–56; future of 56; modern tools  
 59; new NLP tools 58–59; NLP model 51–52;  
 OECD Centre on Philanthropy 51; OECD's  
 Creditor Reporting System 51, 52–54; Official  
 Development Assistance statistics 51; prediction,  
 transparency, and availability to the public 56,  
 56; public accessibility to 57; technological  
 limitations 58–59; workflow 54, 54
- Potluka, O. 3, 97
- Powell Jobs, Laurene 381
- Prahalad, C. K. 363
- precautionary principle 443–444
- precision philanthropy 522
- Pressgrove, G. 293
- Priem, R. L. 371
- privacy: education and awareness 439; ethical  
 audits and certifications 439; legislation 439;  
 partnerships with cybersecurity experts 440;  
 promoting open source and interoperability  
 439–440; support for legal action 440
- probability of sponsorship 300
- prototype indicators (PIs) 35–36
- Prunkl, C. 423
- public-private partnerships (PPPs) 367
- Pymetrics 81–82
- quality of life (QoL) 38, 40, 41, 41, 43, 45–46
- Raddon, M.-B. 497
- Radocchia, S. 134
- Raghavan, M. 398
- Rapaport, W. 503, 504
- Raymond, E. S. 513–514
- Recommended System Fuzzy (RSF) 110
- Reich, R. 15, 394, 481
- Research Institute of Sweden (RISE) 88
- ResNet 350
- resource-based view (RBV): agency theory 369;  
 contingencies for AI 366–367; COVID-19  
 research 371; for donating organizations 370;  
 donor organization 368; inclusion of dynamic  
 capabilities 365–366; lack of enthusiasm  
 368; management literature on firm 315, 363;  
 NGO relationships 367–371; public-private  
 partnerships 367; staff responsible for  
 administering AI donations 369–371; theory  
 development 364; Transaction Cost Economics  
 369–370; VRIN framework 364–365
- responsibility, principle of 444
- responsible AI (RAI) 435–436, 539; vs. AI 540–  
 541; cognitive and behavioral resilience 152–  
 153; emotion regulation resilience: from distrust  
 to trust 153–154; empowering (responsible) data  
 analysis 543–544; enabling ROI 545; fund RAI  
 and organizations and activities 549; lack for  
 philanthropic investors 546; lack in grantmaking  
 542; lack in using, buying, and developing  
 AI 544; lagging behind 544; negative impacts  
 544; positioning philanthropic organizations as  
 leaders 545–546; public resources customized  
 for foundations 549; reputational harm and  
 non-compliance 544–545; resilience capabilities  
 framework 146; scale impact through optimizing  
 internal operations 543; self-educate and

- self-regulate 549; shaping AI ecosystem 545; in storytelling 151–152; tools for public use 543–544; Vet and support vendors, non-profits, and portfolio companies 550
- Responsible AI in Philanthropy Guide* 557
- Responsible Artificial Intelligence Institute (RAII) 218
- Richey, L. A. 21, 22
- Ricoeur, P. 512
- Rinaldi, E. 120
- risk mitigation strategies and safety mechanisms: access 199; accountability 199, 201; AI governance 199–200, 201; bias 198, 200; cyberattacks 199, 201; environment 199; infrastructure 199, 201; talent 199, 201; training 199; transparency 199–201; use of AI in foundations 200–202
- Rober, Mark 297
- robot-me* project 437, 438
- robust dataset 259–261
- Rockefeller Foundation 12, 379, 380, 480
- Rockefeller, John D. 13
- Rogers, Everett M. 61–64, 66–69
- Roy-Chowdhury, R. 546
- Ruixia, Y. 111
- Ruocco, F. 108, 119
- Sappho, Maria 447
- Sætra, H. S. 424
- Sauer, Sina 2
- Saunders-Hastings, E. 394, 395
- Savas, E. S. 368
- Schewick, B. van 346
- Schiff, D. S. 6
- Schipper, M. 3
- Schmidt, J. 367
- Schöbi, Stefan 2
- Schyns, C. 547
- Science and Technology Studies (STS) 474
- Scott, Mackenzie 14, 340
- Searing, D. R. 7
- Searing, E. A. M. 7
- Sellen, C. 7
- Shalini, Shweta 3
- Shamsrizi, M. 3, 133, 141
- Shapiro, R. A. 206
- Shapiro, S. 508
- Sharada, Mohanty 5
- The Sigmund Freud Museum 137–138
- Singer, P. 421, 452, 505–508
- Sloane, M. 399
- Smith, Adam 511
- Smith, W. 107
- Smulowitz, M. 5
- social equality 540
- social impact via Sustainable Development Goals (SDGs) 76
- social justice advocates: Big tech’s influence on AI regulation 547–548; irrelevance, loss of trust, and breach of social contract 548–549; lack of RAI engagement 548; negative social impact 548; politics and geopolitics in AI regulation 548; RAI opportunities for philanthropists 547
- Social Return on Investment (SROI) analysis 34
- social scoring 376
- social sector transformation 331–332
- societal inclusion 437–438
- Society for Automotive Engineers (SAE) 491
- Solaiman, I. 516, 517
- solidarity principle 440–441
- Spandows: AI for good at Amesto NextBridge 278–284; Amesto Group’s expansion 277–278; and family business 275–278; triple bottom line 275–277
- Spiers, D. 4
- Spira, Henry 421
- Srnicek, N. 509
- Stability AI 376
- stakeholders: checklist 529; levels of impacts on 530
- Stallman, R. 512–515
- Stamler, J. 78
- Stix, C. 423
- Stockholm Environment Institute (SEI) 282
- strategic philanthropy 481
- Strubell, E. 183
- Sunstein, C. R. 456, 457
- Susha, I. 314, 316, 317
- Sustainable Development Goals (SDGs) 76, 527
- Sustainable Development Goals Philanthropy Platform (SDGPP) 88
- sustainable development, principle of 444–445
- sustainable investing 266–267
- Sutton, Robert I. 63
- Swiss actors’ concerns 70–72, 71
- Swiss digital civil society 95; and digital technologies 96; financial capacities and efficiency 96–97, 97; fragmentation of networks 97–98; question of use of AI to achieve increased efficiency 98–99
- Swiss Federal Railways (SBB) 440
- Swiss non-profit foundations: definition 193–194; legal principles 193; opportunities and risks 194–195; overview of foundations 192–193
- Swiss philanthropy 65–66, 65–67, 67
- synthesis 495–496, 499
- systemic change: about Altruist League 258–259, 260, 261; field of philanthropy, transformations in 21st century 257; grassroots movements 259; principles of systemic philanthropy 258

- Taddeo, M. 15, 17  
 Takam, Ezekiel K. 7  
 Tan, Garry 505  
 #TeamSeas 297  
 #TeamTrees 297  
 tech for-good philanthropy case studies 165, 167, 168, 168  
 tech integration styles: tech-centered 162, 163–164; tech-implementing 162, 166–167; tech-perpetuating 162, 164–166, 165  
 technology adaptation 61–67; AI adaptation in Swiss philanthropy *see* Swiss philanthropy; common models for innovation diffusion 61–63, 62; four stages in technology diffusion 63–64, 64  
 technology-advancing philanthropic activities 169  
 technology diffusion, stages in 63–64, 64  
 Technology Innovation Institute 376  
 technology stack: data collection 270–271; data processing 271; data storage and databases 271; hardware 271; predictive modeling 271  
 Teece, D. J. 365  
 Tegmark, M. 414  
 Tencent Foundation (2021) 236n9  
 TESCREAL framework 504–505  
 Tescrealist movement 442, 445  
 Textio 83  
 Thaler, R. H. 456, 457  
 Then, V. 2  
 Thiel, Peter 21, 413, 414  
 Thomas, Nisa 7  
 3P (People, Profit, and Planet) 281  
 TikTok 287, 288  
 Tocmacov, Laura 6  
 Todesco, L. 120  
 Topcoder 349  
 Transaction Cost Economics 369–370  
 transformational readiness, China: government regulation of AI 216; leadership buy-in: access to operational funding 214; technology-based service provision 214–215, 215; trust in social sector 215–216  
 Transfr 83  
 Trase 282  
 trending topics 300  
 Trump, Donald 548  
 Tseng, V. 479  
 Tudor, M. C. 2, 7  
 Turner, Ted 378–379, 381, 387  
 Ugazio, G. 7, 137, 476  
 Unabot 81  
 UNDP SDG Impact Standards for Enterprises 30  
 UNESCO 341  
 UN Global Digital Compact 528  
 United Nations Development Programme (UNDP) 88  
 United Nations Foundation (UNF) 378  
 United Nations Fund for International Partnerships (UNFIP) 379  
 United Nations’ Sustainable Development Goals (UN SDGs) 95, 282, 461  
 Université catholique de Louvain (UCLouvain) 282  
 University of Geneva (UNIGE) research 554  
 UN SDG Impact 30  
 user ethics 494–495, 497–498  
 U.S. National AI Research Resource (NAIRR) 385  
 VAIOT 80  
 valuable, rare, inimitable, and non-substitutable (VRIN) framework 315, 363–365  
 Value Alignment Problem (VAP) 462  
 Vanvalkenburgh, P. 78  
 venture philanthropy 481  
 Verhulst, S. 317  
 VGG 350  
 Vial, G. 194, 195  
 virtual aid and support services to beneficiaries 80  
 virtual community assessments 80  
 virtual reality (VR) modules 79  
 Vogel, P. 5  
 Vohra, S. 198  
 Voids, A. 314  
 volatile, uncertain, complex, and ambiguous (VUCA) world 283  
 Wadhvani Institute for AI Foundation (2023) 167  
 Wahed, M. 547  
 Wang, H. 363  
 Waters, R. D. 290  
 Weidinger, L. 149  
 Welbl, J. 397  
 well-being, principle of 436–437  
 Wernerfelt, B. 364, 366  
 Western Balkans: adopting Doc2vec and Word2vec 250; AI-assisted process of data collection 248–251; article processing and model parameters 250; Catalyst Balkans 242–243; CiviGraph 244; cost of AI 252–253; data collection and methodology 245–247; data harvesting 247; embeddings 250–251; future AI integration 251–252; Giving Balkans creation 243, 245; Giving Balkans database 243, 244; Giving Balkans database, problem of 245; insights and implications 253–254; integration of the knowledge 253; limitations of methodology 247–248; problem of categories 246–247; problem of duplicates 247; problem of languages 246; problem of truth 247; problems with AI 252–253; rapid innovation and changes to LLMs 253; relevance classifier 251; resulting solution

## *Index*

- 251; rugged philanthropic landscape 240–242;  
text processing 248–250, 249; topic-based  
clustering and modeling 250
- WhatsApp 288
- Wheeler, Joe 5
- Whittlestone, J. 423
- Wiepking, P. 453
- Williams, A. 509
- Williamson, O. E. 369
- Willig, R. D. 365
- Winnicott, D. W. 140
- Winter, S. G. 365
- Wollstonecraft, Mary 452
- World Economic Forum (WEF) 556
- World Health Organization (WHO) 379
- World Wide Web 344
- World Wildlife Fund (WWF) 526
- “X-risk” movements 380
- X/Twitter 288
- Yescombe, E. R. 369
- Yeung, K. 457
- YouTube 288; AI recommendation system 288–289,  
289; audience engagement with social media  
290, 290; candidate generation 289, 289; levels  
of engagement 290; ranking networks  
289, 289
- Yudkowsky, E. 412, 413
- Zamani, H. 456
- Zenarate 82
- Zittrain, J. 346
- Zuckerberg, Chan 21, 480
- Zuckerberg, Mark 374