

Bernhard Steffen (Ed.)

LNCS 14129

# Bridging the Gap Between AI and Reality

First International Conference, AISoLA 2023  
Crete, Greece, October 23–28, 2023  
Selected Papers

 Springer

OPEN ACCESS

# Lecture Notes in Computer Science

14129

## Founding Editors


Gerhard Goos  
Juris Hartmanis

## Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*

Wen Gao, *Peking University, Beijing, China*

Bernhard Steffen , *TU Dortmund University, Dortmund, Germany*

Moti Yung , *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.


Bernhard Steffen  
Editor

# Bridging the Gap Between AI and Reality

First International Conference, AISoLA 2023  
Crete, Greece, October 23–28, 2023  
Selected Papers



*Editor*

Bernhard Steffen   
TU Dortmund University  
Dortmund, Germany



ISSN 0302-9743

ISSN 1611-3349 (electronic)

Lecture Notes in Computer Science

ISBN 978-3-031-73740-4

ISBN 978-3-031-73741-1 (eBook)

<https://doi.org/10.1007/978-3-031-73741-1>

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

**Open Access** This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This volume contains the postproceedings of AISoLA 2023, an AI-themed sibling of ISoLA, the International Symposium on Leveraging Applications of Formal Methods. AISoLA took place in Crete (Greece) on October 23–28, 2023. It was an in-person event that provided an interdisciplinary forum for discussing the impact of the recent AI developments on research, education, and society. Discussions ranged from philosophical issues that arise from technologies as powerful as indicated by today’s large language models to technical issues and solutions for the responsible uses for AI-applications in safety-critical domains. The program of AISoLA 2023 comprised five keynotes:

Technology and Democracy

by Moshe Vardi, Rice University, USA

Deep Neural Networks, Explanations and Rationality

by Edward Lee, UC Berkeley, USA

Human or Machine: Reflections on Turing-Inspired Testing for the Everyday

by David Harel, Weizmann Institute of Science, Israel

Education and AI – Current Status, Opportunities and Challenges

by Nele McElvany, TU Dortmund, Germany

Graph Neural Networks: Everything Is Connected

Matthias Fey, Kumo.ai, USA

## Special Sessions

Technology and Democracy

organized by Jim Larus and Edward Lee

Beyond ChatGPT: The Impact of AI on Academic Research

organized by Viola Schiaffonati

## Tracks

1. The Nature of AI-Based Systems  
organized by Bernhard Steffen
2. Responsible and Trustworthy AI  
organized by Kevin Baum, Torsten Helfer, Markus Langer, Eva Schmidt, Andreas Sesing-Wagenpfeil, and Timo Speith
3. Democracy in the Digital Era  
organized by George Metakides and Moshe Vardi
4. Digital Humanism  
organized by Viola Schiaffonati and Hannes Werthner
5. Safety Verification of DNNs  
organized by Taylor Johnson and Daniel Neider

6. Verification Meets Learning and Statistics  
organized by Jan Kretinski, Kim Larsen, Nils Jansen, and Bettina Könighofer
7. Health Care  
organized by Martin Leucker
8. AI-Assisted Programming  
organized by Wolfgang Ahrendt and Klaus Havelund
9. Safe AI in the Automotive Domain  
organized by Falk Howar and Hardi Hungar
10. Digital Humanities  
organized by Ciara Breathnach and Tiziana Margaria
11. R@ISE: Research at ISE  
organized by Tiziana Margaria and Mike Hinchey
12. AI-Supported Publishing  
organized by Jonas Spies

The 26 papers of this volume extend the presentation in the AISoLA 2023 on-site proceedings. This means, in particular, that the corresponding track introductions still apply. Only the Health Care and the Digital Humanities track have some dedicated introductions in this volume.

I thank the track organizers, the members of the program committee and their reviewers for their effort in selecting the papers to be presented, the local Organization Chair, Petros Stratis, and the EasyConferences team for their continuous precious support during the entire period preceding the events, and Springer Nature for being, as usual, a very reliable partner for the proceedings production. Finally, I am grateful to Nicolas Stratis and Tim Tegeler for continuous support for the website and the program, and to Steve Bosselmann for his help with the editorial system EquinOCS.

Special thanks are due to the Center for Trustworthy Data Science and Security and the Lamarr Institute for their support in the organization of the event, and to the Technical University of Dortmund, my home institution.

With over 150 international participants and very lively discussions, AISoLA was a very successful event and I am looking forward to seeing many of you in Crete in October for AISoLA 2024.

October 2024

Bernhard Steffen

# Organization

## Program Committee

Ahrendt, Wolfgang	Chalmers University of Technology, Sweden
Baum, Kevin	Saarland University, Germany
Breathnach, Ciara	University of Limerick, Ireland
Havelund, Klaus	Jet Propulsion Laboratory, USA
Helfer, Thorsten	Saarland University, Germany
Howar, Falk	TU Dortmund, Germany
Hungar, Hardi	DLR Braunschweig, Germany
Jansen, Nils	Radboud Universit�at Nijmegen, Germany
Johnson, Taylor T.	Vanderbilt University, USA
Kretinsky, Jan	Brno University, Czech Republic
K�onighofer, Bettina	Graz University of Technology, Austria
Langer, Markus	University of Marburg, Germany
Larsen, Kim	Aalborg University, Denmark
Leucker, Martin	University of L�ubeck, Germany
Margaria, Tiziana	University of Limerick, and Lero, Ireland
Neider, Daniel	TU Dortmund, Germany
O'Shea, Enda	University of Limerick, Ireland
Schmidt, Eva	TU Dortmund, Germany
Sesing-Wagenpfeil, Andreas	Saarland University, Germany
Speith, Timo	University of Bayreuth, Germany
Steffen, Bernhard (Chair)	TU Dortmund, Germany

## Reviewers

Ahrendt, Wolfgang	Chalmers University of Technology, Sweden
Aichernig, Bernhard	TU Graz, Austria
Bao, Tianshu	Guizhou University, China
Baum, Kevin	Saarland University, Germany
Baumann, Jonas	Saarland University, Germany
Bollig, Benedikt	ENS Cachan, France
Bongo, Lars Ailo	UiT, Arctic University of Norway, Troms�o, Norway
Breathnach, Ciara	University of Limerick, Ireland
Chaudhary, Hafiz Ahmad Awais	University of Limerick, Ireland

Ferilli, Stefano	University of Bari, Italy
Floris, Francesco	Università di Torino, Italy
Fradiante, Valeria	Università di Torino, Italy
Havelund, Klaus	Jet Propulsion Laboratory, USA
Helfer, Thorsten	CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
Howar, Falk	TU Dortmund, Germany
Hungar, Hardi	DLR Braunschweig, Germany
Hösterey, Steffen	Humboldt-Universität zu Berlin, Germany
Langer, Markus	University of Freiburg, Germany
Le Roux, Maelle	University of Limerick, Ireland
Leucker, Martin	University of Lübeck, Germany
Margaria, Tiziana	University of Limerick, CSIS and Lero, Ireland
Mitwalli, Daniel Sami	University of Limerick, Ireland
Mourits, Rick	International Institute of Social History, Amsterdam, the Netherlands
Pedersen, Bjørn-Richard	UiT, Arctic University of Norway, Tromsø, Norway
Peled, Doron	Bar Ilan University, Israel
Sachenbacher, Martin	University of Lübeck, Germany
Schmidt, Eva	TU Dortmund, Germany
Schrills, Tim	University of Lübeck, Germany
Sesing-Wagenpfeil, Andreas	Saarland University, Germany
Singh, Amandeep	University of Limerick, Ireland
Speith, Timo	University of Bayreuth, Germany
Teumert, Sebastian	TU Dortmund, Germany
Thoma, Daniel	University of Lübeck, Germany
Walsh, Oonagh	Glasgow Caledonian University, UK
Xiang, Weiming	Augusta University, USA
Yalciner, Mustafa	Fraunhofer IML, Germany
Zafeiridi, Evi	University College Cork, Ireland

# Contents

## Digital Humanities

Digital Humanities and Cultural Heritage in AI and IT-Enabled Environments .....	3
<i>Ciara Breathnach and Tiziana Margaria</i>	
Common Language for Accessibility, Interoperability, and Reusability in Historical Demography .....	10
<i>Rick J. Mourits, Tim Riswick, and Rombert Stapel</i>	
Coding Historical Causes of Death Data with Large Language Models .....	30
<i>Bjørn-Richard Pedersen, Maisha Islam, Doris Tove Kristoffersen, Lars Ailo Bongo, Eilidh Garrett, Alice Reid, and Hilde Sommersth</i>	
Teaching the Specialized Language of Mathematics with a Data-Driven Approach: What Data Do We Use? .....	48
<i>Cecilia Fissore, Francesco Floris, Marina Marchisio Conte, and Matteo Sacchet</i>	
Interoperating Civil Registration of Death and Census Data: Old Age and Marriage as Categories of Analysis .....	65
<i>Ciara Breathnach, Rachel Murphy, Alexander Schieweck, and Tiziana Margaria</i>	
From Data Science to Modular Workflows Changing Perspectives from Data to Platform: DBDIrl 1864-1922 Case Study .....	84
<i>Enda O'Shea, Marco Krumrey, Daniel Sami Mitwalli, Sebastian Teumert, and Tiziana Margaria</i>	
Mapping Madness: HGIS and the Analysis of Irish Patient Records .....	104
<i>Oonagh Walsh and Stuart Clancy</i>	
Digitised Historical Sources and Non-digital Humanists: An Interdisciplinary Challenge? .....	119
<i>Maelle Le Roux and Anna Gasperini</i>	
Using Passive Sensing to Identify Depression .....	132
<i>Evi Zafeiridi, Malik Muhammad Qirtas, Eleanor Bantry White, and Dirk Pesch</i>	

The GraphBRAIN Framework for Knowledge Graph Management and Its Applications to Cultural Heritage ..... 144  
*Stefano Ferilli, Eleonora Bernasconi, Davide Di Pierro, and Domenico Redavid*

**Health Care**

Challenges for AI in Healthcare Systems ..... 165  
*Markus Bertl, Yngve Lamo, Martin Leucker, Tiziana Margaria, Esfandiar Mohammadi, Suresh Kumar Mukhiya, Ludwig Pechmann, Gunnar Piho, and Fazle Rabbi*

Towards a Multi-dimensional Health Data Analysis Framework ..... 187  
*Fazle Rabbi, Bahareh Fatemi, Suresh Kumar Mukhiya, and Yngve Lamo*

Future Opportunities for Systematic AI Support in Healthcare ..... 203  
*Markus Bertl, Gunnar Piho, Dirk Draheim, Peeter Ross, Ludwig Pechmann, Nicholas Bucciarelli, and Rahul Sharma*

CRISP-PCCP – A Development Methodology Supporting FDA Approval for Machine Learning Enabled Medical Devices ..... 225  
*Ludwig Pechmann, Yannik Potdevin, Kai Brehmer, Dirk Nowotka, and Martin Leucker*

Model Driven Development for AI-Based Healthcare Systems: A Review ..... 245  
*Colm Brandon, Amandeep Singh, and Tiziana Margaria*

**Responsible and Trustworthy AI**

Balancing Transparency and Risk: An Overview of the Security and Privacy Risks of Open-Source Machine Learning Models ..... 269  
*Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting*

AI-Related Risk and Uncertainty ..... 284  
*Giacomo Zanotti, Daniele Chiffi, and Viola Schiaffonati*

Leveraging Actionable Explanations to Improve People’s Reactions to AI-Based Decisions ..... 293  
*Markus Langer and Isabel Valera*

From Explanation Correctness to Explanation Goodness: Only Provably Correct Explanations Can Save the World ..... 307  
*Maike Schwammburger*

Thinking Outside the Box?: Regulatory Sandboxes as a Tool for AI Regulation ..... 318  
*Hannah Ruschemeier*

AI and Democratic Equality: How Surveillance Capitalism and Computational Propaganda Threaten Democracy ..... 333  
*Ashton Black*

**AI Assisted Programming**

Safeguarding AI-Based Software Development and Verification Using Witnesses (Position Paper) ..... 351  
*Dirk Beyer*

End-to-End AI Generated Runtime Verification from Natural Language Specification ..... 362  
*Itay Cohen and Doron Peled*

AI-Assisted Programming with Test-Based Refinement ..... 385  
*Bernhard K. Aichernig and Klaus Havelund*

**Automotive Driving**

Safer Than Perception: Increasing Resilience of Automated Vehicles Against Misperception ..... 415  
*Martin Fränzle and Andreas Hein*

Towards ML-Integration and Training Patterns for AI-Enabled Systems ..... 434  
*Sven Peldszus, Henriette Knopp, Yorick Sens, and Thorsten Berger*

**Safety Verification of DNNs**

The Reachability Problem for Neural-Network Control Systems ..... 455  
*Christian Schilling and Martin Zimmermann*

**Author Index** ..... 471



# **Digital Humanities**



# Digital Humanities and Cultural Heritage in AI and IT-Enabled Environments

Ciara Breathnach<sup>1</sup> and Tiziana Margaria<sup>2</sup>(✉)

<sup>1</sup> School of History, University College Cork, IE, Cork, Ireland

[Ciara.breathnach@ucc.ie](mailto:Ciara.breathnach@ucc.ie)

<sup>2</sup> HRI, Lero and CRT-AI, University of Limerick, Limerick, Ireland

[Tiziana.Margaria@ul.ie](mailto:Tiziana.Margaria@ul.ie)

**Abstract.** This track is the first output in Digital Humanities within AISoLA and, while its contributions span a range of diverse topics and approaches, it provides a good representation of the state of the art in the field, stemming from the interdisciplinary collaborations in the DBDIrl project and from the Great Leap COST Action that started in September 2023. It also underpins an ambitious research agenda arising from these collaborations, which aims to foster further international work on data interoperability.

The papers discuss the challenges faced by both computing and historical sciences when addressing on one side some of the most pressing issues of data access, preservation, conservation, harmonisation across national datasets, and governance, and on the other side the opportunities and threats brought by AI and machine learning to the advancement of reasoning, classification and rigorous data analytics.

## 1 The Topic: DH Meets AI and IT

We are in the middle of an AI and IT revolution and at a point of digital cultural heritage data saturation, but humanities' scholarship is struggling to keep pace. In this Track we discuss the challenges faced by both computing and historical sciences to outline a roadmap to address some of the most pressing issues of data access, preservation, conservation, harmonisation across national datasets, and governance on one side, and the opportunities and threats brought by AI and machine learning to the advancement of rigorous data analytics. We welcomed contributions that address the following and other related topics:

- Advances brought by modern software development, AI, ML and data analytics to the transcription of documents and sources (Pedersen et al. [12], Mourits and Riswick [10] and O'Shea et al. [11]).
- Tools and platforms that address the digital divide between physical, analog or digital sources and the level of curation of datasets needed for modern analytics (Le Roux and Gasperini [8], Zafeiridi et al. [16] and Breathnach et al. [3]).

- Design for accessibility and interoperability of data sets, including corpora and thesauri (Le Roux and Gasperini [8], Mourits et al. [10] and Breathnach et al. [3]).
- Tools and techniques for machine-understanding form-based documents, recognition of digits and codes, handwriting, and other semantically structured data (Walsh and Clancy [14], Pedersen et al. [12], Mourits et al. [10] and O’Shea et al. [11]).
- Knowledge representation for better analysis of semi-structured data from relevant domains (diaries, registers, reports, etc.) (Pedersen et al. [12], Mourits et al. [10], O’Shea et al. [11], and Breathnach et al. [3]).
- Specific needs arising from the study of minority languages and populations, disadvantaged groups and any other rare or less documented phenomena and groups (Fissore et al. [5]).
- Challenges relative to the conservation, publication, curation, and governance of data as open access artefacts (Ferilli et al. [4]).
- Challenges relative to initial and continuing education and curricular or extracurricular professional formation in the digital humanities professions (Le Roux and Gasperini [8] and Fissore et al. [5]).
- Spatial digital humanities (Walsh and Clancy [14]).
- Digital humanities aspects concerning occupation, medicine and health (Pedersen et al. [12] and Breathnach et al. [3]).

## 2 The Background: Digital Meets Humanities

Owing primarily to over forty years of born digital content and also because of the digitisation of handwritten and printed documents, we are currently in an age of ‘digital abundance’ [9]. Digitisation can take many forms, if we take a broad view and include photography as a means of preservation (which developed into microforms including microfilm and microfiche) then it has been a critical component of records management in the finance and banking sector since the 1920s [13]. In more recent times, it has been the panacea for conservation and preservation in cultural heritage since the 1940s [2]. While such activity has resolved problems associated with storing newspapers and large datasets like census returns and has taken fragile manuscripts out of the handling environment, it has created several legacy issues associated with obsolescence and discoverability. Indeed Milligan [9] has elucidated the problems associated with the digitisation of newspapers from microfilm, where old problems of shading, depth of field and blurring, have simply been replicated in large scale digital archive projects. It seems that the end-user, or the researcher, has received little priority in the quest for ‘digital abundance’. Greater accessibility has created new vulnerabilities that range from the whims of corporate decision-making to cyberattacks. For example, in the case of the former, executives at the MTV News Website wiped all archival content from the web in June 2024 [7], which has dealt a devastating blow to music history as there are no hard copies of this born-digital content and surrogate copies are limited to a partial capture on the

internet archive on archive.org. Cyberattack has wreaked havoc at the British Library since October 2023, and has negated much of its enormous efforts in the digitisation of manuscripts and other data types over the last century. Clearly there are many old and new problems that scholars must address and here we showcase a range of projects that exhibit the great potential when researchers collaborate in interdisciplinary ways.

### 3 The Contributions in Context

In this Track, we welcomed contributions from interdisciplinarians who work in the field of Digital Humanities, broadly defined. It includes contributions that discuss how scholars can reuse legacy data, the tools and methods they employ to conduct that work, and the role of AI and Machine Learning can have in future research. Further contributions include the potentials of wearable tech and Large Language Models (LLMs) in mental health and in gamification in mathematics pedagogy. Ostensibly what emerged from the conference is a volume about data, its generation in various forms and how we, as interdisciplinarians, exploit its use.

This volume is arranged chronologically in accordance with the evolution of digitization processes and the advancement of

P1: It begins with **Le Roux and Gasperini's** paper [8] discussing the problems that past decision making in digitisation projects create for current research. Taking medical literature pertaining to child health published between 1850 and 1914 in England, France and Italy as their sample they explore the problems that Optical Character Recognition (OCR) generates for digital humanists and how computer science methods might help to resolve them. Adopting a corpus linguistics approach they argue that discoverability remains a problem and that the retrospective application of Free Accessible Interoperable and Reusable (FAIR) [15] principles will require closer collaboration between stakeholders as AI and ML technologies take further root.

Historical Demography features heavily in this track as historical 'Big Data' provide a useful comparative case study in old and new research methods. It showcases some of the work of collaborators in the COST Action CA22116 - The Great Leap. Multidisciplinary approaches to health inequalities, 1800–2022 (GREATLEAP) [1]. Tim Riswick is Chair and Breathnach is Vice Chair of this international network. Locating its work in historical population data The Great Leap examines structural inequalities in health and how they emerged over time, primarily across Europe. It is an interdisciplinary network of scholars and it counts contributors Clancy, Ferilli, Garret, Margaria, Mourits, Reid, Sommerseth and Walsh among its wider membership. Further to experts in history, social sciences, life sciences, computer sciences and epidemiology, it also involves government agencies and data owners such as statistical offices and national archives. Like the contributions to this volume, the Great Leap focuses on individual-level metadata but it has the ambitious aim of creating

interoperability between national datasets. The ultimate aim of the network is to push the capabilities of historical data to have impact in current public health policy and practice.

The Great Leap associated papers in this volume capture a timeline in the development of transcription and machine learning methods over the past 40 years. The earliest contributions to the field were constructed in North America and across Europe (for example, in the Netherlands) but with old technology and its associated methods, comes new challenges.

P2: **Mourits, Riskiwck, and Stapel** [10] deal with the matter of interoperability and common languages in historical demography, they note how modern ontologies are insufficient and cannot be neatly applied to the past. They have gathered data from several international projects to assess its FAIR compliance and found that occupational data was the most standardised category. They point to the success of past research projects like HISCO [6] and its work on standardisation and the encoding of occupations, and present some of their findings here.

P3: Building on Mourits et al., **Pedersen, Islam, Kristofferson, Bongo, Garrett, Reid and Somerseth** [12] discuss the problems surrounding data encoding and how useful LLMs might be in the retrospective application of ICD10 to historical cause of death. Experiments using three AI-models on a random data sample of death registrations from three areas in the UK (that were fully transcribed and encoded by domain experts) showed a varying success rate of correctness. While the authors found the exercise worthwhile they caution that significant levels of fine-tuning (perhaps using Retrieval-Augmented Generation) would be necessary to correct the error rate.

P4: Although they address a different problem, **Fissore, Floris, Marchiso Conte and Sacchet** [5] examine the transformative potential of gamification, LLMs and AI to shift away from didactic approaches to teaching mathematics and create more opportunities for co-production in education. Identifying the specialised language of mathematics as a major obstacle to learning, they recommend a Data-Driven Learning (DDL) methodology to help teachers and students identify and work through difficulties in a scaffolded digital learning environment. They note that while further in-service training is necessary gamification offers massive opportunities to individualise learning while also supporting the educators.

P5: **Breathnach, Murphy, Schieweck and Margaria** [3] describe the difficulties of interoperating various Irish historical datasets uncovered during Death and Burial Data: Ireland 1864–1922 (DBDIrl) a project funded by the Irish Research Council (2018–2023). Taking old age as a case study the authors who are drawn from history and computer science describe the pipelines necessary to automate the process of data linkage using a low-code no-code approach. While age heaping (rounding to nearest 0 or 5) is a perennial problem in census data and disrupts efforts to create linkages between decennial data, marital status was also found to be inconsistently recorded.

P6: In another paper arising from DBDIrl activities, **O’Shea, Krumrey, Mitwalli, Teumert and Margaria’s** work [11] follows the technical prob-

lems that limitations in OCR technology poses and the necessity for precise responses tailored to the problems of each data type. It discusses the benefits and limitations of an AI-ML Data Analytics Pipeline designed as an automated solution to transcription of historical handwritten death registers. In many respects this work shows how countries without longstanding traditions in crowdsourced transcription data (like those described by Mourits and Riswick [10] and Pedersen et al. [12]) can play catch-up through pipelines of segmentation, word detection, and data synthesis, classification and linkage. P7: Spatial epidemiology underpins **Walsh and Clancy**'s paper [14] on the Irish District Lunatic Asylum system. They illustrate the potential that the mapping of individual level data can bring to our understandings of mental health problems in both urban and rural Connaught (a large province in Ireland). The removal of patients from their households of origin was a key component of the 'confinement' model and in this chapter Walsh and Clancy interoperate committal data with civil registration records to plot the lives and movement of patients.

P8: Albeit in a modern context, **Zafeiridi, Qirtas, Bantry White and Pesch** [16] explore how machine learning models and wearable tech can act in a preventive capacity in detecting depression. They found that GPS should not be used in isolation and that passive sensing, which includes activity, sleep data and personal communication (texts, calls, Apps and proximity to other smart devices), offered a more holistic approach to the study of mental health and well being.

P9: With a focus on discoverability **Ferilli, Bernasconi, Di Piero and Redavid** [4] contributes the final paper in this volume. It goes beyond the analytical limitations of the digital archive and shows how the GraphBRAIN framework for knowledge graphs operates in the domain of cultural heritage. In this project they show the realms of possibility for conservation, preservation and, more importantly, discoverability and interoperability.

## 4 The Next Steps

This track is the first output in Digital Humanities for AISoLA and, while its contributions span a range of diverse topics and approaches, it provides a good representation of the state of the art in the field. Several exciting prospects emerge from the use cases presented here, for example the capacity of LLMs and GIS to reconfigure legacy data and offer new insights from old 'Big Data', and the use of tools like gamification drawing on corpus linguistics and DDL to transform mathematics pedagogy. Essentially this volume showcases the importance of collaborative and interdisciplinary approaches to complex research questions, and the enormous potential for future innovative scholarship. We expect further collaborations to arise from the Great Leap context, which started in September 2023 and will provide opportunities of collaboration with its network for the next three years.

## References

1. The Great Leap. Multidisciplinary approaches to health inequalities, 1800 - 2022. EU COST Action CA22116. <https://greatleap.eu/>. Accessed 28 Jul 2024
2. Archives, U.S.N., Service, R.: National Archives and Records Service Microfilm Publications. General information leaflet, U.S. General Services Administration, National Archives and Records Service (1975). <https://books.google.it/books?id=W6sXiHcrHdsC>
3. Breathnach, C., Murphy, R., Schieweck, A., Margaria, T.: Interoperating civil registration of death and census data: old age and marriage as categories of analysis. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
4. Ferilli, S., Bernasconi, E., Di Piero, D., Redavid, D.: The GraphBRAIN framework for knowledge graph management and its applications to cultural heritage. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
5. Fissore, C., Floris, F., Marchisio Conte, M., Sacchet, M.: Teaching the specialized language of Mathematics with a data-driven approach: what data do we use? In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
6. International Institute of Social History: History of work - HISCO. <https://iisg.amsterdam/en/data/data-websites/history-of-work>. Accessed 28 Jul 2024
7. Janine, D.: MTV Deleted! The iconic music tv channel of the 80's & 90's is Gone. <https://stagelync.com/news/mtv-deleted>. Accessed 28 Jul 2024
8. Le Roux, M., Gasperini, A.: Digitised historical sources and non-digital humanists: an interdisciplinary challenge? In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
9. Milligan, I.: The Transformation of Historical Research in the Digital Age. Elements in Historical Theory and Practice, Cambridge University Press (2022)
10. Mourits, R.J., Riswick, T., Stapel, R.: Common language for accessibility, interoperability, and reusability in historical demography. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
11. O'Shea, E., Krumrey, M., Mitwalli, D.S., Teumert, S., Margaria, T.: From data science to modular workflows - changing perspectives from data to platform: DBDIrl 1864-1922 case study. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
12. Pedersen, B.R., et al.: Coding historical causes of death data with large language models. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
13. Teague, S.J.: Microform, Video and Electronic Media Librarianship. K.G. Saur Publishing Inc., USA (1985)

14. Walsh, O., Clancy, S.: Mapping Madness: HGIS and the granular analysis of Irish patient records. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)
15. Wilkinson, M.D., Dumontier, M., et al.: The fair guiding principles for scientific data management and stewardship. *Sci. Data* **3**(1), 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
16. Zafeiridi, E., Qirtas, M.M., Bantry White, E., Pesch, D.: Using passive sensing to identify depression. In: Proceedings of the AISoLA 2023, Special Track, Digital Humanities and Cultural Heritage in AI and IT-enabled Environments, p. (this volume). Springer Nature (2024)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.





The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.







# Common Language for Accessibility, Interoperability, and Reusability in Historical Demography

Rick J. Mourits<sup>1</sup>, Tim Riswick<sup>2</sup>, and Rombert Stapel<sup>1</sup>

<sup>1</sup> International Institute of Social History, Amsterdam, Netherlands  
{rick.mourits,rombert.stapel}@iisg.nl

<sup>2</sup> Radboud University, Nijmegen, Netherlands  
tim.riswick@ru.nl

**Abstract.** One of the biggest challenges in the transition to open science is making data interoperable. Ideally, existing schemas and vocabularies are (re-)used to describe data, but these are generally problematic for historical data, as they exclude historical concepts and are insensitive to temporal variations in meaning. Therefore, the subdiscipline of historical demography has designed its own schemas and vocabularies to standardize historical data, as researchers require them to make and study large-scale reconstructions of populations and life courses. We introduce a web environment called CLAIR-HD that helps researchers to find vocabularies to standardize historical demographic data, and determine lacunae in the standardization of data within the field of historical demography.

**Keywords:** Interoperability · Vocabularies · Historical Demography

## 1 CLAIR-HD

One of the biggest hurdles in data interoperability is communication. Without coordination, database managers tend to come up with different descriptions for the same information, which hurts data interoperability. To tackle this problem, vocabularies and schemas are used to standardize how data in datasets is being described. Sometimes these standardization efforts are very straightforward and apply to very broad contexts, whereas others are of general use to specific communities. For historical data, however, most of these standardization efforts are problematic as they were made to describe contemporary concepts and underestimate how information and meaning can change over time. For example, places and their names change over time, the associations between occupation and social standing shift when labour markets change, and causes of death are coded differently between social and temporal contexts [17–19, 23, 24, 34, 35, 47]. Existing vocabularies standardize these historical concepts at the cost of losing or misinterpreting context-specific information.

In this paper, we use the subdiscipline of historical demography as a case study to see how specialized vocabularies are made and adopted. Historical

demographers from a wide array of countries have built databases to reconstruct the lives of people in Europe, North America, and East Asia [22]. The vocabularies of these databases were designed to “stay true to the source”, so that datasets have sophisticated designs to model local peculiarities and changes in meaning over time. These local efforts have made it possible to standardize defunct phenomena, historical distinctions, and general changes over time, though only within the geographic scope of their projects. Each of these standardization schemes is worth its weight in gold, as it unlocks a wealth of historical data and contains years of insight in the historical sources and context. Yet, standardized communication is necessary to make larger-scale comparisons possible. Currently, the field is in a paradoxical situation where most scholars agree that historical comparisons increase understanding of the “historical context”, but are at the same time afraid to throw out the baby with the bathing water and hesitant to apply schemes not designed for a specific social, spatial, and temporal context [24, 33, 54].

To have fruitful discussions on vocabularies, historical demographers need to have an overview of existing standardization efforts. Yet, this requires too much effort for most scholars, as they need to know the field well, have expertise in presenting data, and invest time in ontology design. To reduce the time and knowledge required to partake in this discussion, we gathered the vocabularies that data centers, projects, and research collaborations developed to make data interoperable and variables comparable. To show how these vocabularies are related, we mapped the relationships between them and created an overview of vocabulary conversion tools. These results are published on the CLAIR-HD webpage, so that researchers and database managers can easily find and reuse existing vocabularies. Gathering and sharing these vocabularies helps historical demographers to learn from each others insights, prevents the re-invention of vocabularies, and ensures that data is interoperable. But most importantly, it serves as a case study on what is required for a move towards open data within history, and perhaps even the humanities and social sciences as a whole, as common vocabularies allow for general-purpose software, make replication studies easier, and are the steppingstone to Linked Open Data. We intend for CLAIR-HD to become an inspiration to other disciplines that face similar challenges.

## 2 Methods

Information on standardization efforts within historical demography was gathered in multiple rounds. Our initial goal was to get a broad outline of the existing vocabularies. Therefore, we contacted the bigger data centers in Asia, Europe, and North America. These data centers were a logical place to start, as they have the most developed infrastructure and are important regional hubs in historical demography. Although all historical data centers responded promptly, it turned out that few disclosed their own standardization lists and often reused existing vocabularies to standardize their data. Nevertheless, there were notable exceptions and paradoxically, the data centers often worked together to develop

shared code books, also known as schemas. Hence, we decided to also gather information on the schemas that they developed. This gave us a feeling for how data standardization efforts in the field were designed and why there has been less focus on developing shared vocabularies than initially anticipated.

We also talked to individual scholars at the major historical demography conferences in Europe and North America: European Social Science History Conference (ESSHC), European Society for Historical Demography (ESHD), and the Social Science History Association conference (SSHA). Here we learned which vocabularies historical demographers use to code their data and how they work together to interpret and codify historical data. In general, researchers seem to prefer using multiple vocabularies in their analyses to test whether differences in interpretation and standardization practices can lead to different statistical associations, either to look for overlap as a robustness check or to tease out differences as an enquiry into underlying mechanisms. Therefore, we decided to also collect information on crosswalks and conversion tools, as they hugely increase data interoperability and give researchers an important tool in their toolbox.

### 3 Existing Standardization Efforts

The outcomes of the enquiry are interpreted in this section of the paper and presented on the CLAIR-HD web page and Sect. 4. Our initial goal was to get a broad outline of the existing vocabularies and show the overlap between them. Once the data came in, it became clear that vocabularies are generally products of scholarly collaborations. Whether these vocabularies are used is dependent on the quality of the product itself, other scholars familiarity with it, and willingness of data providers to implement it. To understand the institutional context within which vocabularies are being provided, we first describe the seven schemas that were developed within and for historical demography. Second, we give an overview of the existing vocabularies and determine lacunae in the standardization of data within the field of historical demography. Finally, we close by discussing crosswalks and conversion tools.

#### 3.1 Schemas

Our inventory of the field showed that each data center uses its own schema. However, there are also seven schemas that provide standardized ways to deliver datasets, 1. the Intermedidate Data Structure (IDS) [1, 2], 2. IPUMS-USA [40], 3. IPUMS-International [25], 4. LINKS-gen [29], 5. MOSAIC [46], 6. North Atlantic Population Project (NAPP) [36–38], and 7. Persons in Context (PiCo) [5].

Although all these schemas are meant to standardize historical data, their intended scope differs. IPUMS-USA, IPUMS-International, MOSAIC, and NAPP were developed to standardize census data. The driving force behind processing census data is IPUMS at the University of Minnesota. In 1991, they

**Table 1.** Census schemas (IPUMS-international, MOSAIC, NAPP) [46]

Theme	Variable	IPUMS	MOSAIC	NAPP
Geography	Country code	CNTRY	country	CNTRY
	Place	–	place	–
	Region	region	place	–
	Urban-rural	URBAN	urban	URBAN
Household	Group quarter status	GQ	gq	GQ
	Household size	PERSONS	hhsz	NUMBERHH
	Household weight	WTHH	hhwt	HHWT
Identifier	Enumeration	SAMPLE	id_enum	SAMPLE
	Household	SERIAL	id_hhold	SERIAL
	Person	PERNUM	id_pers	PERNUM
Individual	Age	AGE	age	AGE
	Literacy	LIT	lit	LIT
	Marital status	MARST	marst	MARST
	Occupational title	–	occupan	OCCSTR
	OCCHISCO	OCCHISCO	occhisco	OCCHISCO
	Present at enum.	RESIDENT	presence	RESIDENT
	Rel. household head	–	relate	–
	Religion	RELIG	relig	RELIGION
	Sex	SEX	sex	SEX
	Weight	WTPER	perwt	PERWT
Person name	First name	–	fname	NAMEFRST
	Last name	–	lname	NAMELAST
Quality	Age	–	qage	QAGEGB
	Household	–	qhold	–
	Rel. household head	–	qrelate	QRELGB
	Marital status	–	qmarst	QMARSTGB
	SEX	–	qsex	QSEXGB
Source	Enumeration type	–	enumtype	–
	Enumeration year	YEAR	year	YEAR

started providing “common-format extracts” with standardized codes and constructed variables for the 1960, 1970, and 1980 US censuses and now maintain standards to exchange census data within the USA (IPUMS-USA) and internationally (IPUMS-International) [39]. In 1999, IPUMS joined up with researchers from Canada, Great Britain, Iceland, Norway, and Scotland with whom they already had established strong ties. As they realised that the original source material was highly compatible and cultural constructs for the measured concepts are similar, they decided to create a machine-readable, census-

based database of, as they put it, the North Atlantic world at the end of the 19th century [36,38]. A similar international census comparison project took place in the early 2010s at the Max Planck Institute for Demographic Research (MPIDR) in Rostock, Germany. The MPIDR schema, called MOSAIC, standardized census data from the 1700s until 1950 for 18 regions in Europe [46]. These schemas to describe census data are very similar as shown in Table 1. Yet, all four schemas are still in use, as they cater to a very specific public and have not been adopted by other historical demographic research projects or databases, as they have been developed for projects with strong institutional boundaries.

Whereas IPUMS-USA, IPUMS-International, MOSAIC, and NAPP focus on standardizing categories between series of cross-sectional census data, other schemas set up standards for sharing person reconstructions. IDS, LINKS-gen, and PiCo are efforts to standardize historical data from other types of historical sources, such as the civil registry, militia registers, parish registers, population registers, slave registers, or tax registers. Of these schemas, LINKS-gen is by far the most limited in its scope and standardizes historical data into a pedigree format - with each row representing a person that is linked to his family by links to one's father, mother, and spouses - and a standardized occupational table that are both ready for statistical analysis [29]. IDS has been around since 2009 and makes different types of data sources available for extraction by explicitly stating for which point or period in time historical information is valid [1,2]. PiCo goes a step further and is developed by the Center for Family History in the Netherlands as a means to store information on persons registrations as well as concomitant records and person reconstructions [5]. Table 2 shows how different these schemas are from each other, as well as from the IPUMS, MOSAIC, and NAPP census schemas. As LINKS-gen and PiCo are relatively new, it is still uncertain to what extent they will be implemented by other historical demographic research projects and databases, which will ultimately determine their longevity.

All schemas within historical demography deal with many similar concepts, but have very limited interoperability. This resemblance is understandable as most schemas are designed for tabular datasets with historical person data, hence one would expect that most schemas contain a standardized variable name and categories for common concepts that describe historical persons and the relations between them. Yet, historical data centers have been deeply rooted in national research traditions and are more focused on disclosing historical sources, matching data, and reconstitution families than on exchanging data [8,9,21,22,42]. As a result, variable names and categories are generally standardized within one institutional context rather than by using a shared vocabulary within the field of historical demography. This is even true for concepts that can be standardized very easily, such as birth or marriages dates, which have no common name or set date format to order day, month, and year. This general lack of a common language for historical demography means that schemas only function within their institutional context, are hard to find for people not actively looking for them, and require a plethora of conversion tools to move between them.

**Table 2.** Date, geography, household, identifiers, individual, and source information in the IDS, LINKS-gen, and PiCo schemas

Variable	IDS	LINKS-gen	PiCo
Baptism	BAPTISM_DATE	–	–
Birth	BIRTH_DATE	B_date	schema:birthDate
Date	TIMESTAMP	Date	–
Death	DEATH_DATE	D_date	schema:deathDate
Divorce	DIVORCE_DATE	–	–
First obs	START_OBSERVATION	–	–
Funeral	FUNERAL_DATE	–	–
Last obs	START_OBSERVATION	LastEntryDate	–
Marriage	MARRIAGE_DATE	M_date_ <i>n</i>	–
Mar. banms	MARRIAGE_PROCLAMATION_DATE	–	–
Stillbirth	STILLBIRTH_DATE	–	–
Baptism	BAPTISM_LOCATION	–	–
Birth	BIRTH_LOCATION	B_location	schema:birthPlace
Death	DEATH_LOCATION	D_location	schema:deathPlace
Divorce	DIVORCE_LOCATION	–	–
Funeral	FUNERAL_DATE	–	–
Marriage	MARRIAGE_LOCATION	M_location_ <i>n</i>	–
Mar. banms	MARRIAGE_PROCLAMATION_LOCATION	–	–
Place	–	Location	schema:address
Stillbirth	STILLBIRTH_LOCATION	–	–
ID father	ID_I.2	Id_father	schema:parent
ID household	ID_C	–	–
ID mother	ID_I.2	Id_mother	schema:spouse
ID partner	ID_I.2	Id_partner_ <i>n</i>	schema:spouse
ID person	ID_I.1	Id_person	pico:personObservation
Age	AGE_YEARS AGE_MONTHS AGE_WEEKS AGE_DAYS	Age	pico:hasAge
Age at death	–	D_age	–
Age last obs.	–	LastEntryAge	–
Alive/dead	ALIVE	–	pico:deceased
HISCO	OCCUPATION_HISCO	HISCO	–
Mar. status	CIVIL_STATUS	–	–
Marriages	–	Marriages_N	–
Nationality	NATIONALITY	–	–
Occ. title	OCCUPATION	Occupation	schema:hasOccupation
Religion	RELIGION	–	pico:hasReligion
Role	–	–	pico:hasRole
Sex	SEX	Sex	schema:gender
Twin	MULTIPLE_BIRTH	Twin	–
Archive	–	–	schema:holdingArchive
Image	–	–	schema:url
Source	–	–	prov:hadPrimarySource

Taking a closer look at PiCo shows that there is a way to solve the problems with findability, accessibility, and interoperability by reusing existing schemas from domains outside historical demography. In line with insights from earlier projects [24], the designers behind PiCo describe concepts in their schema with

vocabularies from existing schemas to make their data more FAIR, as shown in Table 3. They do so by describing their data in different layers. A first layer is very general and uses concepts that are used around the world, such as gender or parenthood. These concepts are derived from schema.org [15]. Domain-specific concepts, such as biological events or data provenance, are borrowed from domain-specific schemas like BIO and PROV-O [6,58]. Finally, concepts that are specific to historical demography or the civil registry are defined in specialized vocabularies, in this case their own PiCo vocabulary. As a result, PiCo contains the best practices from existing schemas and reuses them to be more easily findable, accessible on the internet, and interoperable with other databases.

However, simply reusing vocabularies is not enough. The authors of PiCo run into the same problem as the designers of other schemas when they start to describe concepts unique to the field of historical demography. Following their own design logic, they ought to implement vocabularies from accepted schemas within historical demography. But a widely accepted schema with concomitant vocabularies is simply not around. PiCo tried to solve this issue by defining its own definitions, putting it at risk of also becoming bound to a specific institutional context. Nevertheless, PiCo presents part of the solution to making historical data FAIR and less constraint by institutional boundaries. The other half of the solution is provided by collaborating researchers whose grassroots initiatives have resulted in specialized vocabularies on concepts such as cause of death categories and social status (Sect. 3.2).

**Table 3.** Concentric description of a marriage certificate using PiCo

Schema level	Schema name	Concept name
General	XSD [11]	date, int, string
	Schema.org [15]	spouse, parent, gender, familyName, givenName, ArchiveComponent, dateCreated, locationCreated
Domain-specific	BIO [6]	Marriage, date, partner
	PROV-O [58]	hadPrimarySource, wasDerivedfrom
Specialized	PiCo [5]	PersonObservation, hasRole, huwelijkspartij, huwelijksakte, hasAge

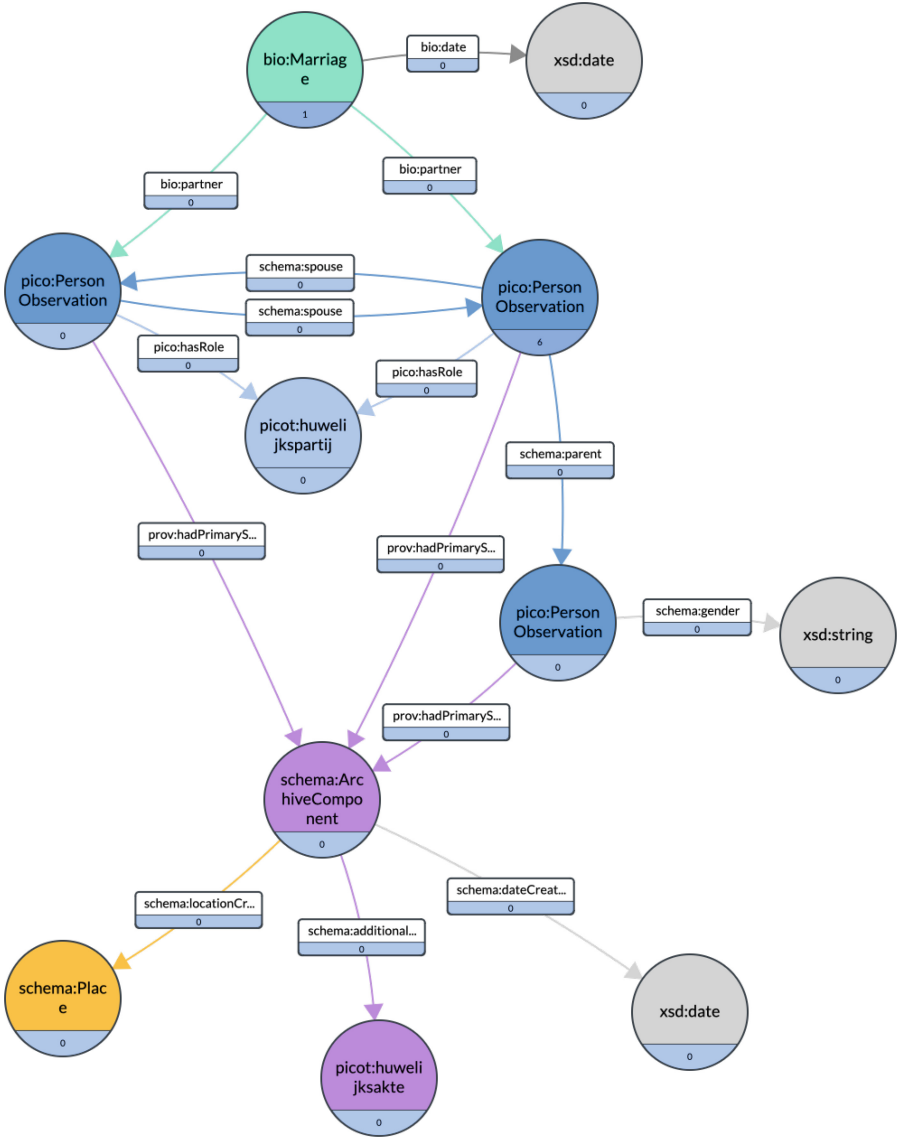


Fig. 1. Marriage certificate according to the PiCo model [48]

### 3.2 Vocabularies

Currently, eleven concepts in historical demography are described with publicly available, standardized vocabularies, 1. cause of death titles, 2. cause of death groupings, 3. occupational titles, 4. occupational groupings, 5. social status codes, 6. place names, 7. grouped geographically defined (administrative)



areas, 8. titles of religious denominations, 9. groupings of religious denominations, 10. (household) relations, and 11. data quality flags, as shown in Table 4.

**Table 4.** Available vocabularies within historical demography

Topic	Standardized titles	Groupings	Status codes
Cause of death	[17, 45]	[17, 45]	–
Occupations	[4, 10, 20, 26, 32, 45]	[37, 49, 55]	[7, 19, 31, 41, 50, 53]
Place names	[12–14, 16]	[43, 44, 52]	–
Religious denominations	[28, 51]	[3]	–
(Household) relations	[1, 25, 27, 40, 46]	–	–
Data quality flags	[1, 25, 27, 40, 46]	–	–

Occupational entries are by far the most standardized. Researchers have been using occupational status schemes for several decades, which has resulted in clear pipelines for processing occupational information. Coding occupational information generally consists of three steps. First, entry errors, abbreviations, and spelling variations are removed to standardize occupational titles [4, 10, 20, 26, 32]. Second, these occupational titles are grouped into occupational groups using intermediate coding schemes. In Europe this is generally HISCO, a system developed by two historical sociologists [49], whereas the standards in the USA are OCC1950 and OCC1990, two systems developed by the United States Bureau of the Census [55]. In a third step, these occupational groups are assigned occupational status codes, such as Duncan’s socioeconomic index [7], HISCLASS [50], HISCAM [19], Nam-Powers-Boyd occupational scores [30, 31], Siegel prestige score [41], SOCP0 [53], and other social status measures. Table 5 provides an example from the SwedPop database.

The coding system behind occupational titles shows that three steps are required for coding historical concepts. 1. standardization of titles, 2. grouping into codes, and 3. assigning status codes. The arduous nature of these steps is shown by the SHiP project, which aims to standardize cause of death titles for multiple countries in Europe and codify them [17]. The project received funding to build a historical causes of death network out of existing collaborations. Over the past six years, researchers from around Europe have been working together to standardize causes of death titles and categorize them. Their goal is to develop a vocabulary that can “deal well with large numbers of historical disease descriptions, from different linguistic areas in Europe, while at the same time it is able to connect to current day disease patterns”. The fruits of their labour are expected to be presented later in 2024.

The amount of effort that was invested in creating a historical International Classification of Diseases (ICDh) also turned out to be its biggest strength. Over the past few conferences anticipation has slowly been building. A year before its launch, ICD10h has already been accepted as the de facto standard

for standardizing and coding historical cause of death titles, as a wide range of researchers contributed. Individual efforts to standardize historical information, such as the Linked International Classification for Religions (LICR) [3], have been far less successful. SHiP shows that specialized vocabularies only succeed if multiple scholars come together as a network, commit time to exchange expertise, and create excitement for their vocabularies.

**Table 5.** Excerpt from the SwedPop standardization, grouping, and status assignment of occupational codes using HISCAM [45]

OCCUPATION STANDARD	HISCO	S	R	P	HISCAM
FARTYGSARBETARE	98100	-9	-9	-9	65
FD TREDJE KLASS FARTYGSARBETARE	98100	-9	21	-9	65
BÅTFÖRMAN	98120	-9	-9	-9	60
BÅTFÖRMANÄNKA	98120	-9	11	-9	60
BÅTKARLFÖRMAN	98120	-9	-9	-9	60
FD BÅTFÖRMAN	98120	-9	21	-9	60
FISKEBÅTSFÖRMAN	98120	-9	-9	-9	60
VATTENBÅTFÖRMAN	98120	-9	-9	-9	60
SJFRMAN	98130	31	-9	-9	53
ANDRA KLASS SJÖMAN	98135	-9	-9	-9	53
ANDRA KLASS SJÖMAN VID FLOTTAN	98135	-9	-9	-9	53
ANDRA KLASS SJÖMANHUSTRU	98135	-9	11	-9	53
...					
BESÄTTNINGSLÄRLING	98140	33	-9	-9	53
BESÄTTNINGSPOJKE	98140	-9	-9	-9	53
BÅTDRÄNG	98140	-9	-9	-9	53
DÄCKMATROS	98140	-9	-9	-9	53
...					
BOGSERBÅTSBESÄTTNINGSKARL	98190	-9	-9	-9	46
BÅTBITRÄDE	98190	-9	-9	-9	46
BÅTFÖRARE	98190	-9	-9	-9	46
BÅTFÖRAREARBETARE	98190	-9	-9	-9	46

*S*, *R*, and *P* refer to the HISCO status, relation, and product code [19, 45, 49]

It is important to realize that work on specialized vocabularies can focus on international as well as longitudinal comparisons. However, local comparisons over time generally require a higher level of detail. Work on standardization of place names and geographically defined (administrative) areas shows that discussions do not necessarily have to take place in international contexts. While modern place name vocabularies are widely available [12, 13, 56], for historical

place names around the world we are less spoilt for choice, although significant effort is being made [14]. These place name vocabularies, or gazetteers, are rarely the best solution to refer to historical spaces, as administrative areas tends to change over time. Therefore, historical demographers have matched place names to geographically and temporally defined geometries, which are generally made available as shape files. In turn, these geometries can be used to calculate spatial statistics [43, 44, 52].

It is beyond the scope of this article to deeply delve into the relationship between place names and associated geographic units. However, researchers with strong ties to spatial demography are doing their best to model the underlying complexities within countries and assign shape files to administrative areas. For example, the Amsterdamse code (AMCO) was developed for the Netherlands to solve issues with applying modern coding systems for municipalities to historical settings [52]. This system works well for the 19th and 20th centuries when administrative areas were more or less fixed. However, a more flexible semantic model was necessary to refer to the more fluid premodern administrative areas within the Low Countries [43, 44]. Such regional or national geographical efforts are generally more useful than one-size-fits-all solutions that span the globe, as they allow users to define how spatial units should be grouped. Yet, such flexibility should not come at the cost of intelligibility. Therefore, it may be worthwhile to explore the use of discrete global grids to create a system for making intermediary layers of historical administrative areas, so that national insights can be translated to the international community. For example, to allow comparisons between Dutch municipalities [52] and Swedish parishes [45].

For other historical concepts, large-scale discussions on how to standardize titles and categorize have not started yet. As a result, data on religion, (household) relations, and data quality flags is much less standardized. Shared vocabularies are either not available or not accepted within the field, so database managers and researchers use their own categories. What is currently needed is to have the local experts join forces in a network to combine these insights in specialized vocabularies. The existence of local coding system means that future efforts to develop specialized vocabularies for religion, (household) relations, and data quality do not have to start from scratch. However, the SHiP project shows that simply having expertise is not enough, as transforming local coding systems into a shared vocabulary takes time and collaborative effort.

### 3.3 Conversion Tools

There are currently no tools available to move between schemas. However, a sizable number of conversion tools exists to move between vocabularies. All but one conversion tools are available for occupation-related vocabularies, making it by far the most vibrant ecosystem. The only other conversion tool groups causes of death titles into the ICD10h.

Table 6 lists the conversions tools that are currently publicly available. There are tools for three different processes: to group titles into categories, to move

between two systems of categorization, and to assign status codes to categories. Different people provide these conversion tools. Tools to categorize titles are generally provided by data centers who produce them as auxiliary data [4, 10, 20, 26, 32, 45]. This allows them to assign status codes using schemas that were developed by scholars [49, 50, 53], while crosswalks between occupational groupings were made by individual scholars in order to compare European, American, or project- specific social status definitions [27, 59].

In order to make conversions possible, conversions should be as straightforward as possible. Nevertheless, conversions from grouped titles to status codes can require additional information on the context. For example, HISCLASS requires information on whether labourers live in an urban or rural environment [50], useful categorizations of cause of death differ by the age of the deceased [17], and conversions from place names to geographically defined (administrative) units require an observation year. The easiest way for conversion tools to assign such context-specific status codes is a rule-based software scripts. However, the problem with scripts is that they need regular maintenance, are not intelligible to all scholars, and can easily become pretty complex. Therefore, the better practice is to make conversion tables, as they are low-maintenance, easy to understand, and limit the number of ways in which data can be split up.

The availability of conversion tools is indicative of an environment in which researchers and database managers share best practices. Sharing conversion tools and their underlying methodology prevents redundant work and ascertains data quality. Nevertheless, there is still a sizable number of conversion tools that are not publicly available. Multiple institutes have crosswalks that are often shared upon request, but hidden from view to possibly interested researchers and database managers. Sharing these conversion tools is low-hanging fruit in improving the interoperability of historical demographic data.

**Table 6.** Publicly available conversion tools in historical demography

Process	From	To	Author
Categorize titles	Cause of death title	ICD10h	[17, 45]
Categorize titles	Occupational title	HISCO	[4, 10, 20, 26, 32, 45]
Convert groupings	HISCO	OCC1950	[27]
Convert groupings	OCC1950	HISCO	[27]
Convert groupings	HISCO	OCCHISCO	[59]
Assign status	HISCO	HISCAM	[20]
Assign status	HISCO	HISCLASS	[20]
Assign status	HISCO	SOCPO	[20]

## 4 CLAIR-HD Website

To help the field in adopting open science practices, we created a website to inform researchers about existing schemas, vocabularies, and vocabularies. The web pages are hosted and maintained by the International Institute of Social History (IISH) in Amsterdam to ascertain that these pages remain accessible in the foreseeable future. Moreover, all provided material is also stored in the IISH data repository to improve its reusability. The website can be found at: <https://iisg.amsterdam/en/blog/clair-hd>.

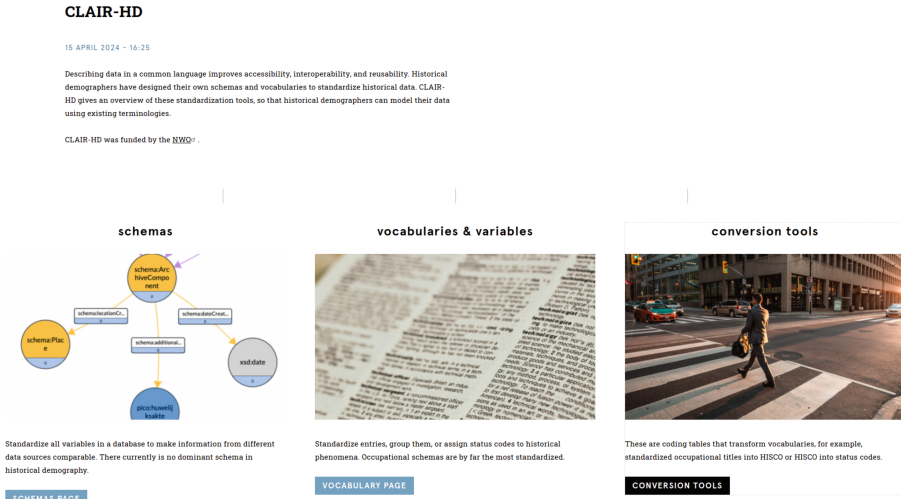


Fig. 2. CLAIR-HD landing page

Theme	Variables	IPUMS-International	MOSAIC	NAPP
<b>Geography</b>	Country code	CNTRY	country	CNTRY
	Place	-	place	-
	Region	-	region	-
	Urban-rural status	URBAN	urban	URBAN
<b>Household</b>	Group quarter status	GQ	gq	GQ
	Household size	PERSONS	hhsize	NUMBERHH
	Household weight	WTHH	hhwt	HHWT
<b>Identifier</b>	Enumeration	SAMPLE	id_enum	SAMPLE
	Household	SERIAL	id_hhold	SERIAL
	Person	PERNUM	id_pers	PERNUM

Fig. 3. CLAIR-HD schema page

# Vocabulary Page (Value Lists)

21 MAY 2024 - 12:24

Historical concepts can be described using standardized vocabularies. Three steps are required for coding historical concepts:

1. standardization of titles,
2. grouping into codes
3. assigning status codes.

For example, "farm lab" gets standardized into "farm labourer", which gets the HISCO code 62110 or OCC1950 code 820. In turn, these codes can be turned into a class score, such as HISCLASS 8 "Farmers and fishermen", or an occupational status score, such as 51 on HISCAM or 50 on Nam-Powers-Boss.

A sizable number of conversion tools exists to move between vocabularies and are listed on the main page.

## Cause of death

Causes of death are being standardized for multiple European countries by the SHiP project. The associated researchers will also provide a crosswalk to convert the standardized causes of death into the *Historical International International Statistical Classification of Diseases and Related Health Problems* (ICD10h).

Standardized occupational titles for Sweden are available at [SwedPop](#)<sup>cf</sup>. Other titles are forthcoming.

## Occupation

Occupational schemas are by far the most standardized. Researchers have been using occupational status schemes for several decades, which has resulted in clear pipelines for processing occupational information.

Standardized occupational titles are available at the [IISH dataverse](#)<sup>cf</sup> and [SwedPop](#)<sup>cf</sup>.

Occupational titles can be grouped into occupational groups using intermediate coding schemes. In Europe this is generally HISCO, a system developed by two historical sociologists, whereas the standards in the USA are OCC1950 and OCC1990, two systems developed by the United States Bureau of the Census.

Occupational groups are assigned occupational status codes, such as Duncan's socioeconomic index, HISCLASS, HISCAM, Nam-Powers-Boyd occupational scores, Siegel prestige score, SOCP0, and other social status measures.

**Fig. 4.** CLAIR-HD vocabulary page

Figure 2 shows the CLAIR-HD landing page which contains a brief explanation on the used terminology and links to pages with more information. This page links to a schema page, a vocabulary page, and a conversion tool page. The schema page summarizes the design principles of the different schemas (see Fig. 3). IPUMS, MOSAIC, and NAPP are compared in one table to see how these

## Conversion tools

21 MAY 2024 - 12:34

These tools are coding tables that transform vocabularies, for example, standardized occupational titles into HISCO or HISCO into status codes.

### Cause of death titles - groupings:

- [Cause of death titles - ICD10h](#)

### Occupational titles - HISCO:

- [\[IT\]](#)
- [\[NL\]](#)
- [\[SE\]](#)
- [\[UK1\]](#)
- [\[UK2\]](#)

### Occupational groupings - groupings:

- [HISCO – OCC1950](#)
- [HISCO – OCCHISCO](#)
- [OCC1950 – HISCO](#)

### Occupational groupings - status codes:

- [HISCO – HISCAM](#)
- [HISCO – HISCLASS](#)
- [HISCO – SOCPQ](#)

**Fig. 5.** CLAIR-HD conversion tool page

schemas name variables relating to the household, identifier, individual, person name, provenance, quality indicators, and the source. A second table provides a similar comparison for IDS, LINKS-gen, and PiCo. Combined, these two tables give a quick overview of the schemas' different design principles.

The vocabulary page explains the three stadiums of standardization and presents standardization efforts per concept (see Fig. 4). Per variable, efforts to standardize variables are introduced as running text and are accompanied by paragraphs on available titles, groupings, and social status codes.

The conversion tool page lists the existing tools and contains links to GitHub, institutional webpages, and repositories where these tools are hosted (see Fig. 5).

## 5 Conclusion and Discussion

There are multiple efforts within the field of historical demography to standardize information using schemas and vocabularies. There are two ways in which the field has developed standardization efforts. On the one hand, there are data centers that developed their own practices in how to standardize data. This information is of great use to other scholars, but only if these “habits” have become accepted practice in the field. Therefore, institutions have worked together on improving the interoperability of historical demographic data. However, the schemas that resulted from these collaborations have limited interoperability and are not widely adapted. There is considerable overlap between the concepts that schemas in historical describe. Yet, most schemas seem to be limited by the institutional context in which they were developed, as they describe concepts with their own terminology, rather than re-using existing vocabularies.

A notable exception is *Persons in Context*, which uses a concentric model to describe data to a wide an audience as possible [5]. A first general layer is used to describe common concepts, domain-specific layer specifies terminology that are well-known within fields, and specialized vocabularies to standardize concepts unique to historical demography. This practice makes data more easily findable, accessible, and interoperable, especially smaller datasets and other “long-tail data” that can easily be obscured from view [57]. Yet, it is currently impossible to effectively describe historical datasets concentrically, due to a lack of specialized vocabularies within historical demography. Therefore, more specialized vocabularies are necessary to make historical demographic datasets (re)usable for a wider audience.

Specialized vocabularies within historical demography are mostly focused on occupational status on occupational titles, groupings, and status. This is indicative of the general process in the field where information is standardized, categorized, and operationalized. Currently, a network of researchers is working hard to standardize historical causes of death titles and simultaneously introduce a historical cause of death classification system. Yet, standardization of other historical concepts, such as place names, religious denominations, and data quality flags is lacking behind. SHiP, the project behind the standardization of cause of death information, shows that the development and adaptation of these specialized vocabularies can succeed by building on existing networks and adapting the principles of team science.

Surprisingly, the development of shared vocabularies to standardize variables is a relatively new phenomenon within the field. Systems to codify occupational clusters or determine occupational status are made by sociologists, rather than historians. The adaptation and use of these vocabularies shows that historical demographers are willing to use standardized variables and that they prove results. The efforts by the SHiP network [17] to standardize historical causes of death titles and codify them is the first example of historians working together to develop their own vocabularies and serves as a blueprint for how the field of historical demography, but also other fields within the humanities and social science, can develop new specialized vocabularies.



CLAIR-HD highlights the hard work done by other scholars. It gives an overview of what has been done, and what work could still be done. The website will be kept up to date for at least the coming 10 years, so that researchers and database managers can publish their data with similar standards or explain why existing standards are insufficient. The information provided by CLAIR-HD facilitates creative discussions and makes enquiries into historical demography easier and more insightful. By doing so, CLAIR-HD will serve as a case study that offers direction for other fields in the humanities and social sciences.

**Acknowledgements.** This publication is part of the project “Common Language for Accessibility, Interoperability, and Reusability in Historical Demography” (CLAIR-HD) with project number 203.001.166 of the research programme Open Science Fund which is (partly) financed by the Dutch Research Council (NWO).

## References

1. Alter, G., Mandemakers, K.: The intermediate data structure (IDS) for longitudinal historical microdata, version 4. *Hist. Life Course Stud.* **1**(1), 1–26 (2014)
2. Alter, G., Mandemakers, K., Gutmann, M.: Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Hist. Soc. Res.* **34**(3), 78–114 (2009)
3. Askhpour, A.: Linked international classification for religions. IISH Data Collection (2017). <https://hdl.handle.net/10622/MHJWRZ>
4. Basten, S.: Basten\_Northern\_English\_Parishes\_1777-1812, IISH Data Collection (2016). <https://hdl.handle.net/10622/YK84PG>
5. CBG.: Persons in Context. GitHub (2023). <https://github.com/CBG-Centrum-voor-familiegeschiedenis/PiCo>
6. Davis, I., Galbraith, D.: BIO: a vocabulary for biographical information. vocab.org (2004). <https://vocab.org/bio>
7. Duncan, O.D.: A socioeconomic index for all occupations. In: Reiss, A., Duncan, O.D., Hatt, P.K., North, D.C. (eds.) *Occupations and Social Status*. Free Press, New York (1961)
8. Edvinsson, S., Mandemakers, K., Smith, K.R., Puschmann, P.: *Harvesting: The Result and Impact of Research based on Historical Longitudinal Databases*. Radboud University Press, Nijmegen (2023)
9. Fauve-Chamoux, A., Bolovan, I., Sogner, S.: *A Global History of Historical Demography: Half a Century of Interdisciplinarity*. Peter Lang, Bern (2016)
10. Fornasin, A., Marzona, A.: HISCO\_Italian\_Formasin\_Marzona\_2006, IISH Data Collection (2016). <https://hdl.handle.net/10622/SRVW6S>
11. Gao, S.S., Sperberg-McQueen, C.M., Thompson, H.: W3C XML schema definition language (XSD) 1.1 part 1: Structures (2012), <https://www.w3.org/TR/2012/REC-xmlschema11-1-20120405/>
12. GeoNames Geographical Database. <https://www.geonames.org/>
13. Getty Thesaurus of Geographic Names® (2017). <http://www.getty.edu/research/tools/vocabularies/tgn/>
14. Grossner, K., Mostern, R.: Linked places in world historical gazetteer. In: 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities 2021), Beijing, China (2021). <https://doi.org/10.1145/3486187.3490203>

15. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: evolution of structured data on the web. *Commun. ACM* **59**(2), 44–51 (2016). <https://doi.org/10.1145/2844544>
16. Huijsmans, D.P.: HSN Gazetteer, IISH data collection (2020). <https://hdl.handle.net/10622/ZDT2DJ>
17. Janssens, A.: Constructing SHiP and an international historical coding system for causes of death. *Hist. Life Course Stud.* **10**, 64–70 (2021). <https://doi.org/10.51964/hlcs9569> <https://doi.org/10.51964/hlcs9569>
18. Janssens, A., Devos, I.; The limits and possibilities of cause of death categorisation for understanding late nineteenth century mortality. *Soc. Hist. of Med.* **35**(4), 1053–1063 (2022). <https://doi.org/10.1093/shm/hkac040>
19. Lambert, P.S., Zijdeman, R.L., Van Leeuwen, M.H.D., Maas, I., Prandy, K.: The construction of HISCAM: a stratification scale based on social interactions for historical comparative research. *Hist. Methods J. Quant. Interdiscip. Hist.* **462**, 77–89 (2013)
20. Mandemakers, K. et al.: HSNDB Occupations. IISH data collection (2020). <https://hdl.handle.net/10622/88ZXD8>
21. Mandemakers, K.: “You really got me”. Ontwikkeling en toekomst van historische databestanden met microdata [Development and future of historical databases with microdata]. Rotterdam: Erasmus University Rotterda (2023). <https://doi.org/10.25397/eur.23256467>
22. Mandemakers, K., Alter, G., Vézina, H., Puschmann, P.: Sowing: The Construction of Historical Longitudinal Population Databases. Radboud University Press, Nijmegen (2023)
23. Meroño-Peñuela, A., et al.: Semantic technologies for historical research: a survey. *Semantic Web* **6**(6), 539–564 (2015). <https://doi.org/10.3233/SW-140158>
24. Meroño-Peñuela, A., et al.: CLARIAH: enabling interoperability between humanities disciplines with ontologies. In: Cota, G., Daquino, M., Pozzato, G.L. (eds.) *Applications and Practices in Ontology Design, Extraction, and Reasoning*, pp. 73–90 (2024). <https://doi.org/10.3233/SSW200036>
25. Minnesota Population Center: Integrated public use microdata series, International: Version 7.3. Minneapolis: IPUMS (2020). <https://doi.org/10.18128/D020.V7.3>
26. Mooney, G.: Mooney\_1866.London\_occupational.codes, IISH Data Collection (2016). <https://hdl.handle.net/10622/ERGY0V>
27. Mourits, R.J.: HISCO-OCC1950 crosswalk. DANS Easy (2017). <https://doi.org/10.17026/dans-zap-qxmc>
28. Mourits, R.J., Mandemakers, K., Laan, F., Munnik, C., Meijer, K.: HSNDB standardisation tables. IISH Data Collection (2024). <https://hdl.handle.net/10622/IKB8HO>
29. Mourits, R.J., Van Dijk, I.K., Mandemakers, K.: From matched certificates to related persons. *Hist. Life Course Stud.* **9**, 49–68 (2020). <https://doi.org/10.51964/hlcs9310> <https://doi.org/10.51964/hlcs9310>
30. Nam, C.B., Powers, M.G.: Changes in the relative status of workers in the United States, 1950-1960. *Social Forces* **47**, 158–170 (1968). <https://doi.org/10.1093/sf/47.2.158>
31. Nam, C.B., Boyd, M.: Occupational status in 2000: over a century of census-based measurement. *Popul. Res. Policy Rev.* **23**, 327–358 (2004). <https://doi.org/10.1023/B:POPU.0000040045.51228.34>
32. Pedersen, B., et al.: Lessons learned developing and using a machine learning model to automatically transcribe 2.3 million handwritten occupation codes. arXiv preprint [arXiv:2106.03996](https://arxiv.org/abs/2106.03996) (2020)

33. Quaranta, L., Sommerseth, H.L.: Introduction: intergenerational transmissions of infant mortality using the intermediate data structure (IDS). *Hist. Life Course Stud.* **7**, 1–10 (2018). <https://doi.org/10.51964/hlcs9288>
34. Reid, A., Garrett, E.: Doctors and the causes of neonatal death in Scotland in the second half of the nineteenth century. *Annales de démographie historique* **123**, 149–179 (2012). <https://doi.org/10.3917/adh.123.0149>
35. Revuelta-Eugercios, B., Castenbrandt, H., Løkke, A.: Older rationales and other challenges in handling causes of death in historical individual-level databases: the case of copenhagen, 1880-1881. *Soc. Hist. Med.* **35**(4), 1116–1139 (2022). <https://doi.org/10.1093/shm/hkab037>
36. Roberts, E., et al.: The North Atlantic population project an overview. *Hist. Methods J. Quant. Interdiscip. Hist.* **36**(2), 80–88 (2003). <https://doi.org/10.1080/01615440309601217>
37. Roberts, E., Woollard, M., Ronnander, C., Dillon, L.Y., Thorvaldsen, G.: Occupational classification in the North Atlantic population project. *Hist. Methods J. Quant. Interdiscip. Hist.* **36**(2), 89–96 (2003). <https://doi.org/10.1080/01615440309601218>
38. Ruggles, S., Roberts, E., Sarkar, S., Sobek, M.: The North Atlantic population project: progress and prospects. *Hist. Methods J. Quant. Interdiscip. Hist.* **44**(1), 1–6 (2011). <https://doi.org/10.1080/01615440309601217>
39. Ruggles, S.: The minnesota population center data integration projects: challenges of harmonizing census microdata across time and place. In: 2005 Proceedings of the American Statistical Association, vol. 9, pp. 1405–1415 (2006)
40. et al Ruggles, S., et al.: IPUMS ancestry full count data: Version 3.0. Minneapolis: IPUMS (2021). <https://doi.org/10.18128/D014.V3.0>
41. Siegel, P.M.: Prestige in the American occupational structure. Doctoral dissertation, University of Chicago (1971)
42. Song, X., Campbell, C.D.: Genealogical microdata and their significance for social science. *Ann. Rev. Sociol.* **43**, 75–99 (2017). <https://doi.org/10.1146/annurev-soc-073014-112157>
43. Stapel, R.J.: Conflating historical population statistics using a historical GIS with a flexible semantic model for premodern administrative units in the low countries: the (Re)counting the uncounted and historical atlas of the low countries projects. In: *GeoHumanities 2023: Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, pp. 56–59 (2023). <https://doi.org/10.1145/3615887.3627756>
44. Stapel, R.J.: Historical atlas of the low countries. A GIS dataset of locality-level boundaries (1350-1800). *Res. Data J. Humanit. Soc. Sci.* **8**, 1–32 (2023). <https://doi.org/10.1163/24523666-bja10033>
45. SwedPop: Swedish population databases for research (2024). <https://swedpop.se/>
46. Szołtysek, M., Gruber, S.: Mosaic: recovering surviving census records and reconstructing the familial history of Europe. *Hist. Family* **21**(1), 38–60 (2016). <https://doi.org/10.1080/1081602X.2015.1006655>
47. Van Erp, M., et al.; More than the name of the rose: how to make computers read, see, and organize smells. *Am. Hist. Rev.* **128**(1), 335–369 (2023). <https://doi.org/10.1093/ahr/rhad141>
48. Van Herck, S., Mourits, R.J.: Upcycling the Dutch Civil Registry using Linked Data, legacy4reuse. Bamberg, Germany (2023)
49. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical International Standard Classification of Occupations. Leuven University Press, Leuven (2002)

50. Van Leeuwen, M.H.D., Maas, I.: HISCLASS: A Historical International Social Class Scheme. Leuven University Press, Leuven (2011)
51. Sang-Ajang, H., et al.: Paramaribo ward registers 1828-1847. IISH data collection (2024). <https://hdl.handle.net/10622/VLN8FD>
52. Van der Meer, A., Boonstra, O.: Repertorium van Nederlandse gemeenten vanaf 1812 waaraan toegevoegd de Amsterdamse code. DANS, Den Haag (2011)
53. Van de Putte, B., Svensson, P.: Measuring social structure in a rural context applying the SOCPO scheme to Scania, Sweden (17 (th)-20 (th) century). *Belgisch Tijdschrift voor Nieuwste Geschiedenis-Revue Belge d'Histoire Contemporain* **40**(1–2), 249–293 (2010)
54. Van Zundert, J.: If you build it, will we come? large scale digital infrastructures as a dead end for digital humanities. *Hist. Soc. Res. Historische Sozialforschung* **37**(3), 165–186 (2012). <https://www.jstor.org/stable/41636603>
55. US Bureau of the Census: Alphabetical Index of Occupations and Industries: 1950. US Bureau of the Census, Washington, D.C. (1950)
56. WikiData. <https://www.wikidata.org/>
57. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**(1), 1–9 (2016). <https://doi.org/10.1038/sdata.2016.18>
58. World Wide Web Consortium: Prov-o: The prov ontology (2013). <https://www.w3.org/TR/prov-o/>
59. Zijdeman, R.L.: OCCHISCO to HISCO (2013). [https://github.com/rlijdeman/occlack/tree/master/crosswalks/occhisco\\_to\\_hisco](https://github.com/rlijdeman/occlack/tree/master/crosswalks/occhisco_to_hisco)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Coding Historical Causes of Death Data with Large Language Models

Bjørn-Richard Pedersen<sup>1</sup> , Maisha Islam<sup>1</sup> , Doris Tove Kristoffersen<sup>1</sup> ,  
Lars Ailo Bongo<sup>2</sup> , Eilidh Garrett<sup>3</sup> , Alice Reid<sup>4</sup> , and Hilde Sommerseth<sup>1</sup> 

<sup>1</sup> Norwegian Historical Data Centre, UiT The Arctic University of Norway, Tromsø, Norway  
{bjorn-richard.pedersen, maisha.islam, doris.t.kristoffersen,  
hilde.sommerseth}@uit.no

<sup>2</sup> Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway  
lars.ailo.bongo@uit.no

<sup>3</sup> University of Edinburgh, Edinburgh, UK  
eilidh.garrett@btinternet.com

<sup>4</sup> Cambridge Group for the History of Population and Social Structure, University of  
Cambridge, Cambridge, UK  
amr1001@cam.ac.uk

**Abstract.** This paper investigates the feasibility of using pre-trained generative Large Language Models (LLMs) to automate the assignment of ICD-10 codes to historical causes of death. Due to the complex narratives often found in historical causes of death, this task has traditionally been manually performed by coding experts. We evaluate the ability of GPT-3.5, GPT-4, and Llama 2 LLMs to accurately assign ICD-10 codes on the HiCaD dataset that contains causes of death recorded in the civil death register entries of 19,361 individuals from Ipswich, Kilmarnock, and the Isle of Skye in the UK between 1861–1901. Our findings show that GPT-3.5, GPT-4, and Llama 2 assign the correct code for 69%, 83%, and 40% of causes, respectively. However, we achieve a maximum accuracy of 89% by standard machine learning techniques. All LLMs performed better for causes of death that contained terms still in use today, compared to archaic terms. Also, they performed better for short causes (1–2 words) compared to longer causes. LLMs therefore do not currently perform well enough for historical ICD-10 code assignment tasks. We suggest further fine-tuning or alternative frameworks to achieve adequate performance.

**Keywords:** Large Language Models · Historical Data · Historical Causes of Death · GPT · Llama · Machine Learning · ICD · ICD-10 · ICD-10h

## 1 Introduction

In historical demography, one of the main tasks is converting the existing knowledge bases of historical registers and microdata into encoded form, so that they can be more easily utilised for research purposes. Many classification systems have been created over the years to help standardise the converted data, such as the Historical International

Standard of Classification of Occupations (HISCO) [1] for occupational information, or the International Classification of Diseases (ICD) [2] for health information. However, a shared, defining trait of all these classification systems is that, when applied to historical data, the encoding requires a tremendous effort in terms of manual work, usually done by domain experts.

The aim of this paper is therefore to explore the possibility of using Large Language Models (LLMs) to automate the encoding of historical causes of death into the ICD-10 system.<sup>1</sup> In the WHO's current ICD-10 classification system, there are over 14,000 unique codes in the base version [4], but several countries have their own, extended versions of previous revisions of the system. Additionally, there are several codes with a high degree of overlap, making it challenging to distinguish one disease or injury from another, for both human experts and automated systems.

LLMs such as GPT (Generative Pre-trained Transformer) and Llama are artificial intelligence models built to both understand and generate human-like text. They are based on the transformer deep learning architecture presented by Vaswani et al. in 2017 [5]. They have multiple uses, in particular within the field of Natural Language Processing (NLP) [6–8], and have shown great potential for understanding text and for classification problems [9, 10]. Prompt engineering is a key feature for LLMs. All models require a prompt, which is the input to the model provided by a user through text, and by combining this with a set of instructions as to how the model should behave we can expect model performance to increase [11]. It is even possible to provide the models with highly specific examples of desired behaviour and outputs [12].

A notable drawback with LLMs is their propensity to “hallucinate” [13], meaning that the models present outputs that are falsehoods. Also, LLMs are black boxes and therefore lack the needed transparency to explain their responses. Other machine learning (ML) models can provide a confidence metric for how confident the underlying algorithm is about a predicted value. This is not possible for an LLM because it does not predict an answer to the input text. Rather, it generates a response text by continually predicting which token should be next in the sentence it is constructing [14].

Despite these drawbacks, LLMs have previously been used to generate ICD-10 billing codes<sup>2</sup> from modern hospital datasets [15, 16], and given the potential upsides in terms of time and resources saved through an automated system, we believe that performing an exploratory study into using LLMs to encode historical causes of death is worthwhile. We evaluate how current LLMs would perform for the specialised task of classifying a dataset containing a mixture of historical and currently used terms for a variety of causes of death.

---

<sup>1</sup> Initially, the primary goal of the study was to assign ICD-10h codes, a version of ICD-10 created for use on historical causes. However, due to the unavailability of the ICD-10h system online, we concentrated on ICD-10. We believe the findings of this study are transferable to ICD-10h when it is accessible to LLMs [3].

<sup>2</sup> Billing codes are standardized codes used in the healthcare industry of some countries to represent various medical procedures, treatments, and services. These codes serve as a means of communication between healthcare providers, insurance companies and regulatory bodies, facilitating the billing and reimbursement process for medical services.

Up until now the work of assigning ICD codes to historical causes of death has been a process done manually by domain experts [17–20], but state-of-the-art ICD classifications for contemporary datasets [21–23] are usually achieved using pre-trained encoder transformers such as BERT [24] or automated coding systems such as ACME [25, 26]. Recently, we have seen an increased interest in exploring the capabilities of LLMs for ICD classification and other health data related tasks. Soroush et al. [15] have assessed the performance of OpenAI’s GPT-3.5 and GPT-4 when generating ICD billing codes, but found that the LLM’s tendency to “hallucinate” key details would present too much of a problem for actual implementation in a healthcare context. Boyle et al. [16] created a novel tree search approach, guided by the LLM and based on the ICD code description, and managed to achieve competitive results without doing any task-specific training.

## 2 Methods

### 2.1 LLMs

We employ OpenAI’s default GPT-3.5-turbo and GPT-4 (Generative Pre-trained Transformer), as well as Meta’s Llama 2 model, specifically the Llama-2-13b-chat-hf. The GPT models were accessed through OpenAI’s API, while the Llama 2 model ran on a local machine. Note that using this API is a paid service [27]. We performed the experiments between November 1st and 8th in 2023. We kept the model parameters at a default level to get the models’ base level performance on the task, but we also employed prompt engineering to explore how this impacts the LLMs’ performance.

### 2.2 Prompt Engineering

The way a prompt is phrased has a massive impact on the LLM output. The models are also very susceptible to the tone of the user’s feedback. If a user says that the answer is incorrect, the model changes its answer. Hence, we found that a set of very clear and concise instructions on how to behave, in a very neutral tone, works best. The structure of a prompt can vary significantly depending on the specific task or desired outcome, as well as the type of LLM.<sup>3</sup> Usually, however, there will be a System prompt and a User prompt. The System prompt defines the base behaviour or sets the context for the model, it can also include specific instructions. In our example below, most of the prompt is a System prompt, defined by writing “role: system”. The User prompt follows next and is where the user interacts with the model by providing a query or input. This part is defined by writing “role: user”. Our prompt for the GPT models is shown below.

---

<sup>3</sup> The prompt shown in this text follows the format required by OpenAI’s GPT models. Other types of LLMs may have different structures and syntax.

```
{'role': 'system', 'content': ""Assistant is an intelligent chatbot designed to help the user assign clinical ICD-10 codes to causes of death.
Instructions:
- Only answer using standard ICD-10 codes, do not use ICD-10-CM billing codes.
- Only return a single ICD-10 code per injury and/or disease found in the given cause of death.
- Each ICD-10 code should not consist of more than 5 characters, the typical format looks like this: 'X01.0'
- Your answers should be in the following format: 'Cause of death: <CAUSE OF DEATH>, ICD-10 code: <ICD-10 CODE>'
- If you are unsure of an answer, do not try to guess. Instead, write the following reply: 'Cause of death: Unknown, ICD-10 code: E99.9'.""},
{'role': 'user', 'content': input_from_user +
'?'}
```

In our case, we are supplying the LLM with a cause of death in the User prompt, and the instructions for how to respond in the System prompt. This would, ideally, result in the following type of exchange between the user and the LLM:

Input – tabes mesenterica?

Output - Cause of death: Tabes mesenterica, ICD-10 code: A18.3.

### 2.3 Historical Causes of Death (HiCAD) Dataset

The dataset used in this project was created by a team at the Cambridge Group for the History of Population (CAMPOP) at the University of Cambridge. It covers three areas of the UK, the town of Ipswich in England, and the town of Kilmarnock and the Isle of Skye, both in Scotland. It spans the period 1861 to 1901 and includes 45,687 individual registered deaths. There are only infant deaths (i.e., a child who died prior to their first birthday) for Ipswich, while Skye and Kilmarnock also have deaths from other age groups, hence roughly 1/3 of the total deaths are those of infants. Historic causes of death are notoriously difficult to code, since there are many blank or illegible entries as well as archaic terms, vague causes, and symptoms. Although it was a legal requirement for the cause of death to be certified by a doctor in both England and Scotland at this time, this did not universally happen, particularly when the deceased had not been treated by a doctor during their last illness. In such cases, a cause of death suggested by the informant (usually a relative) might have been recorded instead. Patterns of medical treatment mean that the deaths of the very young, the very old, and those who died from accidents or acute conditions were less likely to have been allocated a cause by a doctor. On the other hand, the ease of identifying particular causes, such as certain infectious diseases with very characteristic marks like the distinctive rash from smallpox or the



red, swollen tongue of scarlet fever, means that the reporting of deaths from such causes may be more reliable, even if the informant was a lay person [28, 29].

This dataset was manually constructed, in batches, by two domain experts over at least a decade. It has been coded to both the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) as well as a variant of the ICD-10 called the ICD-10h, to accommodate causes of death found in historical populations. The latter was originally developed for the Digitising Scotland project, based at the University of Edinburgh [30]. Currently it is being expanded for European comparisons by the SHiP+ network [3].

The ICD-10h version is currently only available offline and therefore not part of any LLM training set. This makes it unfeasible to directly use current commercial LLMs to classify causes of death into the ICD-10h system without fine-tuning or embedding the knowledge into the model. However, since each cause of death in the dataset was given both a corresponding ICD-10 and ICD-10h code, we can instead use the ICD-10 code as the target for classification.

From this dataset of 45,687 registered deaths, we constructed a smaller dataset of cause of death strings by randomly sampling 19,361 individuals. This was done to reduce the cost of using the GPT models. We will refer to this smaller dataset as the HiCaD (Historical Causes of Death). The original dataset consisted of 21 variables. In addition to personal information about the individual (sex, age at death, length of last illness), it also contained the original cause of death text string from the death registration. This string could contain multiple diseases suffered and/or injuries sustained by the person, which were thought to contribute to their death. Each one of these distinct diseases and/or injuries had been separated out, standardized, and given an ICD-10h code, by the original domain experts. For the HiCaD dataset however, we only kept the variables that were necessary for doing predictions through the LLMs and to analyse the results; the original cause of death text string, the first injury/illness reported as a cause of death and its corresponding ICD-10/ICD-10h code, and finally the historic category of disease (e.g., airborne disease, water- and foodborne disease) assigned.

## 2.4 The Hierarchical Structure of ICD-10 Codes

The ICD-10 coding system represents specific diseases or injuries as an alphanumeric code, usually up to 4 characters long. The ICD classification system is constructed using a hierarchy with 5 levels (Table 1). The first level corresponds to a chapter within the classification system but is not represented as a character in the final code. Levels 2 and 3 serve to narrow down the type or location and are represented by the first 3 characters in the code, called blocks. Most of the three-character categories are subdivided by means of a fourth, numeric character after a decimal point, allowing up to 10 subcategories (0–9); this is the fourth level. The final level is the actual code that is given to the disease or injury.

Within the ICD-10 and ICD-10h coding systems, there exist variables that categorise causes of death by type. For the ICD-10 system this is the chapter, and the ICD-10h equivalent is called Historical category (Histcat). Late twentieth and twenty-first century data can be accurately classified using the ICD-10 chapters, but chapters aren't always appropriate or useful for classifying historic causes of death. This is partly because codes

**Table 1.** Example overview of the hierarchical structure of the ICD-10 system

Level	Level name	Code range	Description
1	Chapters	I–XXII	Each chapter represents a type of disease or injury
2	Blocks	A00–B99	Certain infectious and parasitic types
3	Categories	A15–A19	Tuberculosis
4	Subcategories	A15.0–A15.9	Respiratory tuberculosis, bacteriologically and histologically confirmed
5	Code	A15.1	Tuberculosis of lung, confirmed with culture only

were assigned to words or terms which could change meaning over time and partly because the lack of specificity in historic causes means that many historic causes end up in the “Signs and Symptoms” Chapter of ICD-10. Histcat offers a more historically sensitive classification which ensures codes are grouped in ways which reflect 19th century usage and knowledge.

## 2.5 Experiments

**Correct and Incorrect Classification.** In the experiment, we assessed model performance on the HiCaD dataset by counting the number of codes that were correctly assigned, defined in two ways: a *full match* (exact correspondence with the expert assigned ICD-10 code) and a *partial match* (the first three characters of the model’s output match the expert-assigned code). We accepted partial matches since the fourth character of ICD-10 codes has only ten possible variations, simplifying manual correction. Additionally, we performed an error analysis for the causes of death where the LLMs could not assign a correct ICD-10 code, to determine which types of errors were more common. We subdivided the errors into four categories: 1) The LLM assigned the error code ‘Æ99.9’, which we defined as part of the model prompt; 2) The LLM hallucinated a valid ICD-10 code; 3) The LLM hallucinated an invalid ICD-10 code (here, we define a ‘valid code’ as one found in the WHO’s ICD-10 Version:2019 [31]); 4) The LLM gave a syntactically incorrect output, making it impossible to extract an ICD-10 code for the cause of death.

**Comparison with Alternative Methods.** The second experiment aimed to compare the results of the LLMs with two alternative classification techniques. The first alternative was using traditional Machine Learning-based models, Random Forest [32], and Support Vector Machine (SVM) [33]. To test the models, we created a dataset with two columns: the causes of death text strings from HiCaD and the corresponding ICD-10 code. The preprocessing stage included removal of punctuation marks and extra whitespaces as well as the conversion of all characters into lower case. The sample dataset included 19,360 deaths, which we split into an 80:20 ratio for training and testing. Text feature

extraction was performed using the term frequency-inverse document frequency (TF-IDF) method, and we used GridSearchCV to perform a systematic search over a range of hyperparameters to find the best configuration for our task.

Random Forest is an ensemble learning method that combines multiple decision trees to enhance predictive accuracy. It minimises overfitting and captures complicated relationships well as it builds each tree using bootstrapped data and takes random feature subsets into account.

SVM is a robust algorithm for classification and regression. It seeks the optimal hyperplane to maximise the margin between classes in feature space. For this experiment we used a linear kernel for the SVM, as the kernel's simplicity not only ensures faster training times, but also helps in reducing the risk of overfitting. Additionally, it tends to generalize better when the decision boundary is inherently linear or nearly linear, which is often the case with text data represented in high-dimensional space.

We used 5-fold cross-validation within GridSearchCV to evaluate the performance of different hyperparameter configurations for both models. In 5-fold cross-validation, the data is split into five subsets (folds), and the model is trained on four of them while being tested on the remaining fold. This process is repeated five times, each time with a different fold as the test set. The results are averaged to provide a robust estimate of model performance. This approach helps in ensuring that the model generalizes well to unseen data.

The second alternative was a basic string similarity comparison method, comparing the HiCaD dataset with a dictionary of standardised causes of death, originally created by domain experts.

**Temporal Context of the Causes of Death.** We expected that LLMs would perform worse on terms that only exist in historical registers or have a different meaning from the current understanding of the term. Therefore, our third experiment aimed to compare the number of correct classifications done by the LLMs for both types of terms. We sorted our causes of death into what we refer to as “archaic” and “current” causes, using the manually coded ICD-10h code. If the final character in the ICD-10h code is a 0, it means that this is a cause of death term that exists in both contemporary and historic registers. If the ICD-10h code ends in any other digit, then the cause of death term is only found in historical registers, or the term might be understood differently today.

**Complexity of the Input Values.** Over 80% of the original cause of death text strings within the HiCaD dataset consist of 3 words or less, but the remaining 20% cover a range of 4–41 words. In our fourth experiment, we calculated if the number of words in the input text had any impact on the models' ability to correctly assign the cause of death to an ICD-10 code.

To do this, we grouped the causes of death by word count as short (1–2 words), medium (3–4 words), and long (5+ words) and measured the error rate for each group. These cutoffs were chosen as causes with one or two words are likely to represent just a single cause of death (e.g. “pneumonia”, “scarlet fever”) but causes with more words are more likely to represent accidents and multiple causes of death (e.g. “pneumonia following measles”).

To measure the agreement between the three LLM models, that is, the extent to which all LLMs assigned the same code for each cause of death, the Fleiss Kappa [34]

score was calculated. This was done per word count category and for overall match/no match agreement between the models. Values above 0.2 indicate fair agreement, above 0.4 moderate, above 0.6 substantial, above 0.8 almost perfect.

**Histcat Classification.** Each ICD-10 code can only belong to a single historical category (Histcat). In the fifth experiment we calculated if the distribution of ICD-10 codes within the Histcats in the manually coded HiCaD dataset matched the distribution in the models' outputs. This tells us if there were any particular types of Histcats that the LLMs are worse at classifying than others.

## 3 Results

### 3.1 Correct and Incorrect Classification

For the HiCaD dataset, consisting of 19,353 cause of death text strings, GPT-3.5 coded the causes as a full match in 31% and a partial match in 38%, with no overlap, meaning that 69% of all causes were encoded correctly, and 31% of the causes were given an incorrect ICD-10 code. GPT-4 achieved 58% for full matches and 25% for partial matches, leading to 83% of all causes of death being encoded into the correct ICD-10 code, with an error rate of 17%. Llama 2 coded 9% of all causes as a full match, 31% as a partial match, and had an error rate of 60%. These results show that GPT-4 outperforms GPT-3.5 in the accurate classification of ICD-10 codes, and both models show significantly better performance than Llama 2. These results are summarised in Table 2.

**Table 2.** An overview of how LLMs encoded the historical causes of death.

Model	Full match	Partial match	Correct (full + partial)	Errors
GPT-3.5	31%	38%	69%	31%
GPT-4	58%	25%	83%	17%
Llama 2	9%	31%	40%	60%

When analysing the different errors made by the LLMs, starting with the assignment of the predefined error code 'Æ99.9', GPT-3.5 assigned this code to 19% of its errors, while GPT-4 did so for 7% of its errors. Llama 2, on the other hand, never used the error code. As for hallucinations of valid ICD-10 codes, GPT-3.5 reported these in 50% of cases, GPT-4 in 48%, and Llama 2 showed a higher tendency with 61%. For hallucinations of invalid ICD-10 codes, GPT-3.5 lowest occurrence at 5%, compared to 9% for GPT-4 and a significantly higher 28% for Llama 2. Finally, for syntactical errors, GPT-3.5 had 26%, GPT-4 had 36%, and Llama 2 had the lowest at 11%. These results show that GPT-3.5 was more prone to respond with the error code, whereas more of GPT-4 and Llama 2's errors stem from hallucinations or syntax errors. Additionally, Llama 2 exhibited a higher frequency of hallucinating valid codes compared to the GPT models, but GPT-4 was more likely to hallucinate invalid codes than GPT-3.5. These results are shown in Table 3.

**Table 3.** A breakdown of the prediction errors made by the LLMs.

Error type	GPT-3.5	GPT-4	Llama 2
Error code	19%	7%	0%
Valid code hallucination	50%	48%	61%
Invalid code hallucination	5%	9%	28%
Syntax error	26%	36%	11%

### 3.2 Comparison with Alternate Solutions

Random Forest achieved an accuracy of 87% and SVM achieved an accuracy of 89% on a test set of 19,360 cause of death text strings, sampled from the full dataset of over 45,000 rows.

When using string similarity comparison on the HiCaD dataset, consisting of 19,353 rows, the number of correctly coded causes of death were 6%, meaning that 94% of all causes were assigned an incorrect ICD-10 code.

These results show that both Random Forest and SVM outperform the LLMs, with the best result obtained by an LLM being 83% correctly assigned ICD-10 codes. It does, however, also show that LLMs vastly outperforms string similarity comparison (Tables 4 and 5).

**Table 4.** The results of using alternative methods to assign ICD-10 codes to causes of death.

Method	Correctly assigned	Incorrectly assigned
Random Forest	87%	13%
SVM	89%	11%
String similarity	6%	94%

**Table 5.** Summation of the results from LLMs and alternative methods.

Method	Correctly assigned	Incorrectly assigned
GPT-3.5	69%	31%
GPT-4	83%	17%
Llama 2	40%	60%
Random Forest	87%	13%
SVM	89%	11%
String similarity	6%	94%

### 3.3 Temporal Context of the Cause of Death

Within the HiCaD dataset, we found that 9,773 rows consisted of causes of death that could be categorised as archaic, and 9,580 causes as current. This gives a split of 50.5% to 49.5%. When defining both a full and partial match as correct, we obtained the following results: GPT-3.5 correctly assigned ICD-10 codes to 55% of archaic causes of death with a 45% error rate, and 83% of current causes with a 17% error rate. GPT-4 correctly assigned codes to 75% of archaic causes with 25% errors, and 90% correct with 10% errors for current causes. Llama 2, however, only managed to correctly assign codes to 35% of archaic causes with 65% errors, and 45% of current causes with 55% errors. These results are shown in Table 6.

**Table 6.** Results for LLM classification of archaic and current causes of death terms.

Model	Correct – Archaic	Errors – Archaic	Correct - Current	Errors - Current
GPT-3.5	55%	45%	83%	17%
GPT-4	75%	25%	90%	10%
Llama 2	35%	65%	45%	55%

### 3.4 Complexity of Input Values

Correct assignment of ICD-10 codes for word categories 1–2, 3–4 and 5+ was 74%, 61% and 62% for GPT-3.5, 85%, 82% and 71% for GPT-4, and 45%, 32% and 34% for Llama 2, respectively. Showing that GPT-4 performed better than GPT-3.5 and both outperform Llama 2, this holds for all categories. These results can be seen in Table 7.

When comparing the degree of agreement between the LLMs, i.e. the extent to which all models produced the same output, Fleiss Kappa was lowest for the medium (3–4 words) category with 0.19 meaning slight agreement, for short (1–2) words it was 0.27 which indicates fair agreement. Agreement was fair for the long (5+ words) category with 0.31, meaning that the models tended to give the same output most often for the causes of death that consisted of 5+ words, and the least often for the causes of death that were 1–2 words long. Overall, the models' agreement was measured as 0.27, indicating a fair level of agreement. These results are summarized in Table 8.

**Table 7.** Correct assignment of ICD-10 codes by word category

Model	Word category 1–2	Word category 3–4	Word category 5+
GPT-3.5	74%	61%	62%
GPT-4	85%	82%	71%
Llama 2	45%	32%	34%

**Table 8.** Degree of agreement between models (Fleiss Kappa)

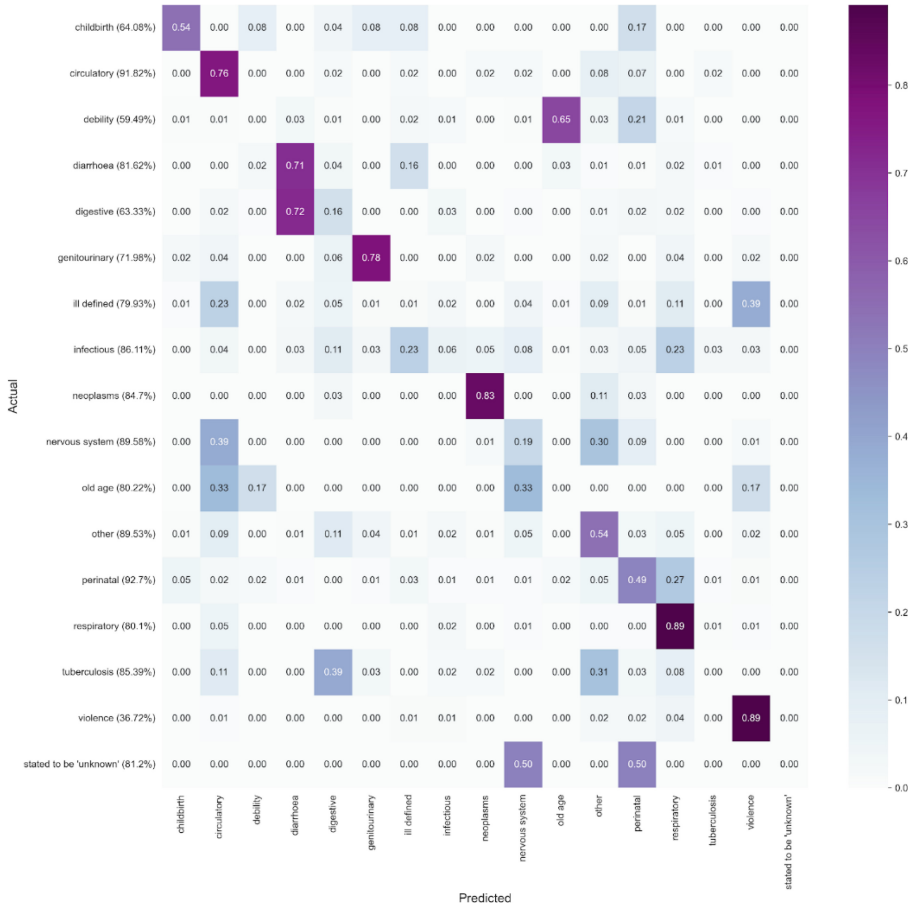
Word category	Fleiss Kappa value	Interpretation
1–2 words	0.27	Fair
3–4 words	0.19	Slight
5+ words	0.31	Fair
Overall	0.27	Fair

### 3.5 Histcat Classification

Figure 1 shows the causes of death which were classified incorrectly by GPT-4, see Table 2, and how these are distributed across the Histcat classification scheme. Despite this figure showing the errors made by the model, it is worth noting that multiple ICD-10 codes can belong to the same Histcat. Therefore, even if the actual ICD-10 code was wrong, it could still be assigned to the correct Histcat, as represented by the clusters along the diagonal of the heatmap. The number of causes that were classified into the correct ICD-10 code, and by extension the correct Histcat, is shown as the percentage value displayed next to each Histcat on the Y-axis. The remaining row values displayed on the X-axis are the causes of death that were classified into an incorrect ICD-10 code as well as an incorrect Histcat. This, all together, means that a desired result in this experiment would be a high percentage value listed after each Histcat on the Y-axis, and alternatively a higher value on the diagonal than the other row values.

For example, in the Histcat “Childbirth”, 64% of causes were assigned a full or partial match by the LLM, as displayed in parentheses following the Histcat name on the Y-axis. Of the remaining 36% of causes which were classified incorrectly, 54% were classified into the correct Histcat but an incorrect ICD-10 code, while 17% were classified into the Histcat “Perinatal”, 8% were placed within “Ill defined”, etc.

GPT-4 showed strong performance across most categories, except for “Violence”, where it only classified the cause to the correct ICD-10 code in 37% of cases. Despite this, it managed to classify those causes into the correct Histcat 89% of the time. The opposite can be seen in the “Debility” category, where GPT-4 managed to find the correct ICD-10 code in 59% of cases, but 0% of the errors for these codes ended up in the correct Histcat.



**Fig. 1.** A heatmap of wrong classifications made by GPT-4 in regard to Histcat. If the model classified the cause of death into the wrong ICD-10 code, but the erroneous code still belongs to the same category, then the errors will cluster on the diagonal. If not, the errors will spread out along the row.

## 4 Discussion

### 4.1 Main Findings

We explored how well out of the box Large Language Models (LLMs) performed on classifying ICD-10 codes for historical death records from three different areas of the UK. GPT-4 performed the best in terms of coding accuracy among GDPR-compliant LLMs, notably surpassing GPT-3.5 and Llama 2, especially for causes of death using current terms. All models showed superior performance with causes of death given in only (1 or 2 words). While model agreement varied with cause length, showing average agreement for longer and poor for medium-length causes, overall substantial agreement was observed. This tendency to make more errors with longer text strings can possibly



be explained by the chance of increased ambiguity and the presence of multiple possible interpretations. We know from the dataset that causes of death from various forms of illness can often only consist of the diseases' names, whereas for external causes of death, like accidents, one is more likely to find complex narratives where interpretation plays a large role in determining the actual cause of death.

Common errors in model classifications were linked to three patterns: abbreviations in the source data (e.g., “con” for “consumption”), coding into the chapter of “Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified” which is known to be more difficult to code, and specific terms like “teething” and “dentition” that are less common in modern datasets but prevalent in historical infant mortality records. We also learned in the breakdown of errors in Table 3 that a significant portion of LLM errors involved hallucinating both valid and invalid ICD-10 codes. This indicates a tendency of LLMs to “guess” plausible answers, which will lead to incorrect classification. We found that GPT-3.5 hallucinated more valid codes than GPT-4 did. A possible reason for this behavior could be that GPT-4 has been trained on more data than GPT-3.5, and as such may have more instances of variant ICD-10 classification systems in its training data. Our inclusion of a predefined error code (“Æ99.9”) was an attempt to include an alternative to hallucinations, but our error analysis shows that this error code was not consistently adopted by all models, with GPT-3.5 using it most and Llama 2 not at all. This suggests variations in how models handle uncertainty, and that this manner of error-handling is not sufficient when dealing with LLMs. In the Histcat classification, the uneven performance we observed may relate to the varying number of codes per Histcat, as categories such as “Violence” which contain more codes (Violence includes 3,278 unique ICD-10 codes) have a higher likelihood of at least partial classification success than a category with less codes like “Debility” which has just 3 unique ICD-10 codes.

Our experiments showed that traditional machine learning models such as Random Forest and SVM outperformed the LLMs in assigning ICD-10 codes. A possible reason for this can be that traditional ML models are known to handle structured classification tasks more effectively than LLMs, which are generally optimized for broader language understanding and generation tasks, without the specialized domain knowledge required for accurately classifying historical causes of death without additional fine-tuning. Despite this, it is worth noting that these ML models were trained and validated on our HiCaD dataset, where the causes of death are not evenly distributed among the ICD-10 codes. As such, there is always a risk of these models becoming overfitted, i.e. not able to generalize to datasets other than HiCaD. This behavior would be less likely in an LLM, due to the increased size of their training data. Nevertheless, all models did surpass the basic string similarity approach using Jaro-Winkler distance.

Based on the experiments, we have learned three main lessons for working with LLMs on this type of classification task:

- The need for domain-specific tuning. Our findings underscore that generalized LLMs cannot compete with models that were trained on domain-specific datasets when it comes to such specialized tasks.
- Importance of error analysis. Detailed breakdowns of error types are crucial for understanding model weaknesses and guiding improvements.

- Consideration of input complexity. As our findings showed that all models performed better on shorter causes of death (1–2 words) compared to longer ones, simplifying inputs wherever possible, or developing strategies to handle complex, narrative inputs, can enhance model performance.

## 4.2 Comparison to Related Work

Our findings mostly correlate with those of Soroush et al. [15], in that we have found the performance of base level LLM models to be inadequate for practical use when it comes to assigning ICD codes to historical causes of death. In our results, most errors did not contain the error code we defined as part of our model prompt, which we instructed the models to use if it did not “know” the answer. Instead, the majority of the errors would fall under the definition of hallucinations, i.e., creating plausible but incorrect statements [13].

We introduced the concept of full matches and partial matches, and we saw that all models had a potential for achieving more correct classifications if they had managed to find the correct subcategory for the cause of death. This could have potentially been done in a similar manner to what Boyle et al. [16] term “meta-refinement of predicted codes”. They used a tree-search approach to find all possible, relevant codes for their data, and then asked GPT-4 to remove false positives. In Sect. 4.3, we discuss the possibility of prompting the LLMs a second time, to find the correct subcategory based on the first 3 characters of the ICD-10 code that were given as output during the first round of prompting.

## 4.3 Future Work

To achieve the best coding rate possible using LLMs on historical causes of death, one technique to use is fine-tuning, a type of transfer learning [35], where pre-trained models are further trained on new, labelled data [36]. This would entail downloading a pre-trained LLM and then fine-tune it on a dataset containing causes of death. Based on previous studies, we expect that this will yield improved results [37, 38].

As seen in Sect. 3, all models gave partial matches as part of their outputs, indicating that if the last character had also been correct, these causes would have been a full match. One possible solution, that can be built upon our current work, is to automate a solution where each cause of death is passed in to a LLM for a second round of classification, this time presenting the model with both the cause of death text string, the almost completed code, as well as a list of possible subgroups that the model is then asked to choose between; this is an approach similar to Boyle et al.’s “meta-refinement of predicted codes”.

Another promising solution is Retrieval-Augmented Generation (RAG) [39, 40]. RAG is an AI framework that can improve LLM responses by grounding the model on external knowledge bases that the users supply. Using RAG, we would be able to use the ICD-10h master list of terms and codes as a knowledge base, and when prompting an LLM to classify a cause of death into an ICD-10h code, it would first retrieve information from the knowledge base before responding. This would increase the probability of getting a correct response and it would give the users a source for the answer.

#### 4.4 Limitations

We intended to use LLMs for assigning ICD-10h codes to the causes of death, as it is tailored for historical terms, but resorted to standard ICD-10 codes due to the ICD-10h's unavailability online, potentially affecting the number of full matches. However, ICD-10h codes have an additional two digits which might have also increased error risks. Furthermore, the LLMs' tendency to align with the ICD-10-CM billing system, despite instructions to avoid it, contributed to the low full match count.

We provided models with original cause of death text strings, which often includes several diseases and/or injuries that the individual was suffering from at the time of death, instructing them to return an ICD-10 code for each. Our analysis compared the first code from the models to the first in our manual dataset, as different causes are not often easy to distinguish. The LLMs may have focused on the true underlying cause mentioned later in the text, as prompted, rather than the first listed condition.

Another limitation is that the analysis was done on records of individual deaths, not on unique strings. This has the advantage of assessing how many deaths were likely to be correctly coded, but not necessarily how well particular commonly written strings were coded. Additionally, the dataset was skewed towards infant deaths.

At the time of the experiment, a more powerful version of the Llama 2 model was trained on 70 billion parameters. However, we could not use this version of the model, as we did not have a powerful enough computer to run it. Hence, we were restricted to the 13 billion parameter model. This has most likely impacted the performance of Llama 2 reported on in this paper.

## 5 Conclusion

In this paper, we explored the use of Large Language Models (LLMs) to encode historical causes of death data. We experimented with the current default versions (The experiments were run in early November 2023) of OpenAI's GPT-3.5 and GPT-4 models, as well as Meta's Llama 2, on a dataset covering the period 1861–1901 gathered from Ipswich, Kilmarnock, and the Isle of Skye in the UK. We found that no LLM performed this task satisfactorily, with GPT-4 achieving a correctness-score of 82.6% at the highest, but it was only able to correctly match the manual coding completely in 57.9% of cases. We found that all LLMs achieved better results for causes of death where more current terms are used, than for archaic terms. They also performed better for causes of death that were composed of shorter text strings. We found that the models made more errors when given longer text strings, but that these errors had a high degree of agreement between models, meaning that the models all tended to make the same error.

We compared the performance of the LLMs to both classical machine learning methods such as a random forest classifier and SVM, and a string similarity matching algorithm. We found that while the LLMs performed much better than string similarity matching, they still perform worse than the machine learning methods.

It is important to keep in mind that these results reflect the use of LLMs that were trained to be chatbots with extensive general knowledge, rather than models that have been specifically pre-trained or fine-tuned with domain expertise in historical causes of death or the ICD-10 system. As such, we have chosen to treat these results as a

baseline that we can use to compare with for future projects where we utilise fine-tuning or frameworks like Retrieval-Augmented Generation. We would not recommend using LLMs for similar tasks without these improvements.

**Acknowledgement.** We wish to thank the General Registry Office for Scotland (GROS), now NRS, for special permission to transcribe the contents of the civil registers of the Isle of Skye and the town of Kilmarnock from 1855–1901, which was carried out under the ESRC funded project Determining the Demography of Victorian Scotland Through Record Linkage (RES-000-23-0128) held at the Cambridge Group for the History of Population and Social Structure, University of Cambridge. We acknowledge funding provided by the Research Council of Norway through the infrastructure project Historical Registers (project number 322231).

**Code and Data Availability.** The dataset used in this study is not publicly available, but all our code and the LLM prompts are available and open source using the MIT licence. It can be found at <https://github.com/HistLab/Coding-historical-causes-of-death-data-with-Large-Language-Models>.

## References

1. Van Leeuwen, M.H.D., Maas, I., Miles, A.: HISCO: Historical International Standard Classification of Occupations. Leuven University Press (2002)
2. Hirsch, J.A., et al.: ICD-10: history and context. *Am. J. Neuroradiol.* **37**, 596–599 (2016). <https://doi.org/10.3174/ajnr.A4696>
3. Janssens, A.: Constructing SHiP and an international historical coding system for causes of death. *Hist. Life Course Stud.* **10**, 64–70 (2021). <https://doi.org/10.51964/hlcs9569>
4. WHO | FAQ on ICD. <https://web.archive.org/web/20041017011702/>. <http://www.who.int/classifications/help/icdfaq/en/>. Accessed 06 Nov 2023
5. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*. Curran Associates, Inc. (2017)
6. Wang, L., et al.: Document-level machine translation with large language models (2023). <http://arxiv.org/abs/2304.02210>, <https://doi.org/10.48550/arXiv.2304.02210>
7. Wang, S., et al.: GPT-NER: named entity recognition via large language models (2023). <http://arxiv.org/abs/2304.10428>, <https://doi.org/10.48550/arXiv.2304.10428>
8. Zhang, B., Yang, H., Zhou, T., Babar, A., Liu, X.-Y.: Enhancing financial sentiment analysis via retrieval augmented large language models (2023). <http://arxiv.org/abs/2310.04027>, <https://doi.org/10.48550/arXiv.2310.04027>
9. Sun, X., et al.: Text classification via large language models (2023). <http://arxiv.org/abs/2305.08377>, <https://doi.org/10.48550/arXiv.2305.08377>
10. Chen, S., et al.: Evaluation of ChatGPT family of models for biomedical reasoning and classification (2023). <http://arxiv.org/abs/2304.02496>, <https://doi.org/10.48550/arXiv.2304.02496>
11. Ahmed, T., Devanbu, P.: Few-shot training LLMs for project-specific code-summarization. In: *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. pp. 1–5. Association for Computing Machinery, New York (2023). <https://doi.org/10.1145/3551349.3559555>
12. Brown, T.B., et al.: Language models are few-shot learners (2020). <http://arxiv.org/abs/2005.14165>, <https://doi.org/10.48550/arXiv.2005.14165>

13. Hallucination Definition & Meaning - Merriam-Webster. <https://www.merriam-webster.com/dictionary/hallucination>. Accessed 20 Nov 2023
14. Bowman, S.R.: Eight things to know about large language models (2023). <http://arxiv.org/abs/2304.00612>, <https://doi.org/10.48550/arXiv.2304.00612>
15. Soroush, A., et al.: Assessing GPT-3.5 and GPT-4 in generating international classification of diseases billing codes (2023). <https://www.medrxiv.org/content/10.1101/2023.07.07.23292391v2>, <https://doi.org/10.1101/2023.07.07.23292391>
16. Boyle, J.S., Kascenas, A., Lok, P., Liakata, M., O'Neil, A.Q.: Automated clinical coding using off-the-shelf large language models (2023). <http://arxiv.org/abs/2310.06552>
17. Bailey, M.J., Leonard, S.H., Price, J., Roberts, E., Spector, L., Zhang, M.: Breathing new life into death certificates: extracting handwritten cause of death in the LIFE-M project. *Explor. Econ. Hist.* **87**, 101474 (2023). <https://doi.org/10.1016/j.eeh.2022.101474>
18. Revuelta-Eugercios, B., Castenbrandt, H., Løkke, A.: Older rationales and other challenges in handling causes of death in historical individual-level databases: the case of Copenhagen, 1880–1881. *Soc. Hist. Med.* **35**, 1116–1139 (2022). <https://doi.org/10.1093/shm/hkab037>
19. Anderton, D.L., Leonard, S.H.: Grammars of death: an analysis of nineteenth-century literal causes of death from the age of miasmas to germ theory. *Soc. Sci. Hist.* **28**, 111–143 (2004). <https://doi.org/10.1017/S0145553200012761>
20. Williams, N.: The reporting and classification of causes of death in mid-nineteenth-century England. *Hist. Methods J. Quant. Interdiscip. Hist.* **29**, 58–71 (1996). <https://doi.org/10.1080/01615440.1996.10112730>
21. Chen, P.-F., et al.: Automatic international classification of diseases coding system: deep contextualized language model with rule-based approaches. *JMIR Med. Inform.* **10**, e37557 (2022). <https://doi.org/10.2196/37557>
22. Edin, J., et al.: Automated medical coding on MIMIC-III and MIMIC-IV: a critical review and replicability study. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2572–2582 (2023). <https://doi.org/10.1145/3539618.3591918>
23. Huang, C.-W., Tsai, S.-C., Chen, Y.-N.: PLM-ICD: automatic ICD coding with pretrained language models. In: Naumann, T., Bethard, S., Roberts, K., Rumshisky, A. (eds.) *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pp. 10–20. Association for Computational Linguistics, Seattle (2022). <https://doi.org/10.18653/v1/2022.clinicalnlp-1.2>
24. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019). <https://doi.org/10.18653/v1/N19-1423>
25. Slik kodes og kvalitetssikres dødsårsaker i Dødsårsaksregisteret. <https://www.fhi.no/op/dodsarsaksregisteret/dodsarsaken-kodes-med-icd-koder/>. Accessed 09 Nov 2023
26. Hernes, E., Johansson, L.A., Fosså, S.D., Pedersen, A.G., Glatte, E.: High prostate cancer mortality in Norway evaluated by automated classification of medical entities. *Eur. J. Cancer Prev.* **17**, 331–335 (2008)
27. Pricing. <https://openai.com/api/pricing/>. Accessed 14 June 2024
28. Reid, A., Garrett, E.: Doctors and the causes of neonatal death in Scotland in the second half of the nineteenth century. *Ann. Demogr. Hist.* **2012**, 149–179 (2013). <https://doi.org/10.3917/adh.123.0149>
29. Alter, G.C., Carmichael, A.G.: Classifying the dead: toward a history of the registration of causes of death. *J. Hist. Med. Allied Sci.* **54**, 114–132 (1999). <https://doi.org/10.1093/jhmas/54.2.114>

30. Scottish Historic Population Platform (SHiPP) | SCADR. <https://www.scadr.ac.uk/our-research/shipp>. Accessed 09 Nov 2023
31. ICD-10 Version: 2019. <https://icd.who.int/browse10/2019/en/>. Accessed 13 June 2024
32. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
33. Support Vector Machines: Theory and Applications. [https://www.researchgate.net/publication/221621494\\_Support\\_Vector\\_Machines\\_Theory\\_and\\_Applications](https://www.researchgate.net/publication/221621494_Support_Vector_Machines_Theory_and_Applications). Accessed 17 June 2024
34. Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971). <https://doi.org/10.1037/h0031619>
35. Spring Research Presentation | College of Physical and Mathematical Sciences. <https://web.archive.org/web/20070801120743/http://cpms.byu.edu/springresearch/abstract-entry?id=861>. Accessed 14 Nov 2023
36. 14.2. Fine-Tuning—Dive into Deep Learning 1.0.3 documentation, [https://d2l.ai/chapter\\_c omputer-vision/fine-tuning.html](https://d2l.ai/chapter_c omputer-vision/fine-tuning.html). Accessed 14 Nov 2023
37. Liga, D., Robaldo, L.: Fine-tuning GPT-3 for legal rule classification. *Comput. Law Secur. Rev.* **51**, 105864 (2023). <https://doi.org/10.1016/j.clsr.2023.105864>
38. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with GPT-4 (2023). <http://arxiv.org/abs/2304.03277>, <https://doi.org/10.48550/arXiv.2304.03277>
39. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks (2021). <http://arxiv.org/abs/2005.11401>, <https://doi.org/10.48550/arXiv.2005.11401>
40. What is retrieval-augmented generation?. <https://research.ibm.com/blog/retrieval-augmented-generation-RAG>. Accessed 16 Nov 2023

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Teaching the Specialized Language of Mathematics with a Data-Driven Approach: What Data Do We Use?

Cecilia Fissore<sup>(✉)</sup> , Francesco Floris , Marina Marchisio Conte ,  
and Matteo Sacchet 

University of Turin, Turin, Italy

{cecilia.fissore, francesco.floris, marina.marchisio,  
matteo.sacchet}@unito.it

**Abstract.** Numerous studies in Mathematics Education have shown that among the causes of disciplinary learning difficulties are the acquisition and understanding of its specialized language. Data-Driven Learning (DDL) is a didactic approach that treats language as data and sees students as researchers doing guided discovery activities. The exploration of corpora can effectively support reflection on the specialized languages of Mathematics. What data should be used? Students daily use the most recent Large Language Models (LLM) and Google, which can be used for linguistic investigations. However, it must be remembered that there is no total control over the data on which searches are carried out, the results, and the type of language they use. Control over the data is important, especially when teachers want to use these tools to design and deliver didactic activities. This paper presents a recent DDL research activity with 80 secondary school students on the specialized language of Mathematics. The students conducted linguistic investigations on a specially designed corpus and carried out corpus-based activities with automatic formative assessment within a Digital Learning Environment. The results show that the students appreciated the proposed activities. They develop linguistic and mathematical skills and become more aware of the importance of the language they use. Students developed digital skills in browsing, searching, and filtering data, as well as in evaluating data, information, and digital content. LLM, such as ChatGPT, could not be used for the same type of activity, but with appropriate design, they can be used as a starting point for investigation and linguistic reflection. In the future, given the notable diffusion of these AI tools, it is essential to train teachers and students on their strengths and weaknesses and how they influence teaching and learning.

**Keywords:** Data-Driven Learning · Large Language Model · Language for Specific Purposes · Mathematics Education

## 1 Introduction

Every subject, from Geography to Mathematics, has a specialized language, which is the typical way of expressing itself in that area. Research in mathematics education has studied how learning difficulties in Mathematics depend on linguistic difficulties, that

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 48–64, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_4](https://doi.org/10.1007/978-3-031-73741-1_4)

is, on the acquisition and understanding of its specialized language [1]. Understanding the language of Mathematics is considered by didactic research to be one of the major obstacles to learning the discipline at all school levels [2]. Furthermore, Mathematics has always been one of the school disciplines where students of all grades encounter the most significant difficulties. For example, the PISA 2022 international triennial survey (to be launched in 2020) shows critical issues in Mathematics for Italian students [3]. Mathematics scores were significantly lower than in 2018 and in previous cycles in almost all types of education. The growth trend seems to have been reversed compared to 20 years ago. For this reason, many researchers in mathematics education study the language that students use in class (with their peers or with the teacher) for questions or collaborative or problem-solving activities that teachers use during explanations, which is present in textbooks, and so on.

It is possible to use the Data-Driven Learning (DDL) methodology to help students develop language skills. DDL treats language as data and sees students as researchers doing guided discovery activities [4]. Students can discover grammatical patterns, word meanings, or other aspects of language by searching linguistic data and investigating large amounts of authentic language. This methodology was born for language teaching and learning but has recently been used for learning specialized languages [5]. DDL is closely linked to the concept of corpus, a collection of texts on which linguistic investigations are carried out.

In recent times, students' use of the Large Language Model (LLM) has spread considerably, also for educational purposes. LLMs are deep neural network models (deep learning) capable of acquiring vast knowledge of a language from enormous quantities of texts, mainly taken from the Web [6]. The impact of LLM on education can be enormous [7]. LLM could change the educational learning goals, activities, and assessment and evaluation practices. The advantages of LLM in education include personalized learning, timely support, continuous assessment, resource delivery, collaboration, and so on [8]. One of the most used LLM by students for learning is currently ChatGPT (<https://chat.openai.com/>). ChatGPT is just one of many ways to use LLM. Research is progressing rapidly in this area, and new models are constantly being developed [6]. The release of ChatGPT as a chatbot based on GPT models by OpenAI marks a pivotal milestone in the development of chatbots and conversational AI [8]. By conversing with ChatGPT, students also try to study and learn Mathematics. There is a lot of research studying the math skills of ChatGPT (see, for example, [9, 10]). Not as much research studies ChatGPT's linguistic skills in Mathematics. LLM represent a significant step towards integrating natural language understanding and generation in information technology and artificial intelligence, with applications that extend to multiple sectors and scenarios (mathematical, chemical, legal, etc.). However, ChatGPT does not always manipulate the specialized language of these disciplines. LLM, as well as Google, can be used for linguistic investigations [11]. However, it must be remembered that there is no control over the corpus data on which they carry out searches, the results obtained from the searches, and the language they use. Control over the data is important, especially when teachers want to use these tools to design and deliver didactic activities.



This paper illustrates a project connecting DDL with a Digital Learning Environment (DLE) integrated with an Automatic Assessment System (AAS) to develop competencies in the language of Mathematics for Italian secondary school students [12]. The DLE is a stimulating environment that boosts motivation and fosters constructivist acquisition while respecting students' learning pace [13, 14]. In this research, The students conducted linguistic investigations on a specially designed corpus and carried out corpus-based activities with automatic formative assessment within a DLE. We define automatic formative assessment as using formative assessment in a DLE by automatically processing student responses and providing feedback [14, 15]. With the AAS, it is possible to create adaptive questions with personalized, immediate, and interactive feedback giving information about how the DDL task was performed and the process to be mastered, thus enabling self-regulation and self-monitoring of actions [16–18]. The DDL approach gives students the language access keys to the content and, in the case of Mathematics, proves its effectiveness in helping them to understand and manage a language that can pose actual obstacles to problem solving activities and exercises.

Section 2 presents the state of the art; Sect. 3 illustrates the research activity and methodology; and Sect. 4 shows the main results emerging from the research activity and discusses the implications and conclusions.

## 2 State of the Art

### 2.1 The Specialized Language of Mathematics and ChatGPT

Numerous studies in mathematics education have shown that the causes of difficulties are the acquisition, understanding, and management of its language [1]. Over time, Mathematics has developed a specialized language that has become increasingly universal, precise, concise, and effective. This language has its semiological code, which is realized in texts in which technical terms, figures and graphs, and symbolic expressions (equations, formulas, algebraic expressions, etc.) coexist; the latter are sometimes inserted in sentences which, for the rest, use the everyday language [19]. Mathematicians and researchers in mathematics education have different attitudes and points of view towards language, which derive from two apparently divergent aspects: the specificity of Mathematics and its language and the role of the context in mathematical communication [20]. Teaching is communication, and one of its purposes is to encourage student learning, using a language that does not hinder understanding and using metaphors to make concepts understood. At the same time, one of the main objectives is to acquire that specialized language.

Often, the linguistic aspects of teaching a language are considered only in a CLIL (Content and Language Integrated Learning) context, but teachers tend to forget that language education is a task that belongs to the training curriculum as a whole [5]. The characteristics of mathematical language often contrast with the students' linguistic habits, which are difficult to understand [2].

Secondary school students are often forced to deal with the languages of the disciplines without adequate linguistic scaffolding that allows them to make the necessary differentiations and categorizations of the language for specific purposes concerning

more general language varieties. Students' linguistic difficulties may concern comprehending or producing verbal texts, symbolic expressions, and figures. In all cases, it is not easy to recognize students' linguistic difficulties because they often compete with other difficulties related to gaps in content, careless mistakes, etc. Language difficulties can cause behaviours that can be identified, such as misinterpretations or the production of incoherent texts, but they can also cause no behaviours, such as not responding, or behaviours that are difficult to interpret, such as random answers [20].

In recent times, the use of LLM by students to learn Mathematics has spread considerably [8]. Consciously or not, this also affects students' learning of its specialized language. However, LLM, such as ChatGPT, do not always use specialized language. If we ask ChatGPT, "Do you know the specialized language of Mathematics?" a response is "Yes, I'm familiar with the specialized language of Mathematics. Mathematics has its technical vocabulary that includes specialized terms, symbols, and notations" [21]. For example, if we ask ChatGPT, "Can you define in 100 characters what "intersection points" are in Mathematics?" the answer is "points where lines or objects meet, sharing coordinates" [21]. The use of terms taken from a common language, such as "object" or "meet", can confuse students because, in a common language, these terms have different meanings. Furthermore, using the specialist verb "intersect" is missing, which students must acquire. Otherwise, students will have difficulty interpreting the meaning of the concept of intersection and tasks such as "at what point does the function intersect the ordinate axis?".

## 2.2 Data-Driven Learning

It is possible to use the DDL methodology to learn specialized languages [12, 22]. The main idea behind DDL is that learners can discover grammatical patterns, word meanings, or other aspects of language by searching for linguistic data. Corpora are proving increasingly influential in language teaching as sources of language descriptions [5]. A corpus is a collection of texts or parts of them in a finite number in an electronic format processed in a uniform way that makes it manageable and searchable using a computer [23]. Students have to deal with a "massive but controlled exposure to authentic input," which is fundamental for language learning [24], and such controlled and contextualized contact fosters more language awareness, noticing, and autonomy.

Tim Johns [4] argues that at the heart of the approach is the use of the machine as a rather special type of informant. Once the informant answers the question, students must make an effort to "make sense of that response and to integrate it with what is already known" [4]. Corpora provide data but do not interpret them: it is up to learners' work and responsibility to evaluate the information found. The easiest way to explore corpus data is through concordancers: end-users may display a list of words with their immediate context. A concordance based on KWIC (keyword in context) can reveal massive information about the language: idioms, collocations, fixed phrases, and frequency data. Since then, literature on the uses and benefits of corpora for language learning has rapidly grown, although there is still little field practice in Italy at least [5].

Recently, the number of corpora and specialised tools available to practising teachers and learners has increased. However, a common criticism is that many of them still require considerable investment in training for learners (and teachers) to understand

their rationale and how to use them effectively. Even accepting the potential benefits of a DDL approach, technology is perceived as a major barrier to the implementation of DDL in classrooms around the world [25].

Google itself can be used for linguistic investigations, but positions on the nature of the Web as a corpus are multiple and conflicting [5]. The exploration of web resources as corpus responds to the lack of corpora available to study increasingly complex specific linguistic problems. At the same time, materials age quickly compared to the continuous evolution of language, as well as new technologies and new means of online communication. Problems often cited against treating the Web as a corpus are its unknown size (no finite dimension), ever-changing composition, hidden pages, etc. Its advantages for language teaching include its size (lots of data), timeliness, variety (whatever you want is probably there somewhere), availability (free), reliability (the Web itself doesn't crash or impose limits on the number of simultaneous users), speed, flexibility, and so on. Importantly, it is already familiar to learners, especially via internet search engines such as Google [23, 26]. Another tool that has recently become very popular with students are LLMs, such as ChatGPT. Although ChatGPT is not designed for linguistic research, it does allow users to do so. Much recent research has explored the potential of ChatGPT for language teaching and learning (see, for example, [27, 28]), analysing the digital skills needed by teachers and students to use this chatbot ethically and effectively. Some research explores the role of ChatGPT in DDL, as the world's largest corpus (in a sense) is publicly available for free querying, with an interactive chatbot available to query this corpus quickly and recursively [29]. LLM can also be used for DDL to generate texts at the appropriate language level and to observe repetitions and variations manually or with corpus analysis tools. It is important to note that when conducting linguistic research in a specialised field, designing prompts to produce the desired results is particularly important. Students, for example, might assume that if they talk about "functions", ChatGPT understands that they are talking about math. However, if we ask ChatGPT, "What are the verbs that have "function" as the subject?", we get an ambiguous answer like: "Verbs that have "function" as the subject are typically used when describing how something operates or works; like functions, operates, works, performs, runs, executes, etc." [21]. If we ask, "What are the verbs that have "function" as the subject in Mathematics?" the answer is radically different (see Fig. 1). In this case, ChatGPT interpreted the task correctly, and the results refer exclusively to the mathematics field. As shown in Fig. 1, ChatGPT reports some verbs as a list, followed by an example sentence in which the verb is present. The presence of the example can help students understand in which context it is appropriate to use the verb, and they can reason about its meaning. However, as we have no control over the data, we do not know what determines the choice of verbs and the order in which they are presented. Furthermore, the concept of function is very transversal in Mathematics and is studied from lower secondary school to university. The results become more accurate if we add more details to the prompt.



In mathematics, "function" is often used as the subject of various verbs to describe its behavior and properties. Here are some common verbs used in mathematical contexts where "function" is the subject:

1. **Evaluates:** "The function evaluates to a specific value at  $x = 2$ ."
2. **Maps:** "The function maps elements from set A to set B."
3. **Transforms:** "The function transforms the input data into a different representation."
4. **Solves:** "The function solves for the roots of the equation."
5. **Graphs:** "The function graphs as a parabola."
6. **Derives:** "The function derives its values from the underlying data."
7. **Integrates:** "The function integrates over the interval  $[a, b]$ ."
8. **Differentiates:** "The function differentiates to find its rate of change."
9. **Computes:** "The function computes the result based on its input."
10. **Approximates:** "The function approximates the solution using a numerical method."

These verbs are commonly used in mathematics to describe how functions operate, the relationships they represent, and the calculations or transformations they perform.

**Fig. 1.** Example of mathematical linguistic investigation with ChatGPT.

### 3 Research Activities

The research activity involved four classes of two Italian secondary schools for 80 students in grade 11 and their teachers. Both schools are science lyceums, secondary schools with a curriculum focused on STEM subjects. It is also not the first experience of the research group in merging STEM assessment and languages [30]. The activities with the students involved two classes as experimental groups, while the other two classes comprised the control group. The students in the control group took a test and a questionnaire before the research activity started and a test and a questionnaire at the end of the research activity. These were the only activities that the control group performed. They did not do any classroom activities with the researchers; they only did normal lessons with their teachers. The experimental group participated also in the research activities compared to the control group.

The didactic experimentation, from November to December 2021, consisted of four two-hour meetings in the classroom. Teachers were involved in selecting topics to be covered and designing the teaching materials to ensure that they were in line with the programme. Teachers were not trained before the start of the experiment, and trainers from the University of Turin conducted the classroom activities. The teachers participated in all the meetings with the students and carried out the activities to study the proposed teaching methodologies. Two specific corpora were created for the activities, selecting texts relating to the concept of function from secondary school textbooks. The first corpus was in Italian, and the second was in English. Corpora were created and validated by

the researchers, who verified that the results of the linguistic research were relevant to the planned research activity.

Corpus-based activities with automatic formative assessment are questions implemented within the DLE using the AAS [12]. From a formative point of view, the students have several attempts to answer each question. The questions are designed to guide students in consulting the corpus and to make them reflect on the mathematical concepts. To carry out the activities, students use the AntConc concordance tool (<https://www.lauranceanthony.net/software/antconc/>) and the DLE (available at the link <https://linguaggi.specialistici.i-learn.unito.it>) side by side. Figure 2 shows an example of a corpus-based activity with automatic formative assessment. The question is divided into two parts: in the first part, students are asked to search for the word “funzion\*” (which means function) in the corpus by inserting 1R, 2R, and 3R as concordance levels and to select the verbs that have “function” as subject; in the second part of the question the students had to complete the proposed statements by inserting the verbs identified in the previous section.

Cerca la parola “funzion\*” inserendo come livelli di concordanza 1R,2R e 3R. Quali verbi hanno funzione come soggetto? Una funzione..

dichiara       interseca

determina       associa

assume       incontra

combina       esprime

Sezione Tentativo 1 di 3

Verifica

Completa le seguenti affermazioni utilizzando i verbi individuati nella sezione precedente:

- Una funzione  agli elementi del dominio un solo valore nel codominio
- La funzione  il valore 3 in  $x=2$
- Una funzione  la relazione tra due grandezze

Sezione Tentativo 1 di 3

Verifica

**Fig. 2.** Example of corpus-based activity.

Students see the second part after answering the first. At the end of each part of the question, there is a “verifica” button (which means “verify”). Students could click this button after answering the question to obtain immediate feedback on the correctness of the answer. If they made a mistake, they could try to answer the question again. In case of an incorrect answer after finishing the attempts, the students see the correct answer that is needed to carry out the next part. To answer the first part of the question, students had to use the AntConc software and enter the word “funzion\*” in the search bar. Inserting the asterisk symbol is used to search for both the word function in the

singular (“funzione”) and the plural (“funzioni”). Figure 3 shows an example of what appears on the screen to students after the search. The levels of concordance specified in the text of the question are placed under the search bar and must be set by the students. Through the levels of concordance, one chooses which words to highlight in the search. In this case, by selecting the levels “1R”, “2R”, and “3R”, the search shows all the phrases of the corpus in which the word function appears, and the following words are highlighted: the words one position to the right concerning the searched word (in red), the words two positions to the right concerning the searched word (in green), and the words three positions to the right concerning the searched word (in purple). In step-by-step processes with multiple response attempts, students earn partial credits for the correctness of their answers. These points act as a motivational lever and, by expressing intermediate levels between “incorrect” and “correct”, also offer teachers and students more precise information about the student’s competence in a particular domain.

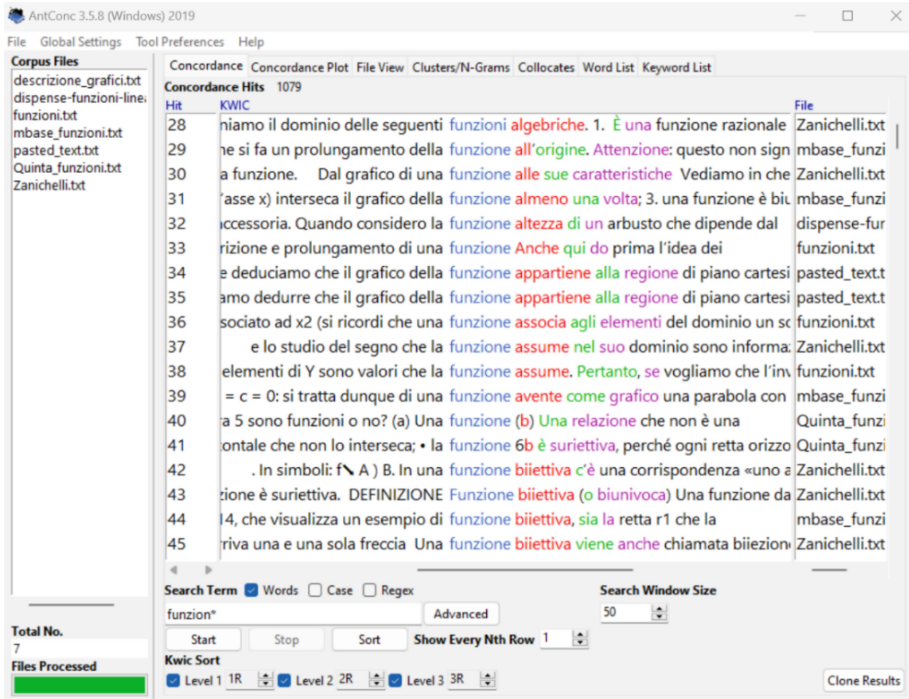


Fig. 3. Example of linguistic investigation with AntConc.

## 4 Methodology

The research methodology of the didactic experimentation included:

- Initial test (before starting the research activity) and final test (at the end of the research activity) to evaluate students’ mathematical and linguistic skills;

- Initial questionnaire (before starting the research activity) and final questionnaire (at the end of the research activity) to evaluate students' opinions about mathematics and mathematics lessons.

All 80 students took an initial test and questionnaire before starting the research activity. The test consisted of ten questions and included linguistic questions and mathematical exercises. The questionnaire was composed of 40 Likert-scale questions (on a scale from "1 = very disagree" to "4 = very agree") covering opinions on Mathematics, mathematics lessons, the study of Mathematics for school, school and extracurricular activities in Mathematics, and students' attitudes towards Mathematics. At the end of the experimentation, all students took a final test structured like the initial test and a final questionnaire. The questions in the final questionnaire were the same as in the initial questionnaire. Questions were added to the questions of the students who carried out the classroom activities to analyze the satisfaction of the proposed activities and methodologies.

Students in both groups could choose whether or not to participate in the educational experiment and change their minds at any time. In any case, their decision did not impact their school evaluations and the educational activities proposed to them. The data were collected through learning tests and paper questionnaires, tests with automatic assessment, and digital questionnaires. Once collected, the data were irreversibly anonymised and processed anonymously. The data were digitised, and the paper copies were destroyed. Qualitative and quantitative analyses can be carried out on the anonymised data in order to pursue the research objectives. Only aggregate statistics will be presented.

The research questions are:

- (RQ1) Did students value the methods proposed?
- (RQ2) Did the research activities increase awareness of mathematical language and mathematics?
- (RQ3) Was the impact of the research different for males and females?

Only the responses to the final survey of 35 students who participated in the experimental activity were considered to answer the research questions.

Questions of the final questionnaire concerning the satisfaction with the proposed activities and methods were analysed. In this question, the students had to indicate how much they agreed with various statements on a Likert scale from "1 = strongly disagree" to "4 = strongly agree". The items concerned:

- The satisfaction with the proposed activities (attention in class, satisfaction, interest in the activities and in Mathematics, appreciation of group activities);
- The usefulness of the activities for the acquisition of awareness of mathematical language and reflection on mathematical content;
- Appreciation of DDL (use of AntConc, interest in text analysis and usefulness for the future);
- Appreciation of automatic formative assessment (immediate and interactive feedback, multiple attempts possible);
- Understanding and appreciation of mathematical topics.



These questions were analysed by reporting the mean and standard deviation. To analyse the correlations between the items, we calculated the Pearson correlation coefficient between all possible pairs of items. To search for useful insights on gender studies, we also investigated the presence of differences in the responses of females and males to the items and the statistical significance of those differences through a two-tailed t-test.

A comparison between the results of the initial and final tests is planned to quantitatively assess the methods' impact, but the analyses are still ongoing.

## 5 Results

The data that we will analyze consists of the answers to the final survey of 35 students who participated in the experimental activity. The remaining students among the 80 students we considered were part of the control group. Among the 35 students, there are 20 females and 15 males. Students appreciated the proposed methodologies and the mathematical activities, which were very different from the traditional ones. The use of the DDL approach had a very positive and motivating impact on the students: according to them, the activities were practical and not theoretical, interactive and non-transmissive, and based on real data. Table 1 presents an analysis of the student's responses to the question: "Thinking about the experimental sessions, on a scale from '1 = strongly disagree' to '4 = strongly agree', to what extent do you agree or disagree with the following statements?".

The students found the activities interesting (3.23) and understandable (3.43). The activities stimulated their interest in Mathematics (2.60) and raised students' awareness of language (3.14). They appreciated the group discussions and the group activity (3.34). According to students, it was easy to use AntConc (3.09) and to understand how to use AntConc to answer questions (3.14). They also believe that learning to use text analysis software is useful for the future (3.14). According to students, the exercises helped them to become more aware of the language to use (3.14), and the exercises on language also made them reflect on meanings and contents (2.77).

Students appreciated the formative aspect of the activities: the possibility of viewing the correct answer immediately after answering a question (3.63) and retrying the exercise in case of a wrong answer (3.40). They agreed that having the assessment immediately after solving an exercise helped them understand how the question should be answered (3.51). The DDL approach gives students the language access keys to the content and, in the case of Mathematics, proves its effectiveness in helping them understand and manage its language that can pose obstacles to problem solving activities and exercises. Students exposed to DDL engaged in an activity that improved their language skills in their actual working practice. The study of language is concerned not only with the choice of words to use in a more specialized language but also with the terms conveying meanings, and by exploring the language, one explores the meanings. The new and interdisciplinary nature of the activity greatly impressed the students and their teachers, who, for the first time, discovered these methodologies and carried out linguistics activities. These activities allow students to better understand mathematical topics and to be more aware of the importance of using correct language to understand and be understood correctly.



**Table 1.** Mean and standard deviations of the 35 respondents to the final questionnaire

Item	Mean	Std. dev.
I paid attention during the activities	3.51	0.56
The proposed activities were interesting	3.23	0.60
The proposed activities were understandable	3.43	0.61
The proposed activities stimulated my interest in mathematics	2.60	0.60
The proposed activities were useful for gaining awareness of the language	3.14	0.65
I enjoyed the group activities	3.34	0.84
It was easy to use AntConc	3.09	0.85
It was easy to understand how to use AntConc to answer questions	3.14	0.85
It was interesting to learn how to use AntConc	3.40	0.74
Learning to use text analysis software is useful for the future	3.14	0.91
The exercises helped me to better understand the topics covered	3.06	0.80
The exercises made me appreciate the topics covered more	2.63	1.00
The exercises helped me to realize my preparation	3.03	0.71
It is helpful to view the correct answer immediately after answering a question	3.63	0.55
When I gave a wrong answer, I would try the exercise again	3.40	0.69
Having the assessment immediately after solving an exercise helped me understand how to answer the question	3.51	0.66
The exercises helped me to be more aware of the language to use	3.14	0.77
The exercises on language also made me reflect on meanings and contents	2.77	0.77

We also looked for correlations among items and calculated the Pearson Correlation Coefficient among all the possible pairs of items. We did not detect any strong correlation among items, and the correlation coefficient was not higher than 0.75. However, we found some interesting weak correlations among the following pairs of items.

Couple #1 is “The proposed activities stimulated my interest in mathematics” and “The exercises helped me to better understand the topics covered”, correlation coefficient 0.65, p-value 0.00002. This correlation shows that interest and understanding grow together; engaging and motivating students can lead to a better understanding of the topics.

Couple #2 is “The exercises helped me to better understand the topics covered.” and “The exercises made me appreciate the topics covered more.” correlation coefficient 0.65, p-value 0.00002. This couple shows that understanding and appreciation are related. However, we can think of this relation mainly in one direction: the more students understand, the more they appreciate what they studied. This is of great importance for

teachers who design activities for their students, as there is a need to focus on the understandability of the activity. Moreover, the immediate and interactive feedback that helped students while performing the activity played an important role in this correlation.

Two couples, couple #3, “It is helpful to view the correct answer immediately after answering a question.” and “When I gave a wrong answer, I would try the exercise again.” and couple #4 “When I gave a wrong answer, I would try the exercise again.” and “Having the assessment immediately after solving an exercise helped me understand how to answer the question.”, both with correlation coefficient 0.63 and p-value 0.00005, concern the formative aspect of the activity: on one side if the student have the chance to retry, they are willing to do it, and this is how a learning environment should be, a place where students can try without the fear of mistakes, on the other side, when having multiple attempts, students understand better how to deal with the activity, viewing the whole process and then going back from the beginning to start a good performance.

As a last analysis, to search for useful insights on gender studies, we also investigated the presence of differences in the responses of females and males to the items of the final questionnaire, together with the statistical significance of those differences through a two-tailed t-test. The results are presented in Table 2. In general, there are no peculiar differences among the two components of the sample. The highest difference ( $-0.67$ ) with the highest significance (0.022) appears in the item “It was easy to use AntConc”. In this specific case, the males signalled a larger agreement on this item. Conversely, the females signalled greater agreement on “The proposed activities stimulated my interest in mathematics” with difference 0.47 and significance 0.026. The other two elements with a p-value lower than 0.05 are represented by “The proposed activities were useful for gaining awareness of the language” (0.039) and “The exercises helped me to better understand the topics covered.” (0.047).

**Table 2.** Mean values of females’ and males’ responses with a t-test significance

Item	Mean (females)	Mean (males)	Difference (females-males)	T-test (p-value)
I paid attention during the activities	3.65	3.33	0.32	0.113
The proposed activities were interesting	3.35	3.07	0.28	0.147
The proposed activities were understandable	3.45	3.40	0.05	0.815
The proposed activities stimulated my interest in mathematics	2.80	2.33	0.47	0.026
The proposed activities were useful for gaining awareness of the language	3.35	2.87	0.48	0.039
I enjoyed the group activities	3.50	3.13	0.37	0.206

(continued)

**Table 2.** (continued)

Item	Mean (females)	Mean (males)	Difference (females-males)	T-test (p-value)
It was easy to use AntConc	2.80	3.47	-0.67	0.022
It was easy to understand how to use AntConc to answer questions	2.95	3.40	-0.45	0.111
It was interesting to learn how to use AntConc	3.45	3.33	0.12	0.658
Learning to use text analysis software is useful for the future	3.05	3.27	-0.22	0.491
The exercises helped me to better understand the topics covered	3.30	2.73	0.57	0.047
The exercises made me appreciate the topics covered more	2.85	2.33	0.52	0.157
The exercises helped me to realize my preparation	3.20	2.80	0.40	0.111
It is helpful to view the correct answer immediately after answering a question	3.70	3.53	0.17	0.403
When I gave a wrong answer, I would try the exercise again	3.55	3.20	0.35	0.159
Having the assessment immediately after solving an exercise helped me understand how to answer the question	3.60	3.40	0.20	0.398
The exercises helped me to be more aware of the language to use	3.25	3.00	0.25	0.364
The exercises on language also made me reflect on meanings and contents	3.00	2.47	0.53	0.050

## 6 Conclusions

In this paper, we present a didactic experiment characterised by the DDL and automatic formative assessment methodologies to study the language of mathematics. The results of the final questionnaire show that the students appreciated the mathematical activities, which were very different from the traditional ones. The use of technology had a very positive and motivating impact on the students.

The students highly appreciated the proposed methods to answer the first research question (RQ1). The DDL approach had a very positive and motivating effect on the students. According to the students, learning how to do linguistic analysis and useful for the future was interesting. They had no difficulty using AntConc and understood the linguistic investigations needed for corpus-based activities with automatic formative assessment. The students also appreciated the automatic formative assessment methodology. They appreciated the possibility of getting immediate feedback on the correctness

of each answer, correcting themselves in case of a wrong answer by having multiple attempts and knowing the correct answer at the end of each activity. All these aspects are closely related.

About the second research question (RQ2), the activities helped students become more aware of the language used. Not all students agree that the proposed activities can increase their interest in mathematics, but they agree that it is useful to do this kind of activity to reflect on the language used. The results show a correlation between students' understanding of mathematical topics and their appreciation of the topics themselves. These couple of items show that understanding and appreciation are related, and this relationship is very important when planning educational activities. The correlation between understanding the proposed activities and understanding the mathematical topics was another interesting result in this respect.

There are no particular differences between males and females (RQ3). The most significant differences concerned the use of AntConc and the impact of the proposed activities on the students' understanding of mathematics and its language. According to the results, males found it easier to learn to use AntConc. The activities further stimulated females' interest in mathematics and their awareness of mathematical language. Furthermore, the activities helped females more to increase their awareness of the proposed topics.

Students acquire language and digital skills using the tools and techniques of corpus linguistics for pedagogical purposes. The use of the digital learning environment and an automatic assessment system enables the generation and collection of data on student learning: about learning processes and not just about results. These data are useful for driving and adjusting the learning path, making choices and decisions, and supporting learning in several ways. It would be important to train teachers on using these methodologies and technologies in the future. In particular, for Mathematics teachers, it could be very interesting and formative to learn the analysis of specialized texts through the consultation of a corpus and the design of DDL activities with formative assessment to make students reflect on the specialized language of Mathematics. It is important to promote reflection among all teachers of all levels on the importance of acquiring specialized languages and how linguistic difficulties can influence disciplinary difficulties.

This research activity, characterised by DDL methods integrated with automatic formative assessment systems, could also be carried out on a larger scale, involving many teachers and students. The main challenge for mathematics teachers might be to reflect on the linguistic aspects. For this reason, training teachers in the proposed methodologies and to help them acquire linguistic skills would be essential. A second important aspect is to provide teachers with ready-made materials (corpus and corpus-based teaching activities with automatic formative evaluation). Indeed, one of the biggest obstacles could be the corpus creation and the consequent selection of representative texts. If teachers wanted to cover a topic other than the concept of function, they would have to create a new corpus and new activities based on this corpus. The situation is similar if one wants to apply this approach to other subjects, such as chemistry or biology, which have specialised language. For the future development of this research activity, the training of mathematics teachers is fundamental, possibly integrated with the training of language teachers and creating databases of ready-made activities.

Using technologies and a student-centred DDL approach allows students to be involved in linguistic tasks, especially in a STEM subject to which they are not accustomed. The exploration of corpora can effectively support reflection on the specialized languages of Mathematics. Many technologies used daily by students support linguistic investigations, but it is important to question the types of data on which they are carried out. Students use the most recent LLM and Google daily, and these tools can be used for linguistic investigations. However, there is no control over the data on which searches are carried out, the results of the searches, and the type of language they use.

Given the considerable diffusion of AI tools, training teachers and students on their strengths and weaknesses and how they influence teaching and learning is important. Students can use these tools critically and profitably even for less standard tasks such as linguistic research in Mathematics if appropriately trained. LLM, such as ChatGPT, could not be used for the same type of tasks since designing tasks with automatic formative assessment to guide students in linguistic investigations requires knowing the correct answer. ChatGPT answers may differ from student to student and may change over time. However, it would be interesting to use ChatGPT, through appropriately designed activities, to make students reflect on the language of Mathematics as a starting point for linguistic and disciplinary reflections. It is important to note that Generative AI has strengths in advancing our understanding of language in se that corpora have struggled to address. By combining corpus and Generative AI approaches, language learners can better understand how language works in different contexts than is currently possible with either approach alone. DDL researchers are well placed to take advantage of this renewed mainstream interest in language data, as we understand both the power of such data for language teaching and the conditions necessary for meaningful learning to take place with such data.

AI tools are changing mathematics education just as they will change the way of researching mathematics education. It is important to reflect on how these tools can influence students' language, both specialist and non-specialist, and their linguistic and not just disciplinary skills.

**Acknowledgments.** The research was carried out within Indam - Istituto Nazionale di Alta Matematica "Francesco Severi". The authors thank all the students who contributed to the research by responding to the questionnaire, and to all the teachers who showed passion for their work and expressed interest in improving their teaching.

## References

1. Ferrari, P.L.: Costruzione di competenze linguistiche appropriate per la matematica a partire dalla media inferiore. *L'insegnamento della matematica e delle scienze integrate* **26A**(4), 469–496 (2003)
2. Sbaragli, S., Demartini, S., Franchini, E.: Le difficoltà di comprensione e di gestione dei termini specialistici della geometria all'ingresso della scuola secondaria di primo grado. *La matematica e la sua didattica* **29**(1), 7–37 (2021)
3. OECD: How did countries perform in PISA? In *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing, Paris (2023). <https://doi.org/10.1787/9149c2f5-en>

4. Johns, T.: Should you be persuaded: two examples of data-driven learnin. Classroom concordancing. *ELR J.* 1–12 (1991)
5. Corino, E., Onesti, C.: Data-driven learning: a scaffolding methodology for CLIL and LSP teaching and learning. *Front. Educ.* **4**, 7 (2019)
6. Attardi, G.: Il Bello, il Brutto e il Cattivo dei LLM. *Mondo Digitale* **22**, 101 (2023)
7. Zhai, X.: ChatGPT user experience: implications for education. SSRN 4312418, 1–10 (2022)
8. Abedi, M., Alshybani, I., Shahadat, M. R. B., Murillo, M.: Beyond traditional teaching: the potential of large language models and chatbots in graduate engineering education. [arXiv:2309.13059](https://arxiv.org/abs/2309.13059) (2023)
9. Frieder, S., Pinchetti, L., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P.C., Berner, J.: Mathematical capabilities of ChatGPT. [arXiv:2301.13867](https://arxiv.org/abs/2301.13867) (2023)
10. Shakarian, P., Koyyalamudi, A., Ngu, N., Mareedu, L.: An independent evaluation of ChatGPT on mathematical word problems (MWP). [arXiv:2302.13814](https://arxiv.org/abs/2302.13814) (2023)
11. Corino, E.: Didattica delle lingue corpus-based. *Educazione Linguistica – Lang. Educ.* **8**, 231–248 (2014)
12. Corino, E., Fissore, C., Marchisio, M.: Data driven learning activities within a digital learning environment to study the specialized language of mathematics. In: O’Conner, L. (ed.) *IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 167–176 (2022)
13. Barana, A., Marchisio, M.: Analyzing interactions in automatic formative assessment activities for mathematics in digital learning environments. In: Csapó, B., Uhomoihi, J. (eds.) *13th International Conference on Computer Supported Education (CSEDU 2021)*, vol. 1, pp. 497–504 (2021)
14. Barana, A., Marchisio, M., Sacchet, M.: Interactive feedback for learning mathematics in a digital learning environment. *Educ. Sci.* **11**(6), 279 (2021)
15. Barana, A., Marchisio, M., Sacchet, M.: Effectiveness of automatic formative assessment for learning mathematics in higher education. In: Domenech J., Merello, P., De la Poza, E. (eds.) *7th International Conference on Higher Education Advances (HEAd 2021)*, pp. 1–8 (2021)
16. Barana, A., Fissore, C., Marchisio, M.: Automatic formative assessment strategies for the adaptive teaching of mathematics. In: Lane, H.C., Zvacek, S., Uhomoihi, J. (eds.) *Communications in Computer and Information Science, CCIS*, vol. 1473, pp. 341–365. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86439-2\\_18](https://doi.org/10.1007/978-3-030-86439-2_18)
17. Marello, C., Marchisio, M., Pulvirenti, M., Fissore, C.: Automatic assessment to enhance online dictionaries consultation skills. In: *16th International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2019*, pp. 331–338, (2019)
18. Corino, E., Fissore, C., Marchisio, M.: Adaptive exercises and formative assessment for English remedial action. In: Ifenthaler, D., Isafas, P., Sampson, D.G. (eds.) *Orchestration of Learning Environments in the Digital World*, pp. 3–19. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-90944-4\\_1](https://doi.org/10.1007/978-3-030-90944-4_1)
19. Sbaragli, S., Demartini, S.: *Lingua e strutture dei testi scolastici di matematica*. Italmatica. Dedalo Editor (2021)
20. Ferrari, P.L.: *Educazione matematica, lingua, linguaggi. Costruire, condividere e comunicare matematica in classe* (2021)
21. OpenAI, ChatGPT. <https://chat.openai.com>. Accessed 10 Dec 2023
22. Fissore, C.: Data driven learning activities in English on the specialized language of mathematics in secondary school. In: Chova, L.G., López Martínez, A., Lees, J. (eds.) *14th International Conference on Education and new Learning Technologies (EDULEARN)*, pp. 9326–9333 (2022)
23. Barbera, E.F.: *Linguistica dei corpora e linguistica dei corpora italiana. Un’introduzione*, pp. 1–113 (2013)

24. Cobb, T., Boulton, A.: Classroom applications of corpus analysis. *Cambridge Handbook of English Corpus Linguistics*, pp. 478–497 (2015)
25. Boulton, A.: Testing the limits of data-driven learning: language proficiency and training. *ReCALL* **21**(1), 37–51 (2009)
26. Boulton, A.: What data for data-driven learning? *EuroCALL Rev.* **20**(1), 23–27 (2012)
27. Kohnke, L., Moorhouse, B.L., Zou, D.: ChatGPT for language teaching and learning. *RELC J.* **54**(2), 537–550 (2023)
28. Baskara, R.: Exploring the implications of ChatGPT for language learning in higher education. *Indonesian J. Engl. Lang. Teach. Appl. Linguist.* **7**(2), 343–358 (2023)
29. Crosthwaite, P., Baisa, V.: Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Appl. Corpus Linguist.* **3**(3), 100066 (2023)
30. Marchisio, M., et al.: Adapting STEM automated assessment system to enhance language skills. In: *Proceedings of the eLearning and Software for Education Conference (eLSE 2019)*, pp. 403–410 (2019)





**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Interoperating Civil Registration of Death and Census Data: Old Age and Marriage as Categories of Analysis

Ciara Breathnach<sup>1</sup>  , Rachel Murphy<sup>2</sup> , Alexander Schieweck<sup>3</sup> ,  
and Tiziana Margaria<sup>3</sup> 

<sup>1</sup> School of History, University College Cork, Cork, Ireland  
Ciara.breathnach@ucc.ie

<sup>2</sup> School of History and Geography, University of Limerick, Limerick, Ireland

<sup>3</sup> HRI, Lero and CRT-AI, University of Limerick, Limerick, Ireland

**Abstract.** As part of the project Death and Burial Data: Ireland 1864–1922 (DBDIrl), a web application was created in DIME, a low-code web application development environment. DIME is based on the popular IDE Eclipse and utilizes three distinct graphical model types (data model, process model and GUI model) as Domain Specific Language (DSL). Web applications are defined by means of these models in the DIME integrated modelling environment in a simple way, and these models are used to generate the web application code that is then deployed and made available to the users. This paper provides an overview of how we use a web application created in DIME, a low-code application development tool to establish interoperability between Irish historical civil registration of death data and Irish cross-sectional decennial census data from 1901 and 1911. The former data collections are complex and require manual interventions like transcription and considerable cleaning to render them machine readable, the latter are fully transcribed. Here we present a case study concerning the analysis of deaths in old age in Dublin City and how they correlate with census returns. We then turn our attention to a use case of marital status and the algorithmic problems caused by irregular replies in census data. Our objective was to test how DIME could replicate the decisions taken by domain experts using old age and marital status as the primary linkage criteria.

**Keywords:** Data matching · civil registration · census · low-code/no-code application development · model driven development

## 1 Introduction

Old age offers a very fruitful case study from the perspectives of life course studies and historical demography. We examine this in the context of Ireland in the early twentieth century, which was subject to British administration from 1801 until 1922. At the close of the nineteenth century increased life expectancy caused a shift in demographic trends, which necessitated the introduction of new social welfare schemes like the old age



pension (OAP). The target cohort were lower-paid working-class people who did not earn enough during their working lives to pay into contributory pension schemes or to save money to support themselves in old age. Bismarck's Germany was the first to introduce a non-contributory pension in 1891 as part of a wave of social welfare reforms, and after extensive investigation and debate, the United Kingdom followed suit [1]. The parameters of old age and its positioning in historiography have been tentative as Johnson argues, and broadly distributed across the themes of labour market participation, welfare and social status [2]. Pat Thane contends that women fared better than men in reaching older age in western societies, but she cautions that the boundaries of older age must be rooted in various cultural constructs and historical contexts [3]. Lines were clearly drawn in the British Old Age Pensions Act of 1908, which stipulated that applicants had to prove that they were 70 years of age [4]. Apart from the chronological markers, applicants had to fulfil two further main criteria: 1) being resident in the United Kingdom of Great Britain and Ireland (as it then was) and a British subject for the previous 20 years, and 2) they had to pass a means test proving that their annual income did not exceed £31 10 shillings. Ó Gráda argues that, apart from its significant impact on national income, the effects of the OAP on household economics placed it as 'the most radical and far-reaching piece of welfare legislation enacted in Ireland in the twentieth century' [5]. From his assessment of post-Famine census reports Gilleard surmises that Ireland had the oldest population in Europe, which he attributes to 'cohort depletion from mass emigration' [1]. Dublin city, like Belfast city, had a more youthful population for two reasons: firstly, mortality rates in younger cohorts (especially infants and children under 5) were high and, secondly, it was where internal migrants gravitated. In rural areas of the West of Ireland traditions of seasonal and permanent migrations were firmly established and cohort depletion led to a larger older populace. Growing old in the city was a tough station and while all sorts of allegations were made about age heaping (rounding up to the nearest 5 or 0) in the 1911 census, it seems that in Dublin it was more difficult to get away with what Ó Gráda terms 'welfare fraud' as civil and religious records were more carefully maintained [5]. The impact of the OAP on Irish fiscal and social life cannot be understated. Family incomes were so low that a steady and reliable weekly flow of 5 shillings for one and 7s 6d for married couples would have been an inestimable boon to both sole-occupancy and multi-generational households at a time when a week's rental in Dublin could have cost up to 3 shillings a week [6].

Until the 1908 Act, 65 was used as the marker for old age for the purposes of the annual reports of the Registrar-General. In 1909, the year in which the old age pension was introduced, civil registration data shows that deaths from old age amounted to 8,870, which included those aged 65–69. Within the pre-established parameters of ages 65 to 74, 839 people died, with a further 5,184 aged over 75, 2,424 aged over 85 and 423 aged 95 and upwards [7]. Historical demographers are interested in understanding the impact of the OAP on these cohorts, and often use a process called data matching to track trends over time. Matching can occur between different record types, but census matching, in which individuals are located in the previous or subsequent census, is one of the most common forms of data matching in the U.S. through IPUMS [8] and in North-West Europe [9].

Apart from the algorithms used by IPUMS and Mandemaker et al. for census returns, there are recognized basic standards for data-matching across all data types. Christensen describes five steps involved in data matching: (i) data pre-processing, (ii) indexing, (iii) comparisons, (iv) record pair classification, and (v) evaluation. Historians and genealogists working with partially indexed data routinely must conduct this record-matching process manually [10].

This aspect of the project had two aims: (i) to take civil registration of death records and match them to census records using a bespoke web application designed in the DIME low-code application development environment<sup>1</sup> and (ii) to ascertain the extent to which it is possible to replicate the historian's thought processes using an algorithmic approach.

We had previously created a bespoke web application with this technology stack to transcribe scanned historical death data [11]. A logical next step was to extend the app to census data using this project as a pilot. While other data-matching solutions exist, they do not align with the full range of Irish records we are analysing both now and in the future. While aiming to replicate the manual process and supporting future work, the matching is limited to some fuzziness, e.g. in names or dates, but no full probabilistic matching at the moment. Also AI is not yet included, as for the moment we are collecting the kind of questions historians would pose, and evaluating the extent to which the approximate matchings we produce without AI are good solutions.

In the remainder of this paper we discuss how we set out to achieve these aims. In Sect. 2 we discuss the datasets we used; Sect. 3 describes the cohorts selected for the case study; in Sect. 4 we introduce the data models and elements of the application designed to conduct the data matching; Sect. 5 discusses the matching algorithm; Sect. 6 reports an analysis of our results, followed by our findings in Sect. 7. We draw our conclusions in Sect. 8.

## 2 Datasets

Some countries can match civil registration records and census data over a long period of time, e.g. England from 1837 for civil registration and census from 1841 to 1921. For the case study presented in this paper we match civil death records to the 1901 and 1911 censuses of Ireland, which are the only fully extant Irish censuses (see Sect. 2.2). Both datasets are described below. We have refined the metadata structure successively, as this structure and its granularity are at the core of the information on which linkage, interoperability and thus matches are based. We explained and analysed the original data structure in relation to civil registration of death data in [11]. Subsequently we refined the granularity further to arrive at the current, stable level of disaggregation described in [12]. As the census data have been transcribed and indexed in full by the National Archives of Ireland (NAI), it did not necessitate the same levels of attention.

### 2.1 Irish Civil Registration Records

Irish historical civil registration data (births, marriages and deaths) are openly available for individual searches on [irishgenealogy.ie](https://www.irishgenealogy.ie). When these datasets were first digitised, a

<sup>1</sup> DIME is an open-source platform available at <https://gitlab.com/scce/dime>.

simple index of a subset of information (place, name and date of death) was created and linked to a TIFF file of the original image. Accordingly, the data are not provided in full machine-readable formats nor is there any national infrastructure to support large scale crowdsourced transcription such as Pedersen et al. describe in their contribution to this volume [13]. A further issue is that ownership of each dataset comes under the remit of various government agencies: for example, Irish civil registration data falls under the auspices of the General Registration Office (GRO), whose primary function is to manage the live population register. The historical census comes under the auspices of the (NAI) and the more modern returns under the Central Statistics Office (CSO).

## 2.2 Irish Censuses

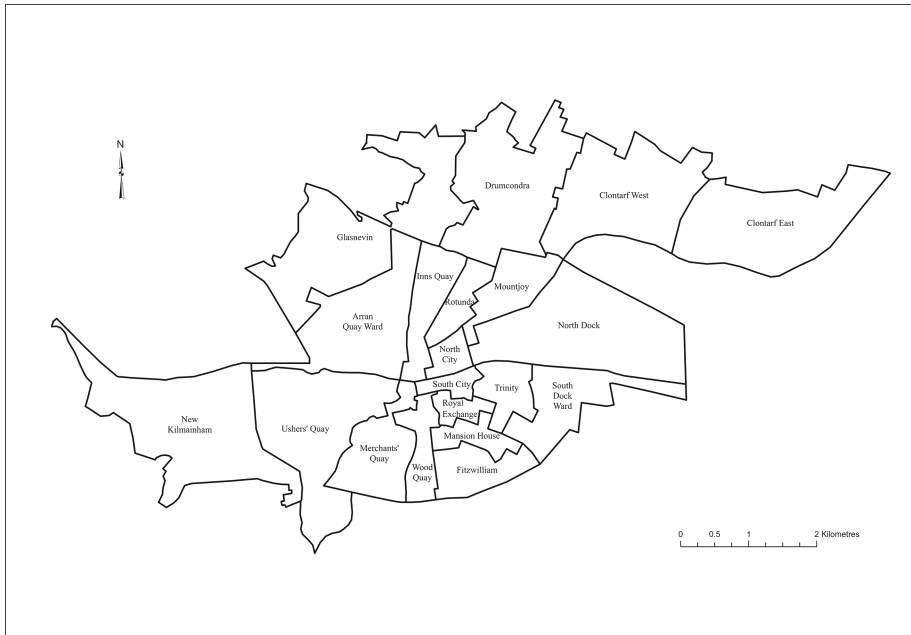
While Irish civil registration records are intact, when it comes to the census (which was taken decennially from 1821 to 1911) Ireland has the unfortunate disposition that some manuscript census returns were legally destroyed by order of the Deputy Keeper of the Public Records in the late nineteenth century, some were pulped for reasons of paper conservation during the First World War effort, and others were a casualty of the Civil War [14]. What survive for the entire country are the full manuscript returns for 1901 and 1911. Owing to the very disturbed state of the country in 1921, no census was held, as martial law prevailed in several counties during the War of Independence, and it was simply unsafe for the Royal Irish Constabulary or the Dublin Metropolitan Police to conduct the work. The next census, which took place in 1926, will be published in accordance with legislation in January 2027 [14].

## 3 Case Study: Old Age in Dublin South City No. 3 Ward

### 3.1 Death Data Used in the Study

We manually transcribed deaths registered in the population over 70 years old for one Dublin ward, South City No. 3 (Fig. 1) in 1911. This year was selected as it was the year of the first census taken following the OAP act. We conducted a data matching exercise with these civil registration of death records and the census returns. Working with the GRO death data provided us with insights into the relatively understudied cohort of older individuals. We took a random sample from this dataset to conduct a manual data matching exercise that in turn assisted our understanding of how best to devise a matching algorithm for this use case. A further aim for the historians was to add to the discourses on the questions of age misstatement in pension applications, which until now have relied on aggregate returns.

Dublin South City No. 3, located to the south of the river Liffey, was a predominantly working-class area, with many individuals working precariously as low-paid labourers or servants. Table 1 shows the recorded deaths of people aged 70+ in 1911 by 5-year cohorts. This data was derived from individual level civil registration of death registers of that year, which were fully transcribed to a machine-analyzable format by the first application we developed with DIME for the history research domain [12]. Analysis of this data shows that this age cohort made up 13% of total deaths in South City No. 3 in 1911.



**Fig. 1.** Map Showing Dublin Wards. (Dublin, North and South Wards, Dublin Metropolitan Police stations shown as points, c. 1900. Map based on information derived from the Dublin Metropolitan Police Committee of Inquiry 1883 [C.3576 C.3576-I], p. 25, and Statistical Tables of the Dublin Metropolitan Police for the Year 1901 [Cd. 1166], p. 33, drawn using OSi historic 25-inch basemap © Tailte Éireann/Government of Ireland Copyright Permit No. MP 003824).

**Table 1.** Deaths of Individuals aged 70+ in South City No. 3, 1911 (based on analysis of GRO death data, South City No. 3, 1911)

Age Cohort	Deaths	Of which Male	Of which Female
70–74	57	23	34
75–79	40	23	17
80–84	22	8	14
85–89	10	2	8
90–94	2	1	1
Total deaths 70+	131	57	74
All deaths in South City No. 3, 1911	998	512	486

As Table 1 shows, few deaths occurred in older age, which forms the basis of our use case. A total of 131 deaths were recorded in the 70 and over category and, in keeping with Pat Thane's [3] assertions, of these 74 (or 56%) were women. Age, conjugal status and socio-economic status are among the primary features of the profile of these pensioners.

The marital status of this group was as follows: 78 were widowed, 25 were returned as married, 15 were spinsters, 11 bachelors, and 2 unknown.

To perform the matching exercise, we had to first identify all those individuals who were alive on the night of the census, 2 April 1911. This meant removing 46 individuals from the total of 131, leaving a sample size of 84 individuals. The breakdown of these individuals is shown in Table 2.

**Table 2.** Deaths of Individuals aged 70+ in South City No. 3 registered 3 Apr.-31 Dec.1911

Age Cohort	Deaths	Of which Male	Of which Female
70–74	41	16	25
75–79	24	13	11
80–84	11	4	7
85–89	7	1	6
90–94	1	1	0
Total deaths 70+	84	35	49

## 4 The Census Application

### 4.1 Background

We follow here the same design methodology used in the initial positive experience with the DIME platform to design, implement and deploy the project application [12], which allowed users located anywhere in the world to transcribe death data records to a curated database at a fine granular level of 64 searchable and indexable fields, compared to the original 11 of the Victorian registers. The aim of using DIME in DBDIrl was to support the use of data from multiple siloed sources, such as the civil registration data and the census data, and to make them interoperable, this way enabling a greater matching potential. This project focuses on phase one of this: census matching.

The next phase of the project was to create a census application called CensusIRL, which enables the tracing of individuals through time using the census records. We decided to use DIME again. The DIME process, defined as a sketch through blueprints, matches entries between two censuses (see Fig. 2). A preliminary approach successfully tested the feasibility using DIME [15], yielding a fully modelled prototype of the application that served as proof of concept. The current, more advanced, census matching application has been tested so far on the two available Irish censuses, the 1901 and 1911 census, but it is built in such a way that in future it can be applied to the 1926 census as well, which is due to be released to the public in 2027.

### 4.2 The Extended Data Models

The NAI provided the project with the full census datasets which were ingested into the application. Each census data record contains the 23 data points detailed in Table 3. All

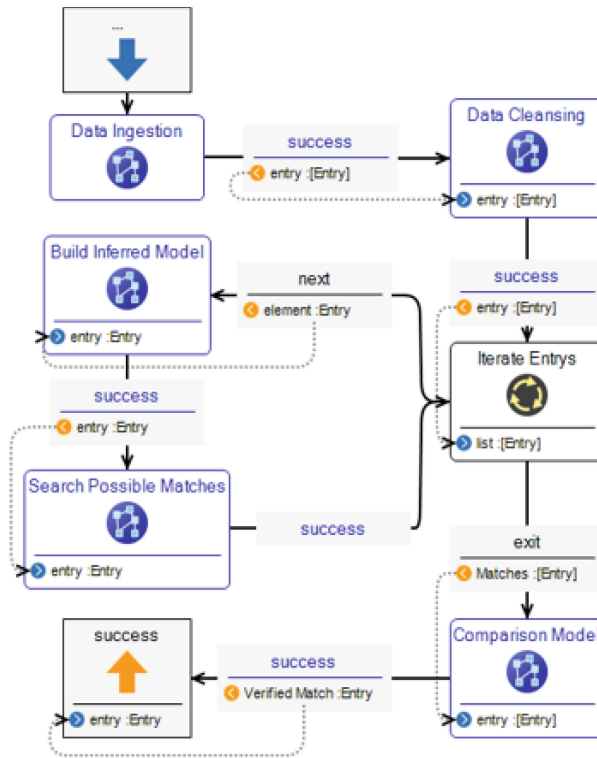


Fig. 2. Search process model sketched using DIME blueprints

census data is included in the scope of the project except crudely anonymised entries of people located in institutions such as workhouses, prisons or asylums. To avoid cases of reputational damage by association with criminality or the ‘taint of pauperism’ only initials were recorded in the original census returns.

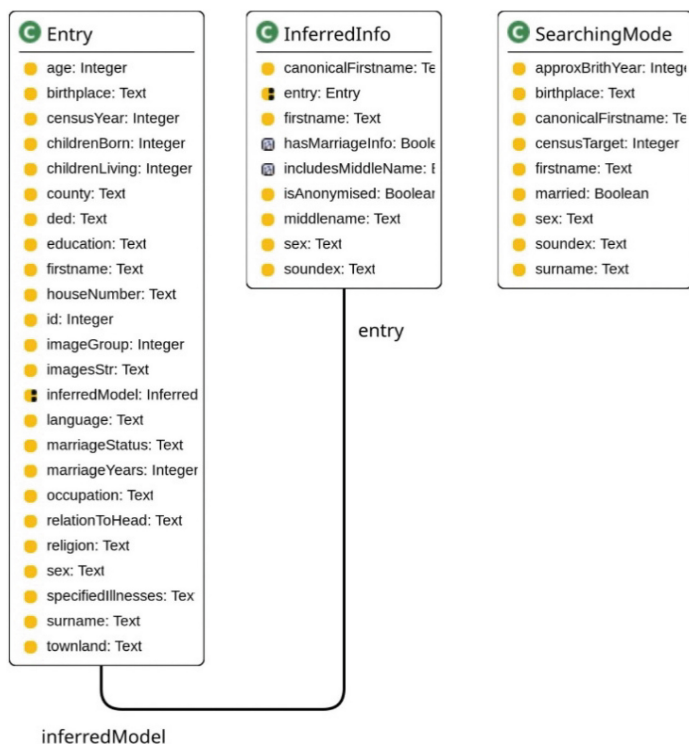
An inferred data model, shown in Fig. 3, was manually defined to enable subsequent data matching based on that extended data model. During pre-processing, some analyses were run on the original data to compute inferred information that is relevant to determine whether a given census record is relevant or not for some matching query. The Entry table (Fig. 3, left) describes the original data entry, while the derived properties are gathered in a separate InferredInfo table (Fig. 3, middle) to enrich the available set of attributes for subsequent searches. For example, this information captures whether one or more middle names are included under the ‘Forename’ field, or whether an individual’s name has been anonymised. This is relevant, as such entries should be excluded, preserving semantic correctness and completeness of the results, but saving computational effort, when trying to match by names. Here we also include Soundex, a phonetic algorithm that relies more on consonants than vowels to index words according to their phonetic pronunciation. Orthography is not always consistent between historical records, particularly where literacy rates varied, and vowel changes are common e.g. between Byrne, Burn and Bourne. In this case these three surnames are all signified by the Soundex code B650.

**Table 3.** Census Data Records Ingested into the application

1901 Census Household Return (Form A)	1911 Census Household Return (Form A)
1. Surname	
2. Forename	
3. Age	
4. Sex	
5. Relation to head	<i>As in 1901 but with three additional data points</i>
6. Religion	
7. Birthplace	
8. Occupation	
9. Literacy	
10. Irish language	13. Years married
11. Marital status	14. Children born
12. Specified illnesses	15. Children living
Enumerator's Abstract (Form N)	
16. County	
17. District Electoral Division (DED)	
18. Townland	
19. House number	
Additional Data	
20. Census year	
21. ID/Image group	
22. Images str	
23. Inferred model	

The *SearchingModel* table (Fig. 3, right) describes the current, fuzzy search and is used by the matching algorithm to find potential candidates in the census records. Some of the fields directly correspond to the original *Entry* elements, like *birthplace*, *firstname*, *surname* and *sex*, but many refer to derived information in the *InferredInfo*, like *canonicalFirstname*, *married* (that connects with the inferred *hasMarriageInfo* attribute), or exist only in the *InferredInfo* table like *Soundex*. Furthermore, only the *approximate-BirthYear* is known due way in which the original data is structured, so we limit the window of years used to approximate matches. The *censusTarget* can be used to selected a specific census, currently 1901 or 1911.

In the model we also see that the *Entry* and the *InferredInfo* tables have mutual pointers, so are connected in 1:1 fashion, but they are kept separate. There are two main reasons for this choice:



**Fig. 3.** The Data model used for the search: original Census Information (left), Inferred Information (middle) and the Searching Model (right) covering the fields that are currently used in the matching process.

- While the original census *Entry* structure is fixed, as the project progresses, we may wish to update and extend the nature and amount of *InferredInfo* made available for the searches.
- Additionally, it is curatorial best practice to keep the original data separate from any additional fields that may be useful for specific purposes but are not in the original format and dataset.

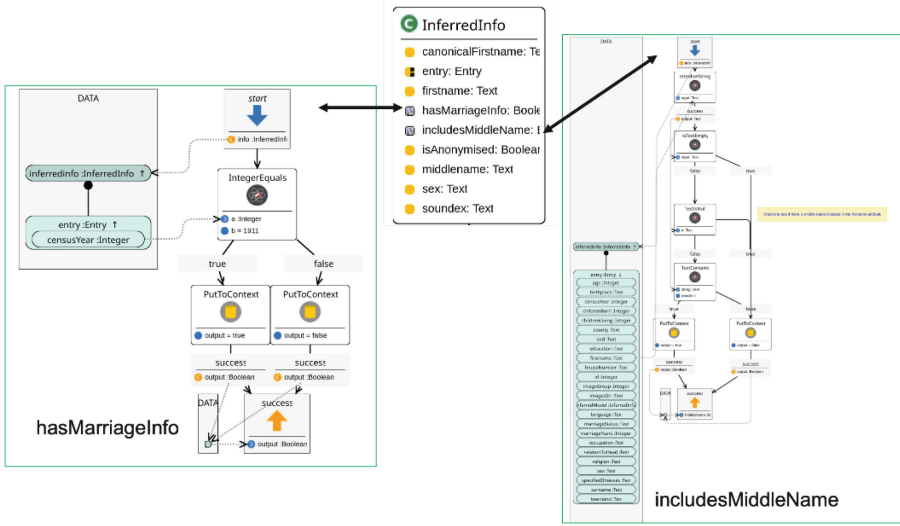
The data is stored in the extended format in a PostgreSQL database.

### 4.3 Embedded Processes for the Inferred Fields

Process models (see Fig. 4) describe the algorithms used to filter the census records in order to “infer” the additional information in the remaining two fields of the *InferredInfo* table: the *hasMarriageInfo* process (left) and the *includesMiddleName* process (right). While originally in the prototype these were fully modelled in DIME, the scalability of the queries has improved by just using database queries. As the data is stored in a standard PostgreSQL database, it is easy to formulate queries and embed them in rather simple DIME processes.



We are thus confident that this technique can be used again to provide in the future additional inferred fields, if needed. Given the similarity of these two processes, we may be able to provide a generic template process with the elements common to all the queries, that can be extended on demand just with a few Service Independent Building-blocks (SIBs) and data fields specific to the query in question.



**Fig. 4.** The embedded process models used to compute inferred information: the hasMarriageInfo process (left) and the includesMiddleName process (right).

## 5 The Matching Algorithm

### 5.1 Matching Properties

The 84 deaths that occurred after 2 April 1911 were matched in the application against the entire census dataset. The matching occurs on the following six properties:

- **First name.** For some individuals the first name may be consistent between censuses, but in many cases the first name may vary because individuals have nicknames. For instance, a woman named Margaret could also be referred to as Madge, Mag, Mags, Maggie, Margarita, Marguerita, Meg, Peg or Peggy, and her name might also be abbreviated to Mgt or Margt. The genealogy website Roots Ireland includes a list of common abbreviations for first names, and we have used the examples in the list to generalise names [16].
- **Last name.** While spelling was becoming more standardised during this period, there can be slight variations, therefore Soundex has been used to account for similar sounding names. Soundex is a phonetic algorithm that indexes names according to their sound. It encodes the first letter of a surname, and subsequent consonants, with

the aim of matching surnames that might have small inconsistencies in spellings. It is not a perfect system, but it has long been used by genealogy companies to underpin surname searches, to a reasonable degree of success. It is based on the English pronunciation of names, so it may have a lower degree of effectiveness for Irish surnames. As will be discussed later, in the case of women, surnames changed upon marriage and so this record field does not help to capture women who have married in the intervening period between the censuses.

- **Sex.** This information should be consistent across the censuses.
- **Approximate birth year.** The birth year should be consistent, but in the nineteenth and early twentieth centuries remembering birthdays was not prioritised as highly as it is now. While some families noted down birthdays in a family bible (probably more a Protestant tradition, or one used by Catholic emigrants) for many this was not information of great significance. Neither was an individual's age, until the coming of the Old Age pension. For these reasons, people may not have known their precise age. So, when matching birth years, we initially broaden the parameters to  $\pm 1$  year, and possibly extend the interval up to  $\pm 5$  years.
- **Birthplace.** This piece of information should be consistent across all records about an individual. Potential issues include inconsistency in spelling of placenames, particularly if the individual concerned was illiterate and relying on another to write the name on their behalf. It is also possible that in rural areas the name of a farm might be given in one census, and a townland in another. Issues such as changing/multiple placenames occur as well; many Irish placenames were Anglicised by the Ordnance Survey for instance.
- **Marital status.** In most cases, following marriage an individual's marital status moves from single to one of the following: married/widow/widower. It would have been highly unusual for a once-married person to describe themselves as single, as divorce was the preserve of the wealthier classes at that time, and even couples who were estranged or those who were deserted would generally maintain their status as married. For instance, returned migrant Mary O'Connor described herself as 'married in America' [17]. However, there are a few examples of individuals who described their status as 'divorced' such as Alfred Manning a 59-year-old retired merchant [18].

## 6 Results

### 6.1 Matches

Table 4 explains how many individuals received a given number of matches for each of the censuses. One individual received 139 matches in the 1901 census, while another received 125 in the 1911 census. Three quarters of individuals had a match in at least one census, but 1 in 4 individuals could not be matched in either census.

**Table 4.** No. Matches per individual in each census

No Matches	1901 Census	1911 Census
No matches	22	25
1–2	17	17
3–10	22	18
11–25	13	12
26–50	7	6
51–100	3	7
101+	3	1

## 6.2 Individuals with One or Two Matches

First, we considered a random sample of all individuals with one or two matches. Using traditional research techniques by the historians we looked at whether we could identify the correct individual in both censuses, and if so, whether this match was correctly identified by the application. The results are shown in Table 5: in all these cases it was possible to find the individual in the other census. In most cases the difficulty arose due to differences in the spelling of names, though in two cases the ages were very different.

**Table 5.** Individuals with one or two matches

Name and age	Located in Census	Correct Match?	Historian's comments
Louisa Malcomson (71)	1901	Yes	Yes, same address in death record and census [19]
William Claffey (72)	1901	No	Incorrectly matched to William Ashley. He was traced to the 1901 census in which he stated he was 59. This data was used to trace him forward to 1911 where he provides an age of 66 – in this case a margin greater than 5 years needed to be applied [20]
Henry Giblin (74)	1911	Yes	1 Bride Road (address on census) was part of Iveagh Buildings (death certificate). There is a potential match in 1901 but this would need to be verified [21]

(continued)

**Table 5.** (continued)

Name and age	Located in Census	Correct Match?	Historian's comments
Charlotte Holmes (76)	1911 only	No	On the death record Charlotte Holmes is living at 29 William Place. A search locates a 'Charlotte Homes' aged 91 living at 29 William's Place. The informant on death certificate is Eileen Beadfield, likely a relative of Martha Bradfield, one of the people Charlotte was living with in the 1911 census. There is a possible match, Charlotte Rachel Holmes in Brooke Street, Enniskillen in 1901 [22]
Hans Murphy (78)	1901 only	Yes	Murphy is also found in the 1911 census; the 1901 census gives his address as Guinness Trust and his death record states Iveagh Buildings, the same location. He was married in both 1901 and at the time of his death so we searched on his wife's name 'Harriet Murphy' and found Hans living there at this address – but his name was written as 'Hands' [23]
James Riordan (82)	1901 only	Yes	In 1901 he was an army pensioner living in Cork, aged 73. In 1911 he cannot be located at his Upper Beechwood Avenue address but was found in the private hospital in which he died under the name James 'Reardon' [24]

### 6.3 Individuals Who Could not Be Matched in the App

Next, we took a random sample of five individuals who could not be matched to see if the historian colleagues were able to match them manually. Historians and genealogists start with what is known and work backwards, so we started with the death record, and then traced individuals to the 1911 census first, followed by the 1901. To emulate the application, we did not look at any other records to assist us in the matching process. In reality, if we could not locate an individual, we would cast the net more widely among available sources to build up a more detailed profile. Our findings will inform future iterations of the app (Table 6).

**Table 6.** Manual matching of individuals who could not be matched in the app

Name, age, occupation marital status and address	Located in Census	Historian's comments
Isaac Hearney (71), Coach Builder, widower St Albans Road	Yes, 1911 and 1901	We were able to match this individual to the census return of Isaac P. Kearney, a 73-year-old coachmaker born in Limerick and living at 19 St Alban's Road. In 1901 he was living at Greenville Terrace, Dublin [25]
Jane Elizabeth Roche (75), widow of a gentleman, 15 Earlsfort Terrace	Yes 1911 and 1901	Jane Elizabeth Redington Roche aged 75 was located at 15 Earlsfort Terrace. The issue in identification is down to the fact that her surname was indexed under Redington Roche. Using this surname, she was also easily located at Adelaide Road in the 1901 census [26]
David John Fitzgerald (76), Grocer, 14 Emor Street	Yes, 1911 and 1901	David Fitzgerald was easily located in the 1911 census, but his surname was written FitzGerald. A retired grocer's assistant aged 76, he was born in Wexford and lived at Emor Street, so a positive identification. The David Fitzgerald located in Royal Canal Bank in the 1901 census is likely to be the same person. He was born in Wexford, aged 60 and a 'packer' (mistranscribed as 'paller') [27]
Leibe Schein (83), widowed schoolteacher, 11 Walworth Road	Yes, 1911 and 1901	In the 1911 census Louis Shein aged 82 is a boarder in the Jaffey household (10 Walworth Road). While no occupation is given this is likely to be the correct person. The first name does not correspond, but members of the Jewish community often Anglicised their names so Russian-born Louis is likely to be Leibe. This is confirmed in the 1901 census when Russian-born Lewis Shine of 10.1 Walworth Road, is described as a teacher in Hebrew born in Russia [28]
Mary Townsend Saunders (85), Lady, Spinster Rest for the Dying	Yes, 1911 but not 1901	In the 1911 census Mary T Saunders aged 85 was located living in 8 Ulverton Road, Dalkey, Dublin. She was an annuitant (living on a sum of money she received each year) and single, born in Co. Cork. All this information matches her death record perfectly [29]. We searched the 1901 census for a Mary Saunders but were not able to locate her

#### 6.4 Individuals with a High Number of Matches in the App

Finally, we reviewed a random sample of 5 individuals with a high number of matches to show how we might be able to narrow down the search results in the application. Typically, when doing this kind of research historians start narrow and broaden their search when no positive findings are forthcoming (Table 7).

**Table 7.** Manual matching of individuals who could not be matched in the app

Name, age, occupation marital status and address	Were they located by the algorithm?	Historian's comments
William Brown 70, married, clerk, Meath Hospital/69 Brighton Square	No	Located at 69 Brighton Square under the surname Browne. Age 68 (2 years less than death record so not located in search) and married. Occupation is 'commission agent, retired' rather than clerk [30]
John Obrien 70, married, army pensioner, 5 Mclean's Lane	?	No John O'Brien at 5 McLean's Lane in the census. Death cert shows his wife was Mary. Possible candidate is John O'Brien of 86 Crumlin, Terenure, whose wife was Mary, but further research required [31]
Margaret Byrne, 73, widow of coach driver, 1 Earlsfort Terrace	Yes, 1911	Margaret was located at 1.2 Earlsfort Terrace, aged 73. Spelling of name, age and address are all precise matches to death record. She was not located in the 1901 census [32]
Michael Byrne, 71, widowed labourer, 36 Upper Kevin Street	No	Michael Byrne located at 36.3 Upper Kevin Street in the 1911 census aged 77, widower, labourer. The algorithm did not identify him due to too great an age discrepancy. He did not live at this address in 1901; Michael Byrne of 73.4 Francis Street is a possible match, but more research required [33]
Mary Breen, 75, widow of horse dealer, 12 Golden Lane	Not in 1911 but in 1901	Mary Breen, aged 78, was located at 11.2 Golden Lane. Age is 3 years above that in death certificate, and street name is identical (number different). She was located at 11.1 Golden Lane in 1901 aged 65. This is exactly 10 years younger than the age in the death certificate which is why she was matched in 1901 [34]

## 7 Discussion

This exercise allowed us to identify several challenges with the data matching:

**Dataset Size.** While we only used a small sample of data for this trial, the subsequent attempt to run the matching application with a larger data set of ca. 8 million records (the full 1901 and 1911 census) pushed our low-code technologies to their limits, and we needed to use custom database code.

**Anonymised Entries.** Of the full dataset, over 155,000 records relate to anonymised entries. It will be very challenging to be able to match these records which only utilise initials of first name and last name.

**Birth Year.** Some individuals were highly accurate in recording their birth year, but others much less reliable. For some individuals this is because knowing their own age or date of birth was not as important as it is now, but for others this might be due to age heaping, where dates are rounded. In other cases, this imprecision or differences may have been intentional (particularly after the old age pension had been introduced). A search of plus or minus one year is too limited a time window – a range of ten years either side is likely to capture most eventualities.

**Inconsistent Responses.** Even where specific responses were required, some individuals still provided an alternative response. For instance, though there were 4 options for marital status ('married', 'widower', 'widowed' and 'single'), across the entire census dataset we found 227 distinct status descriptions, which included non-standardised responses such as 'no', 'still single', 'unmarried (unfortunately)', 'on the lookout' and 'not yet'.

**Name Changes After Marriage.** This is a major issue when tracing women. However, it could potentially be solved by also incorporating the marriage certificates into the data pool.

**Transcription Inaccuracies.** Even though marital status should be relatively easy to transcribe, some issues were identified. For instance, in the case of Edward Flynn, Cookstown his marital status is mis-transcribed as 'less than one year' yet in the original census return he stated 'married' [35]. This additional information should have been transcribed into the 'years married' data field.

**Record Sets Containing Different Information.** All these records were designed for highly specific purposes, and being able to link them was not a factor in their original design. Thinking of future applications for data matching, one issue will be identifying common information between all these records. For instance, the birthplace is very useful in census matching, but it is missing from the marriage and death records.

We also identified some areas that could be further enhanced. The historians noted that the address can be a very useful property to include in searches, particularly where a death occurs one or two years after a census. In the cases where the historians checked the data manually, using an address to match allowed for a positive identification. If the address does not match, it makes sense to search first in close proximity (e.g. within the same District Electoral Division), and only then broaden the search to a county, or Ireland as a whole. This gradual spatial approach should lead to more accurate results.

Inconsistent spellings of names posed a problem as shown in the tables above: e.g. Hans was modified to Hands, Riordan to Reardon. A system other than Soundex might be used, or we could develop a solution that handles the specifics of Irish names. Finally, in many cases the standard plus or minus 5 years window for age searches was not broad enough. A plus or minus 10 years would work better in some instances. We also noted that the highest number of matches appeared with common surnames, but that using Soundex many similar surnames were returned, e.g. Brien, Breen and Brown. These are distinct surnames in terms of pronunciation, so Soundex may not be the best solution.

## 8 Conclusions

While the initial findings are promising, more refinement is required for the application to work well. Rather than having a single standardised set of search criteria it may work better to create a decision tree to better emulate the flexibility of the knowledge discovery mode in which historians would progressively work their way through the manual matching process. This is the case because at the start of any search historians look at all the information available to them and decide which starting point (name plus address, birthplace, birthdate or even occupation) seems most likely to give the right results. In the future, we hope to expand this process within the application using some kind of AI or machine learning technique, which could help to identify similarities in the entries, which have not been obvious to the human eye. Such innovation will deepen understandings of the extent of ‘welfare fraud’ [5] in early twentieth-century Ireland and will add to the growing body of work in age-cohort studies using individual level data [36].

**Acknowledgements.** “Death and Burial Data: Ireland 1864–1922” is a project funded by an Irish Research Council Laureate Award IRCLA/2017/32. This publication has emanated from research supported in part by a grant from SFI R@ISE SPP under grant number 21/SPP/9979.

## References

1. Gilleard, C.: *Old Age in Nineteenth-Century Ireland*. Palgrave Pivot, London (2017)
2. Johnson, P.: Historical readings of old age and ageing. In: Johnson, P., Thane, P. (eds.) *Old Age from Antiquity to Post-Modernity: From Antiquity to Postmodernity*, pp. 2–3. Routledge, London (1988)
3. Thane, P.: Social histories of old age and aging. *J. Soc. Hist.* **37**(1), 93–111 (2003)
4. *Old Age Pensions Act 1908*, 8 Edw. 7. c. 40
5. Gráda, C.Ó.: ‘The greatest blessing of all’: the old age pension in Ireland. *Past & Present* **175**, 124–161 (2002)
6. Breathnach, C., Murphy, R.: Fine lines: locating commercial sex work in official data, Dublin 1901 and 1911. *Urban History* **51**, 435–456 (2023). <https://doi.org/10.1017/S096392682300010X>
7. *Annual Report of the Registrar-General*, p. 137 (1909)
8. Helgertz, J., Price, J., Wellington, J., Thompson, K.J., Ruggles, S., Fitch, C.A.: A new strategy for linking U.S. historical censuses: a case study for the IPUMS multigenerational longitudinal panel. *Hist. Methods* **55**(1), 12–29 (2022)



9. Mandemakers, K., Alter, G., Vézina, H., Puschmann, P.: *Sowing: The Construction of Historical Longitudinal Population Databases*. Radboud University Press, Nijmegen (2023)
10. Christen, P.: The data matching process. In: *Data Matching. Data-Centric Systems and Applications*. Springer Berlin, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-31164-2\\_2](https://doi.org/10.1007/978-3-642-31164-2_2)
11. Breathnach, C., Ibrahim, N.M., Clancy, S., Margaria, T.: Towards model checking product lines in the digital humanities: an application to historical data. In: ter Beek, M.H., Fantechi, A., Semini, L. (eds.) *From Software Engineering to Formal Methods and Tools, and Back*, pp. 338–364. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-30985-5\\_20](https://doi.org/10.1007/978-3-030-30985-5_20)
12. Schieweck, A., Murphy, R., Khan, R., Breathnach, C., Margaria, T.: Evolution of the historian data entry application: supporting transcriptions in the digital humanities through MDD. In: *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, Los Alamitos, CA, USA, pp. 177–186 (2022). <https://doi.org/10.1109/COMPSAC54236.2022.00033>
13. Pedersen et al (in this volume)
14. History of Irish Census Records. <https://www.census.nationalarchives.ie/help/history.html>
15. Doherty, A.J., Murphy, R.A., Schieweck, A., Clancy, S., Breathnach, C., Margaria, T.: CensusIRL: historical census data preparation with MDD support. In: *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, pp. 2507–2514 (2022). <https://doi.org/10.1109/BigData55660.2022.10021106>
16. Roots Ireland First Names. <https://www.rootsireland.ie/help/first-names>
17. Mary O'Connor. <http://tinyurl.com/Mary-O-Connor-1911>
18. Alfred Manning. <http://tinyurl.com/Alfred-Manning-1911>
19. Louisa Malcomson. <http://tinyurl.com/Louisa-Malcomson-1911> and <http://tinyurl.com/Louisa-Malcomson-1901>
20. William Claffey. <http://tinyurl.com/William-Claffey-1911> and <http://tinyurl.com/William-Claffey-1901>
21. Henry Giblin. <http://tinyurl.com/Henry-Giblin-1911> and <http://tinyurl.com/Henry-Giblin-1901>
22. Charlotte Holmes. <http://tinyurl.com/Charlotte-Holmes-1911> and <http://tinyurl.com/Charlotte-Rachel-Holmes-1901>
23. Hans Murphy. <http://tinyurl.com/Hans-Murphy-1911> and <http://tinyurl.com/Hans-Murphy-1901>
24. James Riordan. <http://tinyurl.com/James-Riordan-1911> and <http://tinyurl.com/James-Riordan-1901>
25. Isaac P Kearney. <http://tinyurl.com/Isaac-P-Kearney-1911> and <http://tinyurl.com/Isaac-P-Kearney-1901>
26. Jane Elizabeth Redington Roche. <http://tinyurl.com/Jane-Redington-Roche-1911> and <http://tinyurl.com/Jane-Redington-Roche-1901>
27. David John Fitzgerald. <http://tinyurl.com/David-John-Fitzgerald-1911> and <http://tinyurl.com/David-John-Fitzgerald-1901>
28. Leibe Schein. <http://tinyurl.com/Leibe-Schein-1911> and <http://tinyurl.com/Leibe-Schein-1901>
29. Mary Townsend Saunders. <http://tinyurl.com/Mary-Townsend-Saunders-1911>
30. William Brown. <http://tinyurl.com/William-Brown-1911>
31. John O'Brien. <http://tinyurl.com/John-O-Brien-1911>
32. Margaret Byrne. <http://tinyurl.com/Margaret-Byrne-1911>
33. Michael Byrne. <http://tinyurl.com/Michael-Byrne-1911> and <http://tinyurl.com/Michael-Byrne-1901>
34. Mary Breen. <http://tinyurl.com/Mary-Breen-1911> and <http://tinyurl.com/Mary-Breen-1901>

35. Edward Flynn. <http://tinyurl.com/Edward-Flynn-1911>
36. Breathnach, C.: Parish pump politics and the introduction of Old Age Pensions in Ireland. Accepted and forthcoming History Workshop Journal







**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# From Data Science to Modular Workflows Changing Perspectives from Data to Platform: DBDlrl 1864-1922 Case Study

Enda O'Shea<sup>1,2</sup> , Marco Krumrey<sup>1,2,3</sup> , Daniel Sami Mitwalli<sup>1,2,3</sup> ,  
Sebastian Teumert<sup>1,2,3</sup> , and Tiziana Margaria<sup>1,2,3,4</sup>  

<sup>1</sup> University of Limerick, Limerick, Ireland

{enda.oshea,krumrey.marco,mitwalli.daniel,teumert.sebastian,  
tiziana.margaria}@ul.ie

<sup>2</sup> CRT-AI: SFI Centre for Research Training in Artificial Intelligence, Cork, Ireland

<sup>3</sup> Lero: The Irish Software Research Centre, Limerick, Ireland

<sup>4</sup> HRI: The Health Research Institute, Limerick, Ireland

**Abstract.** Many historical data collections foot on handwritten documents and registers, whose consultation is often very difficult due to the conservation state of the physical artefacts, and whose comprehension is also made difficult by the handwriting, difficult to interpret, and the language used, different from the modern terminology. Therefore significant research efforts by historians, demographers, population health scientists and others have been started in the past with the aim of making such data collections digitally available, first on the basis of images and then as readily available repositories of transcribed data in electronically queryable formats. For the purpose of extracting data from the Irish Civil registers of deaths in the DBDlrl 1864-1922 project (<https://www.dbdirl.com>), an AI-ML Data Analytics Pipeline was proposed as a working approach validated on a subset of the data. However, the pipeline requires manual steps and it is not applicable as is on similar datasets without significant modifications to its inner workings.

We are currently transforming this prototyped, single purpose product to a modular, fully automated workflow, intended to be used and reconfigured for new data in a low-code/no-code fashion by domain experts like historians. We explain our adopted analysis and refactoring process, illustrate it on part of the pipeline, including how we faced obstacles and handled pitfalls. We also evaluate its potential to become a methodical approach to transforming an interactive program to a fully automated process, in a low-code/no-code workflow style, that can be easily reused, reconfigured and extended to be able to tailor it to other datasets as needed.

**Keywords:** Historical data · Data science · Model driven development · Low-code/No-code · Digital Thread · DIME

# 1 Motivation

Numerous factors can shape a programmer’s coding style. In contemporary times, programmers enter the field with diverse educational backgrounds and experiences. According to a survey conducted by Stack Overflow, approximately 80% of professional programmers in 2021 have successfully obtained a bachelor’s degree, leaving one in five without this qualification<sup>1</sup>. Concentrating on educational background, around 38.41% of programmers who held at least a bachelor’s degree between 2018 and 2021 were predominantly from the field of “Computer and Information Science and Support Services”<sup>2</sup>. Thus, these students are often exposed to traditional programming paradigms such as object oriented programming and well established, state-of-art programming patterns. Another large share of code, however, originates from non-CS STEM fields such as engineering, mathematics, statistics, biology/medicine and business related studies. These programmers generally have been taught the bare minimum programming skills to achieve or satisfy their tasks, mostly use interpreted or scripting languages, and they operate on a paradigm where code is written down instead of properly designed, and it is seen as “disposable” instead of a product. Other factors that influence a programmer’s coding style are their level of experience, the company or organizational culture at which they have been employed, and the associated programming language(s) they used.

Further important factors that influence coding style are the purpose of the project and the role that the code plays within it. In a traditional software project, the code is used to produce a running program which is ultimately “*the product*”. In data science projects, however, scientific methods and algorithms are used to extract knowledge and insights from (possibly semi- and un-)structured data collections. In this context, the code is not primarily the product itself, but it serves as a mere tool to derive “*the product*”, which is in this case information: the output of running the code.

Being perceived as an ancillary artefact, the code involved tends to be written pragmatically and used as a short-lived and temporary tool, e.g. to explore and experiment with the data. This approach is more akin to a rapid prototyping mindset, where the code is kept pragmatic and not really curated, and its purpose is more of presenting ideas, as a proof of concept, and produce quick results. In some cases, such projects move from exploration to larger production or collaboration, where a clean and consistent style guide and well-documented code become increasingly important: that’s where this prototyping style shows deficiencies, and a different mindset is required to take over.

In the DBDlrl 1864-1922 project<sup>3</sup>, described in more detail in Sect. 2, there was a core need to investigate the feasibility of automating the information extraction from the Irish civil registers of deaths, provided as a large collection

---

<sup>1</sup> <https://insights.stackoverflow.com/survey/2021#education-ed-level-prof>.

<sup>2</sup> <https://datausa.io/profile/soc/computer-programmers?employment-bl-geo=statesEBL&majors-select=commonMajors>.

<sup>3</sup> <https://www.dbdirl.com>.

of TIFF images page by page. This need gave rise to a “proof of concept”-style project that investigated the AI-supported extraction of information from the images and its conversion into usable data for statistical analysis. The project leaders were aware from the beginning that the medium-term goal was to produce a reusable and repurposable platform, but the starting point was from zero, with only one skilled computer scientist joining the team at a later point, with the personal aim of introducing AI and ML technologies and demonstrating that “yes, we can”. The modules developed in this project are designed as interactive programs that allow the user (so far a skilled CS expert who knows AI, ML and the data very well) to follow the data processing process step by step. The results were excellent, but not transferable to other users nor datasets.

In the context of the just started GREATLEAP EU COST Action<sup>4</sup>, however, the goal is to analyze and compare death data from historical civil registries of over 15 countries in Europe, and link them to other data collections and sources of information. In [1] and more recently in [2] we address the linkage problem of the death data with the 1901 and 1911 Irish census data. Additionally, [3] addresses coding historical causes of death data with Large Language Models and [4] introduces and discusses a common language for accessibility, interoperability, and reusability in Historical Demography. In this large scale European effort, Working Group 3 addresses the creation of analytical tools, specifically computational and visualization tools for the application to new research questions. It is clear that the process of extracting information and providing usable data is to be automated, making it easily accessible to users that are historians, and repurposable, e.g. in order to handle other collections of images of datasets with “register”-like appearance. This means that the “product” shifts from the *derivation of data* and *understanding the process*, to a possibly parameterized *automated software pipeline* that effectively embodies rather complex workflows (i.e., it is neither linear nor simple). It must be able to efficiently execute these modular sub-steps and transform the provided information into usable data, covering a range of inputs and outputs.

In the following, Sect. 2 provides some background on the DBDIrl 1864-1922 project and on the general class of questions such research elicits. Section 3 describes the AI-ML data analytics pipeline and its structure, then Sect. 4 presents the issues encountered that need resolution for a full automation. Section 5 describes the transformation needed to move to a reusable, modular application that is part of a model driven platform. Section 6 provides a comparison of the mindsets behind the two different styles of programming, on the basis of the example. Lastly, Sect. 7 summarizes and concludes the work.

## 2 DBDIrl Project Background

The “Death and Burial Data: Ireland 1864-1922” project (DBDIrl, 2017-23)<sup>5</sup> delved into the historical civil registration (CR) documents of Ireland to reveal

<sup>4</sup> See <https://greatleap.eu>.

<sup>5</sup> <https://www.dbdirl.com>.

and analyze the biopower dynamics across this time period. From 1864 to 1922 the Irish General Register Office (GRO), operating under the authority of the British government in Ireland, systematically recorded vital life events such as births, deaths, and marriages. By thoroughly examining these records, the project aimed to reveal the dynamics surrounding life and death during this significant period in Irish history, with a particular emphasis on the conclusion of life paths in a nation marked by high emigration.

## 2.1 The Dataset

The General Register Office in Ireland facilitated the provision of CR death records from this period to the DBDIrl team, encompassing approximately 4.3 million distinct deceased individuals. These records consist of scans, as shown in Fig. 1, serving as digital reproductions of the original writings made by the registrar at the time of death, provided in both PDF and Tagged Image File Format (TIFF).

Following the partial digitization of birth, marriage, and death records by the GRO department, the accessibility and searchability of these records have been greatly enhanced on the official Irish Genealogy website<sup>6</sup>. Users can now explore these records using criteria such as the individual's name, location, and the time-frame of their life event. While this vast collection proves to be an invaluable resource, the digitization efforts extended to death records within the DBDIrl project demanded a more detailed approach. These records not only capture essential information about the deceased individual but also delve into details about the informant (responsible for validating the deceased's identity) and the registrar involved in the registration process. This heightened level of detail allows for a more profound understanding of the context and intricacies surrounding each documented death event. Analyzing the scanned record depicted in Fig. 1, each document can contain up to 10 individual death records, each featuring 11 columns of data. Through a collaborative effort between the DBDIrl team and computer scientists, these 11 fields were initially expanded to 22 data categories, as detailed in [5]. They were subsequently expanded to 63 categories [6] that reflect the level of granularity essential in the transcription processes for each distinct subject.

---

<sup>6</sup> <https://www.irishgenealogy.ie/en/>.

Superintendent Registrar's District		Registrar's District										
1881 DEATHS Registered in the District of		in the Union of										
		in the County of										
No. (1.)	Date and Place of Death (2.)	Name and Surname (3.)	Sex (4.)	Condition (5.)	Age last Birthday (6.)	Rank, Profession, or Occupation (7.)	Certified Cause of Death, and Duration of Illness (8.)	Signature, Qualification, and Residence of Informant (9.)	When Registered (10.)	Signature of Registrar (11.)		
468	1881 Sixth September Claneool	Edward Cottonel	M.	Bachelor	18 years	Son of Edward Cottonel, owner Mineral Water Manufactory	Diphtheria 11 Days Certified	Ellen X Cottonel Mother Present at Death Claneool	Five October 1881	Charles Cooper Assistant Registrar.		
469	1881 Sixth September Work House	Richard Mahane	M.	Bachelor	74 years	Labourer	Heart Disease Mitral Valve 2 years Certified	Abram Hayes Occupier Randon Workhouse	Sixth October 1881	Charles Cooper Assistant Registrar.		
470	1881 Eighth October Ridge Street	Ellen Walsh	F.	Married	55 years	Wife of Mr. Michael Walsh Publican	Childbirth 12 Days Certified	Michael Walsh Husband Present at Death Bridge Street	Eight October 1881	Charles Cooper Assistant Registrar.		
471	1881 Fifth October Kilbragan Street	Honora Pinehorn	F.	Married	24 years	Daughter of Mrs Madony	Phthisis 1 year	Honora O'Sullivan Householder Present at Death Kilbragan St.	Six October 1881	Charles Cooper Assistant Registrar.		
472	1881 Eleventh October Convent Hill	Ellen Ryan	F.	Spinster	27 years	Daughter of Timothy Ryan	Mitral Regurgitation 12 years Certified	Kate Ryan Sister Present at Death Convent Hill	Fifteen October 1881	Charles Cooper Assistant Registrar.		
473	1881 Sixteenth October Claneool	Hannah McCarthy	F.	Married	52 years	Wife of Denis McCarthy Farmer	Childbirth 12 Days Certified	Dennis X McCarthy Husband Present at Death Claneool	Seventeenth October 1881	Charles Cooper Assistant Registrar.		
474	1881 Fifteenth October Work House	Bridget Driscoll	F.	Widow	65 years	Pauper	Diphtheria 4 days Certified	Abram Hayes Occupier Randon Workhouse	Twenty October 1881	Charles Cooper Assistant Registrar.		
475	1881 Eighteenth October Work House	William Murphy	M.	Bachelor	63 years	Labourer	Phthisis Pulmonary 8 months Certified	Abram Hayes Occupier Randon Workhouse	Twenty October 1881	Charles Cooper Assistant Registrar.		
476	1881 Nineteenth October Ridge Lane	Wrenish Fitzpatrick	M.	Bachelor	5 years	Son of Wm. Fitzpatrick Shoemaker	Scarlet Fever 25 Days Certified	Wm X Fitzpatrick Mother Present at Death Ridge Lane	Twenty October 1881	W. Cooper. Assistant Registrar.		
477	1881 Twentieth October Inverbragan	Timothy Hawley	M.	Widower	78 years	Labourer	Septic Pustul Abscess 65 Days Certified	Timothy X Hawley Mother Present at Death Inverbragan	Second November 1881	J. Walsh Registrar.		

Fig. 1. Example of Scanned CR Death Record from 1881

### 2.2 Historical Record Transcription Approaches

Manual transcription has been, and remains, a predominant approach for transcribing vast volumes of handwritten historical documents. Strategies such as crowd-sourcing [7,8], transcription workshops [9], and employing expert transcription houses [10] are the most prevalent methods for annotating large quan-



tities of data. These approaches enable researchers to maintain control over the environments while ensuring a certain level of quality assurance. However, a clear challenge arises as the time and resources available to different researchers may vary.

Looking at automated transcription systems in the study by Thorvaldsen et al. [7], the researchers utilized two distinct versions to handle both historical Norwegian census (1891) and Barcelona marriage records (1451 to 1905), with each system catering to different layouts and data type representations. The Norwegian census data were presented in a uniform, structured document, enabling the utilization of anchor regions for layout analysis and the isolation of word regions based on these anchors. This process was executed using open-source tools and libraries, resulting in the identification of 99.3% of underlined fields within the document and the correct classification of 96.7% of these underlines. Subsequently, they segmented data regions into individual images, achieved up to 70% accuracy in classifying single digits, and generated name clouds by gender to aid future manual transcription endeavors.

Conversely, the Barcelona marriage records were recorded as lines of sequential handwritten text, necessitating additional segmentation compared to the Norwegian census data. The team developed a three-tiered, web-based, crowd-sourced platform, gaining access to over 150 transcribers who annotated two years' worth of data and provided essential ground truth data for image analysis. Following this, a handwriting recognition system was integrated to support the transcribers during the manual transcription phase. The handwriting recognition module places a focus on the segmentation of text blocks into individual paragraphs using probability maps to identify textual features. Subsequently, these paragraphs are further segmented into singular lines through the application of graph-based continual path measures.

Thorvaldsen subsequently extended this research [11] by utilizing 353,000 commercially annotated images. These images were employed to train a deep learning network with the goal of classifying first names listed in the Norwegian 1950 census, yielding the ten most probable results for each name classification. Focusing on the Norwegian 1950 census, [12] furthered this research by creating the 'Ocode' end-to-end machine transcription pipeline. This pipeline was designed to transcribe 2.3 million handwritten three-digit occupation codes embedded within the census data. To capture the occupation codes, an ad hoc segmentation program was applied, and multiple processing methods for the codes were tested. The sequence model demonstrated outstanding performance, achieving an automated transcription accuracy of 97%. The residual 3% of the data was subjected to manual transcription, underscoring once again the necessity for a hybrid approach in handling such handwritten data.

In automated transcription pipelines, such as those outlined above, a key recurring theme is the necessity to segment various types of documents into smaller, more detailed data units, such as table cells or individual words in a text line. This segmentation is crucial for the effective processing of these documents by language models. Only after this segmentation can tasks like word



spotting, word classification, or character-level classification be effectively undertaken. Additionally, the integration of lexicons is a common practice in these pipelines, as they considerably narrow down the search space for word classification, playing a vital role in achieving desirable results.

This approach is exemplified in [13–15], and [16], where vertical and horizontal projection filters, skew adjustments followed by 2D mask generation, and the use of anchor points from OCR transcription software are utilized to identify cell regions. More modern methods that incorporate deep learning, including semantic and instance segmentation, are used in [17, 18], and [19]. These sophisticated strategies emphasize the automated extraction of features followed by the identification and extraction of crucial regions of interest. Subsequently, classification networks and techniques can be applied to these extracted elements.

### 2.3 Manual Transcription in DBDIrl

In the early phases of the DBDIrl project, the data underwent manual transcription by domain experts, specifically members of the DBDIrl team. Despite being a painstaking and time-intensive procedure, it yielded high-quality results, ensuring the accuracy needed to conduct quantitative research. Challenges emerged, however, when attempting to expand this approach to include non-domain experts, such as students, and special interest groups. These new stakeholders necessitated extensive training to accurately extract and transcribe documents with the required level of granularity, adhering to strict guidelines to steer their progress. These guidelines were established through the development of a dedicated Web application, the Historian DIME App (HDA) [9], featuring individual pages that guide the transcription of records based on a page for the information provided by each role (registrar, informant, deceased) involved in the registration of that death. The software itself was developed using the eXtreme Model Driven Development approach (XMDD) [20]. The transcription covered all 63 data elements, equipped with predefined correctness checks (expressed by syntactic and semantic constraints) that ensured correctness and plausibility of the input data. Furthermore, a set of classifiers was developed [6] to identify errors in previously transcribed data, initially captured in Excel sheets. The classifiers analyzed the data record by record and field by field, providing feedback on each entry as well as aggregated percentages of accurately transcribed sets for the fields pertaining to each of the three roles.

### 2.4 Automating the Transcription Through an AI/ML Pipeline

While the manual transcription process demonstrated a commendable level of precision, the sheer volume of over 4 million individual death records warranted a move towards automation. Given the monumental scale of processing, involving nearly 60 years’ worth of data, each comprising approximately 12,000 scanned documents, it became imperative to adopt a systematic pipeline approach. This method entailed a series of sequential steps: initially, the documents were meticulously narrowed down to table regions containing death records, then further

refined to individual row and column regions. Subsequently, the focus shifted to the identification and classification of individual words within these refined regions.

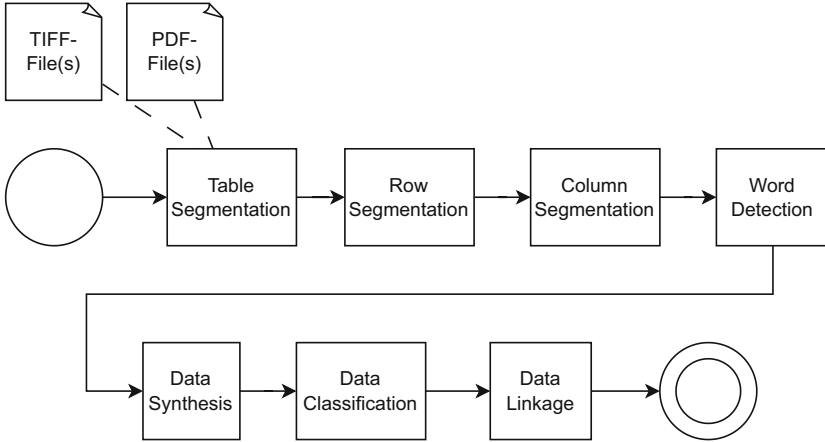
The decision to segment the documents only at the word level was influenced by the handwritten cursive nature of the texts, which posed significant challenges for character segmentation. Moreover, the lexicon within multiple column regions was limited to specific domains, simplifying the word-level classification process. To streamline the workflow, the pipeline was constructed as a series of distinct modules, each dedicated to a specific aspect of the transcription process. Each stage of the pipeline was tasked with processing an entire year's worth of data before passing it on to subsequent stages. The advantage of this approach is that upon completion of a year's worth of data, and its quality have been assured, the outputs could then be integrated into the training sets for the machine learning models, ensuring consistent updates and improving transcription accuracy as the project expanded.

The central objective was approached from a data science standpoint, aiming to assess the feasibility and overall capability of the project. Although the project was developed in separate modules, each targeting specific concerns with varied techniques, the ultimate goal remained to create a unified product capable of seamless operation throughout its entirety. This automated end-to-end process must effectively manage both yearly volumes of data and individual distinct records. At present, several endpoints within the pipeline necessitate manual intervention, whether it's for data inspection or processing to enable subsequent stages to commence. The forthcoming sections will delve into these limitations and outline the prerequisites for fully automating the pipeline, thereby eliminating the need for such manual interventions.

### 3 The AI-ML Data Analytics Pipeline

Figure 2 illustrates the implicit flow graph underlying the modules of the automated data analytics pipeline. It is a linear pipeline, where the output information produced by one stage is passed on to the successive module for further processing. It takes images provided as TIFF files as input and starts with the recognition of the shape of the table within these images through the use of image segmentation, followed by the extraction of said table.

After extracting a table image, the rows containing individual deceased records undergo segmentation and extraction in the *Row Segmentation* stage. Following this process, for each individual row its 11 columns are segmented in the *Column Segmentation* stage, providing access to the single cells. This top-down approach decomposes the structure of the original TIFF file image to its finest granular level, enabling the subsequent content analysis, through identification of each individual word for each cell. This approach is very general, and it is applicable to all document types containing the same tabular structures. This enables a wide reusability, and it can lead to processing at scale using these procedures.



**Fig. 2.** Flow graph underlying the logic flow of the prototype in [21].

Within these cell regions, the words and phrases can now be identified using object detection networks in the *Word Detection* stage of Fig. 2. Depending on the cell region and expressiveness of the lexicon within this region, data synthesis may then be used to create training data for a series of classifiers (*Data Synthesis* stage). Finally, models are trained and data are classified (*Classification* stage) to generate a normalization of the resulting data, to counteract the subjective nature of human formulation in writing.

Disregarding aspects such as data annotation and model training, a closer examination of this flow graph reveals a simple, unidirectional data flow, suggesting that, in theory, it could be readily transformed into an automated process. However, if we look for example at the internal organization of the *Table Segmentation* module (see Fig. 3), we can recognize some obstacles.

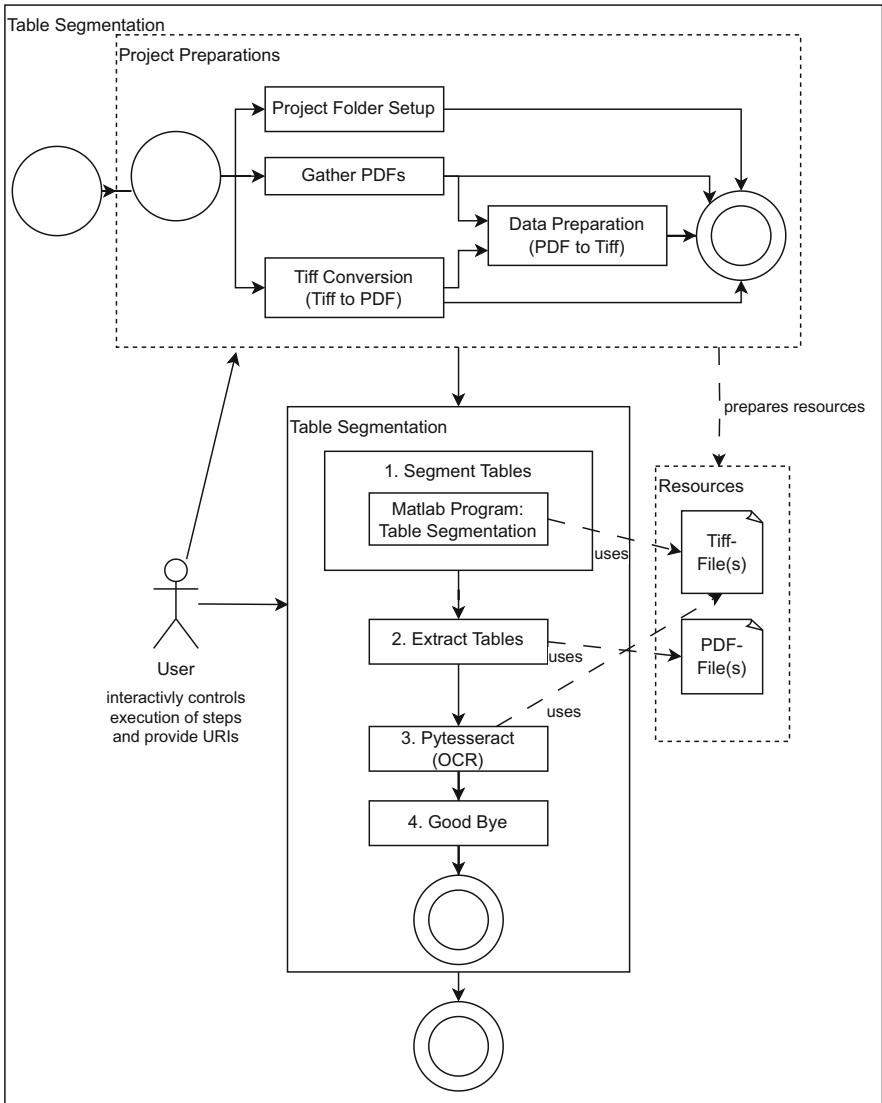
### 3.1 The Complex *Table Segmentation* Module

The *Table Segmentation* module is the initial step in the pipeline that takes image files of tables in the form of TIFF files as input but is not actually an atomic step: it has a complex internal structure with multiple sub-modules. As shown in Fig. 3, there are two main consecutive steps: the *Project Preparations* and the actual *Table Segmentation*.

### 3.2 The *Project Preparations* Submodule

The *Project Preparations* submodule consists itself of three parallel steps that prepare the projects input data.

- Before any operation, the *Project Folder Setup* program establishes a folder structure utilized by the other modules in Fig. 2. This setup is performed



**Fig. 3.** A deeper look at the functionality of the *Table Segmentation* module of Sect. 2.

only once per year of data and ensures a consistent structure throughout the project. This consistency enables the utilization of static location-based arguments when processing artifacts such as image and text files.

- The data provided to the DBDirl team comprised numerous images in PDF and TIFF file formats, organized annually and by district. Within these folders, around 20% were in PDF format, while the majority, approximately 80%, were in TIFF format. A decision regarding the preferred data type for further

processing needed to be made. TIFF files were selected due to their recognized capability to store high-quality images with a resolution of 300 Dots Per Inch (DPI), ensuring optimal detail. This is especially crucial for classifying handwritten text images. This choice is particularly significant as the focus involves narrowing down the document scope to individual words, emphasizing the importance of preserving high-quality image types. However, once again, the process of constructing a dataset of TIFF files is not straightforward and necessitates additional handling. The TIFF files supplied by the GRO vary in size and have a bit depth level of 1 (i.e., just black and white), posing potential issues in the table segmentation module. This is because the network for this module requires input with 3 RGB channels and a consistent size structure. Therefore, there is a need for processing the TIFF files at scale. The most straightforward approach involves adding additional channels to the original TIFF image. However, this method resulted in blurring and a reduction in quality [21]. To address this issue, all available TIFF files were first converted to PDF format, unifying the dataset to 100% PDF. These PDFs were then converted back to the final TIFF format with 3 RGB channels. While the ideal size of the images would be  $[6400 \times 4800 \times 3]$ , this caused memory issues with certain AI network types. It was then determined that the input to the network could be 1/16 of this size and still operate adequately. This process ensured the required input format for the subsequent stages of the data analysis pipeline while preserving the necessary quality.

Note that this process involves user interaction: the user must select the folders containing the specified images to be processed. Once all the data is converted into the conforming formats, it becomes ready for use to extract the tables from the images: the *Table Segmentation* in Fig. 3.

### 3.3 The Actual *Table Segmentation* Submodule

The submodule *Table Segmentation* is also realized as an interactive program. It gives the user the possibility to execute all its actions in the Phases 1 to 4 as desired, even if they should be executed consecutively.

1. In Phase 1, the *Segment Tables* action carries out the actual segmentation of the tables. This is achieved by initiating a child process implemented in Matlab, i.e. as a platform-dependent program on the Windows operating system. The process prompts the user to specify the location of the TIFF images whose table information needs segmentation. The segmentation utilizes a deep neural network model that identifies the coordinates of the specific *region of interest* representing a table. It outputs results as key-value pairs where the key is the absolute file path of the image’s location, and the value is an array with the coordinates of the four points defining the table’s location inside the image. These coordinate values must be scaled by a factor of 16 when performing extraction.

2. The *Extract Tables* stage performs the actual extraction by cropping the segmented regions out of the images. This is done on a page by page basis, where a PDF file is first converted into a TIFF image with an ideal size of  $[4800 \times 6400 \times 3]$ . The *regions of interest* identified by the output of *Segment Tables* and scaled by a factor 16 are then extracted. Since the death records were scanned from a physical register source, their orientations and ranges may vary from page to page. To address this, a de-skewing process is employed before applying the Lanczos interpolation method, all aimed at ensuring the standardization of the resulting table image at a size of  $[3600 \times 4800 \times 3]$ . This table image is not saved to disk, rather it is immediately sent into the Row Segmentation module for separation. Upon successfully extracting the table region and performing the row segmentation operation, the original  $[4800 \times 6400 \times 3]$  image is deleted from memory before processing the following page.
3. The *Pytesseract* action applies the Pytesseract Optical Character Recognition (OCR) tool to the extracted table region images. This serves as a qualitative assessment, ensuring the quality of the extracted table regions by verifying the presence of the printed word ‘*Superintendent*’ in the appropriate region, as depicted in the top-left position of Fig. 1. By creating a set region box across this area, serving as an anchor point, an Intersection over Union test is performed to compare the location of the ideal region with the resulting bounding box produced by the Pytesseract system. If the Intersection over Union falls below a specified threshold when applied across a subset of a year’s data, it indicates an issue in the Table Segmentation process. Consequently, the pipeline cannot proceed until the problem is addressed.
4. If all the tests have passed and the user is satisfied with the outcome, the user can close the program by selecting *Good Bye* to complete the module *Table Segmentation* and exit.

## 4 Issues Encountered for Full Automation

Looking at the internal structure and characteristics of this module, one can notice some pitfalls and obstacles that complicate the transformation of the interactive module into a fully automated process.

1. **No Interactivity/Convention over Configuration:** Each sub-module of the given module prompts the user for some input (e.g., file paths) or to manually execute parts of the process. This interactivity is useful when designing the pipeline step by step, exploring the data, prototyping ideas, and learning to understand the proposed procedure of the *Machine Transcription Pipeline* of [21]. To automate the procedure, however, this interaction is disruptive and must be removed and replaced by agreed conventions. While a folder structure that serves such a conformance is created in the *Project Folder Setup*, it seems to be more suited for the other modules seen in Fig. 2 and its root needs to be specified manually by the user.

2. **Separation of Concerns:** In data science projects, scientific methods and algorithms are used to extract knowledge and insights from data that is often semi-structured or unstructured, like in our case the images of register pages. The code merely serves as a tool to derive information, and it is not seen as a primary product itself, as is customary in software engineering projects. The clean organization of the code is frequently affected. The *Project Folder Setup* sub-module establishes a folder structure that affects the whole procedure seen in Fig. 2. As it covers a much larger scope than where it is located, two nesting levels under the top-level pipeline, it should not be part of this module, but instead constitute itself a new module, initialising all subsequent stages in the Fig. 2 flow graph.
3. **Data Conformance and Optimization:** The *Table Segmentation* sub-module expects two file formats, TIFF and PDF, each chosen for specific purposes. Upon conversion from PDF to TIFF file images, the size is considerable at  $[4800 \times 6400 \times 3]$  with a 300 DPI. The TIFF image employed in the sub-module is a scaled down  $[300 \times 400 \times 3]$  version, which is only used for segmentation by a neural network (see *Matlab Program: Table Segmentation* in Fig. 3) to rapidly determine the corner points. The sub-module *Extract Tables* expects the PDF files containing the original information, which are converted into temporary full sized TIFF images (see Fig. 3), which are then used to extract the region of interest. Such a converted TIFF image takes up a considerable amount of memory and does not scale well with the number of PDF files that need to be processed: just this experimental subset of data consists of 2000 TIFFs at 300DPI with a resolution of  $4800 \times 6400 \times 3$  taking approximately 176 GB of storage space. Therefore, only a limited number of files can be processed in parallel, after which they are deleted to recover storage. The choice of file formats to be used should be evaluated to optimize the performance of the whole process. In this case, the number of file conversions and the size of the data should be reduced to increase the number of files that can be processed in parallel. This could be achieved by finding a more appropriate file format for this particular problem domain. A good candidate for this case could be PNG. While both TIFF and PNG are lossless, PNG has a reduced file size due to compression by limiting the number of color spaces supported.
4. **Remove Dead Code:** The action *Pytesseract* in the *Table Segmentation* sub-module is an ancillary procedure to perform qualitative assessment upon completion of the Table Segmentation process, but it is irrelevant in the scope of the automated pipeline shown in Fig. 2. Additionally, a series of modifications and transformative methods are repeated throughout the pipeline. These would need to be refactored into a single process to avoid code redundancy and enhance maintainability.
5. **Platform Independence:** The regions of interest inside the images are identified by a Matlab program executed by the *Segment Tables* process (see Phase 1: *Matlab Program: Table Segmentation* in *Segment Tables*). In an early version of this project, the Matlab program was platform-dependent on the Windows operating system. Initially, the pipeline operated on a file by file basis,

employing the executable to perform segmentation when required. This approach restricted the executability (at least) for this module to only one operating system/platform. To give better support for a possible migration to other platforms, such a procedure should be executed in a platform-independent way. Fortunately, the Matlab runtime environment offers the possibility to run the Matlab program as a script, platform-independent. This functionality was not used in this project, but rather the explicit execution of a Windows executable of the Matlab program from within the Python code. It should be noted that this singular processing approach was later changed to operate at scale by writing all segmentation elements to text files at once, avoiding the repeated usage of the Matlab executable [6].

This analysis concerns only the first module of the flow graph in Fig. 2, therefore the issues identified so far might not be the complete set of problems to be considered when migrating the program based on a data science project to an automated process: we expect to find more kinds and variety of migration problems. They are however indicative of the spread of concerns that play a role when moving towards a “platform” thinking, with robust, efficient and easily reusable modules, which is our next goal.

## 5 From a Single-Purpose Monolith to General-Purpose Modules

The workflow as outlined above is a custom tailored collection of scripts geared towards the specific use-case of these death-and-burial records and it works around the idiosyncrasies, problems and oddities that are specific to these kinds of documents. However, many of these problems are of a general nature. By swapping just parts of the workflow, it should be possible to adapt the workflow to other sets of register-like sources, extending the applicability to problem domains with different kinds of records with different kinds of information. Furthermore, automating the workflow should allow batch processing of the whole collection of all death-and-burial records, especially for all years, without human intervention. Note that this way there is no restriction on the transcription to a specific language if there are suitable trained models for other languages.

Thus, there are two independent drivers for automating the workflow. In the short term, automating the workflow would quickly generate more usable results for other years of the same dataset, that can be used by Irish historians to advance their research interests. Automation would also pave the way for improvements in modularization and (re-)composition.

However, the pipeline as developed is monolithic in nature, with many implicit dependencies between the steps, which require careful manual orchestration. There are many interaction points in the current work, which require human intervention and/or input at the time of execution. For a full automation it would be necessary to remove these interaction points and instead collect the required inputs as one input vector that can be provided ahead-of-time, prior



to execution of the workflow. This can be implemented as a configuration file, command-line parameters or any other form of parameterization from which these values can be collected and then injected at the proper points during runtime.

Similarly, a desirable property for each step is purity. Ideally, a single step would only consume a defined set of inputs and generate a defined set of outputs. While this is a desirable property in theory, practice shows that it is not always feasible. Nonetheless, side-effects onto the runtime-environment, the system or data should be minimized and formally captured in an appropriate representation that allows for this data to be formally reasoned about and to be tractable for a tool that automatically orchestrates the workflow. We have found that many steps in the current workflow leverage implicit dependencies on the state of the environment and in-memory state, and manipulate state via side-effects, i.e. through global variables that change during loop iteration. This does not pose a problem if the steps are executed one-by-one, as single processes, but it severely restricts the ability for automated tools to freely re-compose the steps into single scripts that have a shared runtime lifecycle.

In a similar vein, redundancies exist in the current workflow where similar code occurs in multiple places. By treating the whole workflow as a model, these redundancies can and should be eliminated and transformed into components that can be re-parameterized and re-used at various locations, and then woven into the finished workflow by means of an automated orchestrator.

Aside of purity or the formal capture of side effects, idempotency [22] (and, related to that, binary reproducibility) is another desirable property for any such workflow. While purity is not practically attainable in many cases, idempotency is. In this context, an idempotent operation may have side-effects (it is not pure), but given the same inputs, the same operation always produces the exact same side effects (for example, generating the same files with the same content). Thus, idempotency can be leveraged for caching, cache reconstruction and many other desirable features.

The current workflow is platform specific and only works on a Windows-platform. It also has hard-coded mitigation strategies for some limitations of that platform. Removing some of these platform dependencies is rather trivial, for example replacing calls to Windows `.exe` files with generic command-line calls or even moving to a containerized execution. Other platform-specific aspects, for example to combat limitations on the size of file handles, need more careful consideration. Again, a model-based approach could be leveraged to auto-generate this code for each relevant platform, instead of requiring it to be baked into the user-level code.

A model-based approach would also allow for better scheduling. The workflow as presented is linear in nature, as it is executed by a human step-by-step. But, if viewed from an information science point of view, it represents a dataflow that can be modeled as directed-acyclic-graph. It forms a dependency-tree with multiple fork- and join-points, and has several paths that could be executed in parallel and then synchronized at the join points. An orchestrator could freely re-

arrange the execution steps to improve execution times by parallelizing execution and smart caching of dependencies.

In summary, redefining the workflow as a proper process model and re-shaping the scripts from the prior work into pure, idempotent, clearly defined and interaction-free modules would enable a great deal of optimization techniques to be applied. Modelling the workflow as dataflow would also enable security and quality assurance aspects to be checked statically, for example with model checking [23]. It is also an essential step in providing a low-code/no-code graphical environment to freely re-compose the workflow. Such a graphical mIDE [24, 25] can enable non-experts to better understand and even change parts of the workflow [26]. Automating the workflow would furthermore allow interfacing with existing low-code/no-code tools that are already used in the field today, e.g. the Historian DIME App [9]. Thus, it would make the prior work much more accessible and usable for Irish historians and enable even closer cross-domain and interdisciplinary collaboration between historians, data scientists and computer scientists.

## 6 Analytical Overview

After detailing the challenges and issues present in the current state of the pipeline in Sect. 3, and outlining the requirements and general approach to transforming it into general-purpose modules, this section aims to provide an assessment of the individual pipeline steps with respect to their current state regarding the individuated obstacles and pitfalls. We look at three major categories of quality for the outcome of the automation process:

- **Maintainability/Usability:** This category describes problems such as *Separation of Concerns* and *Dead Code* [27] but also incoherent code style in general. While these problems may seem a minor concern, they should be taken into consideration when rebuilding the pipeline as they impact the future maintenance and usability of the platform.
- **Performance:** Efficiency is a significant concern, particularly in the realm of data science where handling substantial amounts of data is common. This category encompasses issues stemming from data inconsistencies, redundancies in processing steps, and broader performance concerns such as unnecessary cyclomatic complexity.
- **Automation alignment:** The final category comprises problems directly affecting the transformation of the pipeline into an automated process. Current instances of interactivity, such as prompting for data input and configuration settings, must be either fulfilled by injecting the necessary information or entirely replaced by predefined configurations. Additionally, all process steps should be platform-independent to remove restrictions on supported operating systems.

Table 1 provides an overview of the number of major technical issues that require immediate remedying within each pipeline module. The numbers differ across the categories.

**Table 1.** Occurrences of issues in each pipeline segment, per category

	Maintainability	Performance	Alignment
Table Segmentation	3	1	2
Row Segmentation	1	0	1
Column Segmentation	2	2	2
Word Detection	0	0	1
Data Synthesis	0	0	1
Data Classification	4	2	5
Total	10	5	12

- *Performance* issues are the least observed, with 5 instances. This is positive, as it indicates that the initial purpose of the pipeline, the efficient extraction of data, is already quite satisfied.
- *Maintainability* issues (10 instances) are mostly due to the prototyping nature and thus quickly changing code base.
- *Alignment* issues are the most frequent, with 12 instances. This was to be expected, because the transformation into an automated process was not in the initial scope of the project and thus of no concern until now.

## 7 Conclusion

We have shown that there is a big difference between the code organization and the orchestration choices in a typical data analytics and AI based project, aimed at solving a specific problem for experts, and the organization needed for a platform based reusable, extensible domain specific implementation that makes the single components platform-agnostic, manages dependencies in an organised, separate way. Similar approaches have been successful in the past for generic tool wrapping [28], in applications to, e.g., bioinformatics [29] and climate change analysis [30], and in Semantic Web applications [31]. This approach supports an easy reuse and repurposing of the components, together with an ease of evolution and reorganization of the processes, substituting the hard-wired pipeline or highly interactive manual execution.

The next steps will transform the pipeline into an automated process. This reconstruction and transformation is the major goal going forward, and in this course several concerns must be addressed. First, all manual interactions need to be served either by externally injecting needed information or by predefined configurations. Furthermore, all the current instances of platform-dependent process steps need to be replaced by platform-independent operations.

Besides these, a major concern lies in the way big data sets and models are processed. Transferring data between independent process steps needs to be avoided as it causes unnecessary overhead regarding performance and traffic load. Thus, data should be processed in place.

So far the main data transcription was done with the Historian DIME App [9], which enables a fast but manual input of transcribed text or data. As a first step towards automation, this app should be enhanced with the content recognition workflow in order to pre-fill automatically identified data into the corresponding fields. This way the users could either confirm (if correct) or otherwise amend the prefilled data, and save time while providing a better training data for the automatic recogniser.

As we have seen in the GREATLEAP context, the sheer volume of TIFFs available to archives, collections, and online is astounding. Considering that their high quality transcription is a once-only activity, and that it is essential to unleash a wealth of research relevant to many disciplines, it is definitely worth to research how to systematically produce excellent AI and ML supported capabilities for fully or semi-automatic transcription.

We intend to do it in an XMDD and LC/NC fashion as such a platform has the best chances to leverage the community of practices and empower the researchers to “self-assemble” or evolve and modify workflows for their data sources and aims.

This modular workflow has huge potential not just for historians: government documents, handwritten patient records in hospitals, old business records, legal records and other documents exist in large numbers, and they are unused or hardly accessible. Eventually, the goal is to achieve a fully configurable graphical transcription environment that can quickly adapt the transcription process to other types of documents.

**Acknowledgements.** This work was conducted with the financial support of Science Foundation Ireland (SFI) under grants number 18/CRT/6223 574 (SFI Centre of Research Training in AI) and 21/SPP/9979 (R@ISE).

## References

1. Doherty, A.J., Murphy, R.A., Schieweck, A., Clancy, S., Breathnach, C., Margaria, T.: CensusIRL: historical census data preparation with MDD support. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 2507–2514 (2022)
2. Murphy, R.A., Breathnach, C., Schieweck, A., Margaria, T.: Interoperating civil registration of death and census data: old age and marriage as categories of analysis. In: Proceedings of the AISoLA 2023. Springer (2024)
3. Pedersen, B.-R., et al.: Coding historical causes of death data with large language models. In: Proceedings of the AISoLA 2023. Springer (2024)
4. Mourits, R.J., Riswick, T., Stapel, R.: Common language for accessibility, interoperability, and reusability in historical demography. In: Proceedings of the AISoLA 2023. Springer (2024)
5. Breathnach, C., Ibrahim, N.M., Clancy, S., Margaria, T.: Towards model checking product lines in the digital humanities: an application to historical data. In: ter Beek, M.H., Fantechi, A., Semini, L. (eds.) From Software Engineering to Formal Methods and Tools, and Back. LNCS, vol. 11865, pp. 338–364. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-30985-5\\_20](https://doi.org/10.1007/978-3-030-30985-5_20)

6. O'Shea, E., Khan, R., Breathnach, C., Margaria, T.: Towards automatic data cleansing and classification of valid historical data an incremental approach based on MDD. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 1914–1923, (2020)
7. Lladós, J., et al.: A tale of two transcriptions machine-assisted transcription of historical sources. *Hist. Life Course Stud.* **2**, 1–19 (2015)
8. archives.gov. 1950 US Census Transcription Projects (2023). Accessed 25 Sep 2023
9. Schieweck, A., Murphy, R., Khan, R., Breathnach, C., Margaria, T.: Evolution of the historian data entry application: supporting transcriptions in the digital humanities through MDD. In: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 177–186 (2022)
10. findmypast.co.uk. How was the 1921 census transcribed (2023). Accessed 25 Sep 2023
11. Thorvaldsen, G.: Automating historical source transcription. *Hist. Life Course Stud.* **10**, 59–63 (2021)
12. Pedersen, B.R., et al.: Lessons learned developing and using a machine learning model to automatically transcribe 2.3 Million handwritten occupation codes. *Hist. Life Course Stud.* **12**, 1–17 (2022)
13. Mas, J.J., Fornés, A., Lladós, J.: An interactive transcription system of census records using word-spotting based information transfer. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 54–59 (2016)
14. Sibade, C., Retornaz, T., Nion, T., Lerallut, R., Kermorvant, C.: Automatic Indexing of French Handwritten Census Registers for Probate genealogy. In: HIP 2011, Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 51–58. Association for Computing Machinery, New York, NY, USA (2011)
15. Sibade, C., Retornaz, T., Nion, T., Lerallut, R., Kermorvant, C.: Thanatos: automatically retrieving information from death certificates in Brazil. In: HIP 2011, Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 146–153. Association for Computing Machinery, New York, NY, USA (2011)
16. Nion, T., et al.: Handwritten information extraction from historical census documents. In: ICDAR, pp. 822–826 (2013)
17. Dahl, C.M., Johansen, T.S.D., Sørensen, E.N., Westermann, C.E., Wittrock, S.F.: Applications of machine learning in document digitisation. arXiv preprint [arXiv:2102.03239](https://arxiv.org/abs/2102.03239) (2021)
18. Lin, W., et al.: TSRFormer: table structure recognition with transformers. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 6473–6482 (2022)
19. Petitpierre, R., Kramer, M., Rappo, L.: An end-to-end pipeline for historical censuses processing. *Int. J. Doc. Anal. Recogn. (IJDAR)* **26**, 419–432 (2023)
20. Margaria, T., Steffen, B.: Service-orientation: conquering complexity with XMDD. In: Hinchey, M., Coyle, L. (eds.) *Conquering Complexity*, pp. 217–236. Springer, London (2012). [https://doi.org/10.1007/978-1-4471-2297-5\\_10](https://doi.org/10.1007/978-1-4471-2297-5_10)
21. O'Shea, E.: Machine transcription pipeline for death and burial data - Ireland 1864-1922 (dbdirl) (2024)
22. Peirce, B.: *Linear Associative Algebra*, pp. 97–229 (1881)
23. Steffen, B.: Data flow analysis as model checking. In: Ito, T., Meyer, A.R. (eds.) *TACS 1991*. LNCS, vol. 526, pp. 346–364. Springer, Heidelberg (1991). [https://doi.org/10.1007/3-540-54415-1\\_54](https://doi.org/10.1007/3-540-54415-1_54)

24. Steffen, B., Gossen, F., Naujokat, S., Margaria, T.: Language-driven engineering: from general-purpose to purpose-specific languages. In: Steffen, B., Woeginger, G. (eds.) *Computing and Software Science*. LNCS, vol. 10000, pp. 311–344. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-91908-9\\_17](https://doi.org/10.1007/978-3-319-91908-9_17)
25. Lybecait, M., Kopetzki, D., Zweihoff, P., Fuhge, A., Naujokat, S., Steffen, B.: A tutorial introduction to graphical modeling and metamodeling with CINCO. In: Margaria, T., Steffen, B. (eds.) *ISoLA 2018*. LNCS, vol. 11244, pp. 519–538. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-03418-4\\_31](https://doi.org/10.1007/978-3-030-03418-4_31)
26. Lamprecht, A.-L. (ed.): *User-Level Workflow Design*. LNCS, vol. 8311. Springer, Heidelberg (2013). <https://doi.org/10.1007/978-3-642-45389-2>
27. Romano, S., Vendome, C., Scanniello, G., Poshyvanyk, D.: A multi-study investigation into dead code. *IEEE Trans. Software Eng.* **46**(1), 71–99 (2020)
28. Margaria, T.: Web services-based tool-integration in the ETI platform. *Softw. Syst. Model.* **4**, 141–156 (2005)
29. Lamprecht, A.-L., Margaria, T., Steffen, B.: Seven Variations of an alignment workflow - an illustration of agile process design and management in bio-jETI. In: Măndoiu, I., Sunderraman, R., Zelikovsky, A. (eds.) *ISBRA 2008*. LNCS, vol. 4983, pp. 445–456. Springer, Heidelberg (2008). [https://doi.org/10.1007/978-3-540-79450-9\\_42](https://doi.org/10.1007/978-3-540-79450-9_42)
30. Al-areqi, S., Lamprecht, A.L., Margaria, T., Kriewald, S., Reusser, D., Wrobel, M.: Agile workflows for climate impact risk assessment based on the ci: grasp platform and the jABC modeling framework (2014)
31. Petrie, C.J., Margaria, T., Lausen, H., Zaremba, M.: *Semantic Web Services Challenge: Results from the First Year*, vol. 8. Springer Science & Business Media (2008). <https://doi.org/10.1007/978-0-387-72496-6>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Mapping Madness: HGIS and the Analysis of Irish Patient Records

Oonagh Walsh<sup>1</sup>(✉) and Stuart Clancy<sup>2</sup>

<sup>1</sup> Glasgow Caledonian University, 70 Cowcaddens Road, Glasgow G4 0BA, Scotland, UK  
oonagh.walsh@gcu.ac.uk

<sup>2</sup> University of Limerick, Limerick V94 T9PX, Ireland

**Abstract.** The Connaught District Lunatic Asylum (CDLA) opened at Ballinasloe, Co. Galway in 1833 as one of the first of a nationwide network of Irish District Asylums. Intended to serve the curable pauper lunatics of the counties of Mayo, Sligo, Leitrim, Galway, and Roscommon, the institution found itself at the heart of significant social, economic, and political change in the West of Ireland. From its opening, the asylum maintained a full and complex series of records that provide an exceptional level of detail on a cohort – the very poor and illiterate population of Connaught – who otherwise often lived and died unrecorded on the margins of Irish society. The CDLA admission records include information on age, sex, occupation, education, religion, marital status, places of origin and residence, migration, and family structures as well as the medical information, both mental and physical, required for treatment in the asylum.

This paper will examine the potential benefits of implementing spatial epidemiological methods into historical studies of mental illness. Using a database of patient records this paper will conduct a demographic analysis of a sample of the population of the CDLA. The paper will outline the process of transforming the data extracted from these records into visual maps using Historical GIS (HGIS). Using the geographic co-ordinates of these two locations, the unique patterns of movement of those that entered the asylum can be mapped using GIS. These maps enable the examination of the socio-spatial processes which affected the marginalised population of the asylum. (The sources used in this paper come from the Connaught District Lunatic Asylum records, held at the National Archives of Ireland in Dublin. As the material is drawn from committal warrants of patients admitted in 1889, it falls outside of the ‘100 year rule’ for accessing sensitive historic patient information in Ireland.)

**Keywords:** History · Irish Insanity · Digital Humanities · GIS · Demography

## 1 Introduction

The history of psychiatry is often a history of loss: of status, of family, and of course, of mind. Patients ‘lost themselves completely’ when ill, behaving in ways that set them apart from community and society. It is no coincidence that the original word for a psychiatrist was ‘alienist’, derived from the French term *médecin aliéniste*, a person who became

responsible for the treatment of individuals who were separated both from society and themselves as a result of mental ill health. Patients were aliens in every sense of the word, as their delusions and strangeness of behaviour positioned them as unknown and unpredictable beings. Long-stay patients in asylums were in great danger of losing their connections to the outside world, as families visited infrequently, and in at least some cases, preferred to quietly deny their link with a person stigmatised by mental illness. This paper seeks to restore some of the lost connections between patients, their families, and the broader society from which they came by directly mapping the co-ordinates of their origin. This enables further verification of key demographic factors such as the patient's socio-economic circumstances. Using spatial epidemiological methods, we seek to reconstruct the patient's non-asylum social world, and re-think the historiography of mental illness. Rather than focus on an individual experience of committal, HGIS supports cohort analysis that engages with the non-medical factors in admission, thereby situating the asylum population in the complex world of nineteenth century Ireland. It opens up the record body to a new form of analysis, with record linkage helping to track patients across a longer period of their life cycle. It permits large bodies of data to interact, facilitating the recreation of patient movement to and from institutional care, as well as the world of work. This fine-grained connection has enormous potential to deepen our understanding of patient life beyond the asylum, which has been largely invisible in the past.

## 2 The Asylum: Reach and Importance

The Connaught District Lunatic Asylum (CDLA) opened at Ballinasloe, Co. Galway in 1833 as one of the first of a nationwide network of Irish District Asylums. It was situated in an impoverished rural region in the west of Ireland, one that endured chronic underemployment, periodic food crises, and a growing pauper population for the entirety of the nineteenth century. The CDLA had been intended to serve only the curable pauper lunatics of the counties of Mayo, Sligo, Leitrim, Galway, and Roscommon, and to implement the optimistic development in mental health provision that was 'moral therapy'.<sup>1</sup> However, the institution rapidly found itself at the heart of significant social, economic, and political change in the West of Ireland. As Irish society came increasingly to accept and exploit such institutions, admissions to the asylum were often driven by considerations that had little to do with mental ill-health.<sup>2</sup>

Although Michel Foucault described the eighteenth century as a period of the so-called 'Great Confinement', it was in fact the nineteenth that saw the sharpest rise in admissions to the national system of institutions for the care of the mentally ill in Ireland and Britain. From 1800, both countries established a regional network of lunatic asylums

---

<sup>1</sup> Moral Therapy refers to the belief that if mentally ill patients were treated gently and permitted to live in the manner to which they had been accustomed at home, they would recover their senses. Although a far more humane principle than the earlier, largely custodial, regime that preceded it, it proved expensive and less successful than had been hoped.

<sup>2</sup> This paper uses original terminology for mental illness, as it appears in the records. Although jarring to the modern ear, the terms were accepted medical discourse in this period, and their usage in this paper protects the integrity of the historic sources.



that grew exponentially, to address what was perceived to be the growing problem of mental ill-health. The development was part of an increasing centralisation of state power that is mirrored in the growth of workhouses (established under the Poor Law), prisons, specialist institutions for the care of the intellectually disabled, and a general hospital system. The second half of the nineteenth century also saw a significant rise in the numbers of general hospitals, and substantive improvements in public health that included compulsory vaccinations, slum clearance, public sewage systems, and care for the elderly and infirm.

These developments were intimately associated with a broader European concern with what was known as 'the national stock'. In an era of 'land grabs' in Africa, as well as increasing tensions between European Imperial powers, governments saw a tightly regulated health and welfare system as a means of ensuring a healthy domestic and overseas army to take and hold territories, as well as producing replacement populations. Mental illness, especially on the scale that appeared to be evident in the British Isles, was a direct threat to such an ambitious global positioning. Throughout the nineteenth-century, the British government became increasingly concerned with what was regarded as an apparently inexorable rise in mental illness in Ireland in particular which saw institutionalized patient numbers rise to over 21,000 by 1900: there were an additional 9,000 lunatics in Workhouses (mostly inmates with dementia), giving a total of over 30,000 individuals in institutional care designated insane. This growth is all the more extraordinary when placed against a sharp population decline in Ireland in the second half of the nineteenth century, from 8.2 million in 1845, to 4.4 million in 1901. The CDLA opened in 1833, to serve the entire population of the province of Connaught. The provincial population stood at almost 1.5 million in this period, so the asylum's 150 bed provision was by any standard a modest provision. Within a year of its opening, it housed 300 in-patients, an inexorable rise that continued for over 100 years until its peak of over 2,000 residential inmates in the 1950s. The asylum shaped, and was itself shaped by, local and national politics, becoming a crucial source of not only medical treatment but employment, therapy, and refuge for a vulnerable population. In the 1860s, the catchment area for the CDLA reduced from the original five counties of the province of Connaught to the sole counties of Galway and Roscommon. Despite this shrinkage, the asylum population maintained a steady increase, apparently confirming the government's belief that the Irish character was inherently unstable.

But how to evidence this ethnically-based belief? The answer lay in a sophisticated data-gathering exercise that was conducted at every moment in the asylum admission, treatment, and discharge process. Irish insanity was uniquely regulated by a piece of legislation that created a dangerously destructive association between mental ill health and criminality, through the 1838 enactment of the Dangerous Lunatics Act. This extraordinarily liberal (or more accurately repressive) piece of legislation operated to eliminate even the basic civil rights of the nineteenth century, permitting the arrest and imprisonment of any individual accused of 'dangerous insanity'. Intended for emergency use only, it became the default means of committal of the vast majority of patients to the CDLA throughout the century, and well into the twenty-first.<sup>3</sup> The Act permitted any

---

<sup>3</sup> The DLA accounted for over 75% of all male admissions in the early 1890s, despite being a provision was intended for exceptional and emergency use.

person to make an accusation of dangerous insanity against another, without corroborating evidence or witnesses. The accused was arrested, brought before two Justices of the Peace, who without benefit of any medical expertise determined if the person before them was indeed a dangerous lunatic: in the vast majority of cases they agreed that this was the case. The unfortunate accused was then transported to gaol, from where after a period of detention (from a matter of days, to a month or more) they were transferred under armed constabulary escort to the nearest District Asylum. The Act required that the asylum accept every person brought to them under the DLA, even in those cases where non-medical factors had precipitated the arrest. There were frequent cases at the Ballinasloe Asylum where the accompanying constables told the asylum Physician that a dispute over land or other resources lay at the heart of an accusation of insanity, but the Public Order imperative of the DLA overrode all other considerations. The Act was uniquely open to exploitation, and proved an irresistible temptation to relatives who sought to rapidly admit a family member, whatever the primary motivation.<sup>4</sup>

### 3 Database Creation and Demographic Analysis

To date there have been few studies conducted which analyse spatial patterns relating to mental health in the present and even fewer analysing these patterns in the past. Folasade Lyun has argued that the need to conduct psychiatric epidemiologic studies arises from the correlation between mental illness and socio-cultural environmental factors. The analysis of historical medical records using digital methods provides a unique set of challenges, particularly in contrast to the analysis of civil records from the same period. Civil records such as birth, marriage and death records are standardly uniform in nature and specific guidelines were frequently used for the record's initial population. The records that were utilised for this research contain a wealth of qualitative and quantitative information as is visible in Fig. 1.

The information present on.pdf copies of these records were manually input into an excel database which had been broken down into fifty individual categories. These categories were designed to allow the data to be analysed at a granular level. The process of creating the database initially centred around the extraction of key demographic factors. These include the patient's age, sex, residence, and occupation. Several categories were included which indicated the nature of the patient's conditions such as 'species of insanity' and 'probable cause of derangement'. The inclusion of the patient's 'friends' or family members was intended to aid the mapping process and improved the ability to link the patient records with additional datasets. In a number of cases, it was possible to link the patient's original residence to the records of the census in 1901. The inclusion of family members allowed for the potential to confirm that the household was correct. A 'census' category was included in the database for this purpose. This category included a link which could be easily used to refer back to the census registers on the National Archives of Ireland website. The census records are vitally important to demographic

---

<sup>4</sup> Because the DLA guaranteed admission, it was deliberately used by families to access immediate treatment for their relatives, in the hope of a rapid cure and discharge. This usage however led to gross overcrowding of all of the District Asylums, and a deterioration in care.

The following Forms must be filled up by the Medical Officer who has personally examined the Lunatic or Idiot:—

**I. MEDICAL CERTIFICATE.**

I certify that \_\_\_\_\_, whom I visited on \_\_\_\_\_ day of \_\_\_\_\_ and into whose case I specially and personally inquired, is now a dangerous opinion, from the nature of his malady, that he is a fit subject for speedy admission into Lunatic Asylum, under the provisions of the Act 30 and 31 Vict., c. 119, s. 16.

Date, \_\_\_\_\_ 18\_\_\_\_

Signature of Medical Officer \_\_\_\_\_  
Residence, \_\_\_\_\_  
Dispensary District, \_\_\_\_\_

**2. STATEMENT OF PARTICULARS OF CASE.**

Species of Insanity.	Probable Cause of Derangement.	Prominent Symptoms.	Whether affected with Bodily Disease.	Whether Idiotic or Epileptic.	Facts indicating that the Patient is a Dangerous Lunatic or Idiot.

I hereby certify that this Form is filled up correctly, to the best of my opinion and belief

Date, \_\_\_\_\_  
Signature of Medical Officer \_\_\_\_\_

The following Forms must be filled up by the Friends of the Lunatic or Idiot.  
If no friends of the Lunatic or Idiot are known this Form may be filled up by the Police so far as their information will enable them to do so.

**NAME OF THE TWO NEXT KIN TO THE LUNATIC OR IDIOT.**

Relative's Name.	Residence and Post Town.	Occupation.	Degree of Relationship.

Age of Lunatic or Idiot.	Religion.	Place of Birth.	Place of Abode.	Occupation of Family, and whether name of his own.	Whether Single or Married, and if a Female, whether she has had Children.	Whether and how long he has been Insane.	How long ill, and if violent.	Habits of Life, Temperate or Intemperate, &c.	Education.

Date, \_\_\_\_\_  
Signature, \_\_\_\_\_

**Fig. 1.** The second page of the Committal Warrant of a Dangerous Lunatic or a Dangerous Idiot.

studies of this period as they can provide a strong indication of a person’s class and background (Table 1).

**Table 1.** All categories present in project database.

Date	Patient Name	Age	Male or Female	Dangerous	Religion	Place of Birth
Means	Marital Status	No. of Children	Whether any near Relative had been Insane	Length of Illness Years	Length of Illness Months	Length of Illness Days
Probable Cause	Education	Species of Insanity	Temperate or Intemperate etc.	Prominent Symptoms 1	Prominent Symptoms 2	Whether Affected with Bodily Disease

(continued)

**Table 1.** (continued)

Date	Patient Name	Age	Male or Female	Dangerous	Religion	Place of Birth
Residence and Post Town 1	Occupation	Degree of Relationship 1	Place of Abode	Residence and Post Town 2	Relative's Name	Degree of Relationship 2
County	If Violent	Habits of Life	Whether Idiot or Epileptic	Additional Info	Medical Officer	Facts Indicating that the Patient is Dangerous

Difficulties with uniformity and the varying nature of cases created a number of challenges throughout this process. The database is required to constantly evolve throughout this period of transcription to accommodate the needs of the researcher and to include new elements that appear when analysing the records. During the initial process of transcription, the categories were populated using the exact text present. There were no generalised standards for the population of these records. As a result, the contents of the records had a variety of linguistic variations for a common term, dependent on the medical officer's preferences. Once the database was completed these terms were standardised. This involved deviation from the exact text for the purpose of analysis in certain categories. For example, in the 'probable cause' column of the database, there was a high degree of variation in language used to describe an unknown cause of insanity. These included 'not known', 'cannot assign cause' and 'no assignable cause'. To aid with the analysis of the data these were all reclassified as 'unknown'. Similarly, there were issues with the education category. In the majority of cases patients were described as having the ability to 'read and write' or that they had 'none'. In certain cases, this category was used to describe the patient's level of education. Terms such as 'fair' and 'average' were used. For the purpose of this research all patients with some degree of education were classified as literate and those without education were classified as illiterate.

The co-ordinates of the CDLA were input into 'AsyLat' and 'AsyLong' categories. The inclusion of this category creates the ability to understand the distance travelled by patients as well as the expansive reach of the asylum. The patient's original residence was input into the 'OriginLat' and 'OriginLong' categories which were required for the data to be processed correctly by ArcGis. The co-ordinates of the original residence for all thirty-three were then added to the additional 'decimalLatitude' and 'decimalLongitude' categories in the excel database.

The analysis of this database can provide a wealth of information regarding the demographic of asylum patients in this period. This pilot study examined thirty-three patients which had been admitted to the asylum in June and July 1889 -three of these patients were deemed to be 'dangerous lunatics' as defined under the Lunacy (Ireland)

Act 1838<sup>5</sup>. This act saw a major overhaul to the legislation in place and introduced new classifications for patients who were deemed to be dangerous. This act stated that an individual who was considered to be a ‘dangerous lunatic’ could be committed to an asylum indefinitely by two justices of peace. The remaining two were committed to the asylum as pauper lunatics. There were no private patients present in this dataset. Three of the primary indicators of the patient’s class and background are religion, occupation, and literacy levels. The majority of the patients admitted to the CDLA were from poor backgrounds. All thirty-three patients admitted to the institution were Roman Catholic. As Table 2 shows, over 80 per cent of the patients admitted in this period were employed as or dependents of a farmer, a labourer, or a person of no means. The analysis of this data provides strong evidence indicating a prevalence of admissions amongst the poor, as one would expect in a pauper asylum. However, the data also shows that individuals of relative means – farmers, landowners, and tradesmen – declared themselves paupers to avoid the imposition of fees for their care:

**Table 2.** Patient Occupations from June-July Committal Warrants.

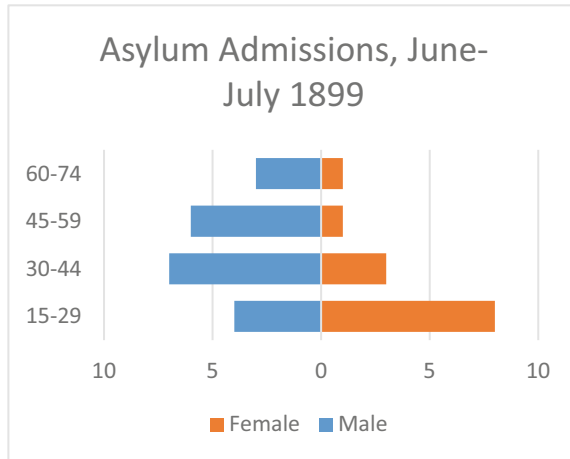
Occupation of patients admitted to the CDLA, June-July 1899	
Occupation (or dependent of)	No. of patients
Farmer	10
Labourer	10
None	7
Craner	1
Herdman	1
Landholder	1
Miller	1
Saddler	1
Shoemaker	1

This information can be used in conjunction with further demographic analysis to gain an understanding of the general profile of the asylum’s population. Table 3 and Fig. 2 show the age of admission to the asylum. The categories were chosen to create four individual demographic categories that were evenly divided. These categories begin at the age of 15 which was observed as the lowest age admission. The highest age of admission was 69 but 74 was chosen to enable the inclusion of equal incremental demographic categories. It is notable that the patients admitted to the asylum were of a younger demographic. Each demographic shows a steady decline. It is possible that this coincides with population decline of the period, further analysis using these methods in conjunction with civil records may be able to show a potential correlation in these trends.

<sup>5</sup> All cases taken from committal warrants for patients admitted to the Connaught District Lunatic Asylum from June-July, 1889.

**Table 3.** Age demographic of patients admitted to the CDLA, June-July 1899

Age Group	Number of Patients
15–29	12
30–44	10
45–59	7
60–74	4

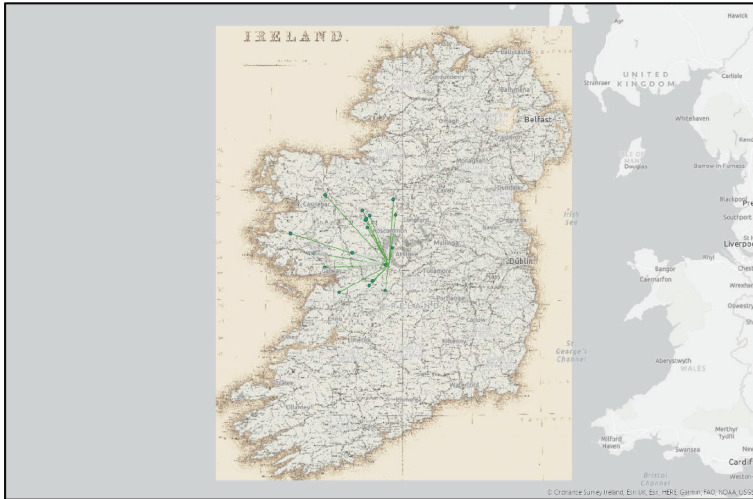
**Fig. 2.** Female/Male Admissions, June/July 1899

## 4 Mapping Process

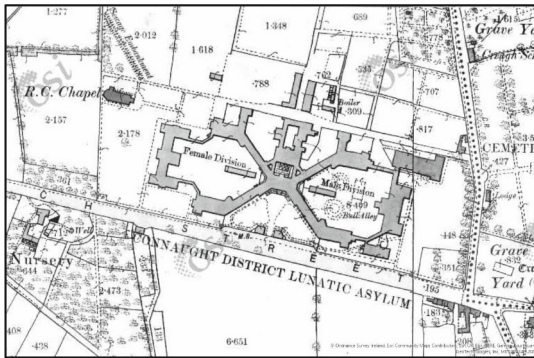
The contemporary maps mentioned earlier also played a large role in this aspect of the mapping process. The modern map that was referenced was the Open Street Map (OSM) Humanitarian Model. The older maps Ordnance Survey of Ireland (OSI) were overlaid onto the OSM. The OSI maps used for this project were created in Ireland between 1892 and 1911. This iteration was used as the period of research fell within this time period, meaning very few changes in the landscape would have occurred. This mapping process allows for historic villages and townlands to be mapped accurately, even if they no longer exist. This overlay is visible in Figs. 3 and 4 below.

Programmes such as ArcGIS have analysis tools which allow for the calculation of unique geographic factors that provide further insights into the expansive influence of the asylum. A key spatial factor that was derived using these methods was the distance from the asylum to the patient's original residence. The co-ordinates of the CDLA were input as additional columns in the excel database. This allows a visualisation of the distance from the patient's origin to the CDLA as seen in Fig. 5 below.

Each of these new databases were then formatted correctly with the appropriate titles to be inputted into ArcGIS. Each was exported from Microsoft excel as a.csv file and



**Fig. 3.** Complete map of Ireland overlaid with Historical 25 inch



**Fig. 4.** The CDLA as it appears on the Historic 25 Inch OS Map

inputted into ArcGIS. This allows the point that is generated in ArcGIS to retain the all of the original data in the excel and associate it with the individual point. The GIS programme itself can be seen as an additional database used to process this information. GIS is better viewed as a type of database than exclusively a mapping tool. What makes this database unique is that each item of data has an allocation stored. The data is then presented in a variety of chosen formats. This transforms the data into an electronic tool that allows the researcher to instantly answer a variety of research questions relating to the topic. Following the input into GIS the data was transformed from the initial data to new formats. The final representation consists of two types of data. The first is the attribute data which was presented previously as quantitative data in the form of a chart or a graph. The next is spatial data which locates the data using a line, a pixel or a polygon. This is known as vector data and is used to represent discrete features in GIS. The pixels which are known as raster data represent continuous surfaces. Attribute

data represents the ‘what’ when analysing the data whilst the spatial data represents the ‘where’. Using the data derived from the admission records in conjunction with maps emphasises the great extent to which mental illness was primarily a rural issue or indeed the reprimanding of the patient was to be more likely.

The utilisation of ArcGIS in combination with classical historical research methods can address several issues that cannot be resolved with the strict use of empirical data or qualitative analysis alone. The blending of these methods enables the ability to contextualise and frame the data in unique yet understandable manner. The human organisation of space often creates methodological problems when solely utilising GIS outputs as it is not always uniform and requires contextualisation. The output of the GIS databases such as these have the potential to be an invaluable source for historians. However, like all sources the GIS output cannot be taken for granted. While the GIS database is accurate for the most part, there is little way to be certain that the original source has a 100 per cent accuracy rate and thus so leaves the GIS output open to similar interrogation.

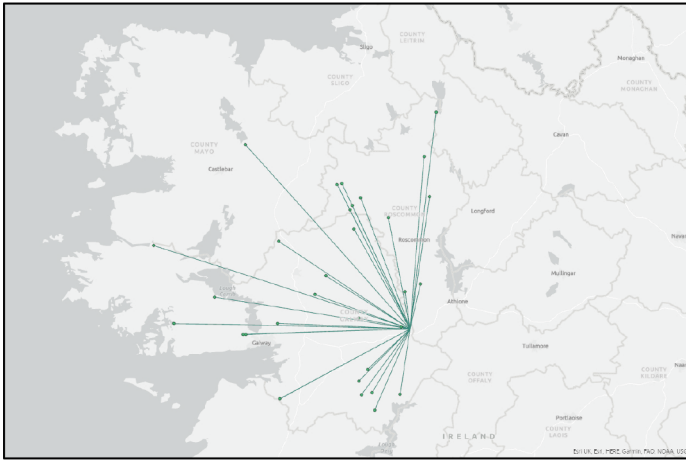


**Fig. 5.** Map displaying original residences of patients admitted to the CDLA.

Programmes such as ArcGIS have unique analysis tools that allow for the calculation of unique geographic factors that provide further insights into the expansive influence of the asylum. A key spatial factor that was derived using these methods was the distance from the asylum to the patient’s original residence. The co-ordinates of the CDLA were input as additional columns in the excel database. This allows a visualisation of the distance from the patient’s origin to the CDLA as seen in Fig. 6 below.

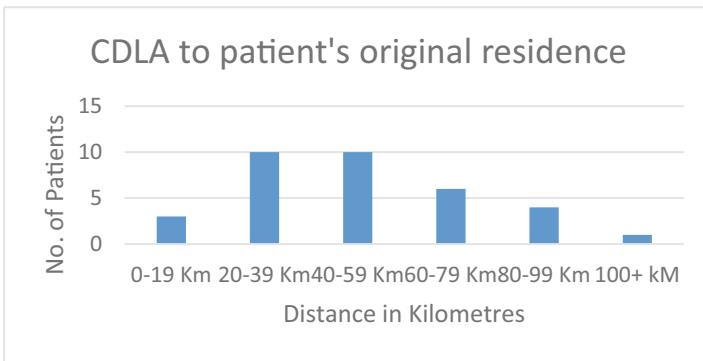
The visualisation of this data shows the influence of the asylum in Connacht. In July 1899 admissions arrived to the asylum from three counties. All of the admissions were rural, no patients were admitted from Galway city. Once this distance was input it was possible to use one of the spatial processing tools in ArcGIS to calculate the exact distance. Once this has been calculated it is possible to export the new data to an excel file to conduct further analysis. The initial findings of this to process show that the distance of admissions ranged from 3 kms to 102 kms as visualised in Fig. 7. The average patient was calculated to have resided 50 kms from the institution. The ranges included





**Fig. 6.** Map showing distance between the CDLA and patient’s original residence.

in Table 4 were chosen to ensure a clear visualisation of the demographic distribution while incorporating the minimum and maximum distance from which patients were admitted to the asylum.



**Fig. 7.** Distance to CDLA from Residence

## 5 Initial Findings

One of the striking findings in this sample is the lack of admissions from Galway city. This is highly unusual: there were relatively few large urban areas in the counties of Galway and Roscommon in this period, and Galway was the county capital. Moreover, male admissions to the Ballinasloe asylum were most often arrested some distance from their homes, and accused of offering violence to strangers. Women were on the other hand more likely to be admitted from their homes, as a result of alleged suicidal ideation

or violence towards immediate family members. Thus, men were frequently arrested in the city of Galway, where they had been observed behaving peculiarly in public, or had become embroiled in arguments in shops and at marts and fairs in the city. However, their place of residence was frequently recorded as elsewhere in the county, even when they had been living in the city for some time. This may be seen in the case of the 26-year-old Patrick T., arrested in Galway on June 8<sup>th</sup>.<sup>6</sup> Described as resident in the city where he worked as a labourer, he was arrested after he ‘unlawfully threatened to kill his Father’ following an altercation in Shop Street in the heart of the city. The magistrate who accepted the medical evidence that Patrick suffered from Mania and represented a danger to the public put his place of residence as Ahascragh, Co. Galway, that being the home of his father who swore evidence against him. This is not merely a confusion in terms of origin, but a deliberate elision of residence to avoid liability for Patrick’s support. Irish District Asylums were intended for the care of pauper lunatics, and each city and county paid for the support of inmates through taxation. It was therefore very important to correctly identify a patient’s place of residence in order to ensure that the appropriate Grand Jury, and after 1898, County or City Council, was charged for their upkeep.<sup>7</sup> By recording Patrick’s place of residence as Ahascragh – some 45 miles east of Galway – and not the city, Galway City Council could refuse to pay for his upkeep, and that expense fell on the County of Galway instead. This was a constant battle between local power bases, and for almost 20 years after the asylum’s opening Galway city had refused to pay for the support of city-resident patients. In a philosophical interpretation of the regulations that echoed the dual-state of Schrödinger’s cat, the city fathers argued that as Galway city was in the County of Galway, all city residents were therefore county residents, and appropriate charges on the county Grand Juries. It is when admissions are viewed through the ArcGIS programme that the absence of city patients becomes clear, and tensions between regions that are hidden in the record body are suddenly illuminated. Mapping in this case reveals lacunae, and offers a visual insight into the non-medical imperatives that were part of asylum admissions.

The mapping process used in this pilot also deepens our understanding of the unofficial adaptations made by asylum staff and the authorities to accommodate patients. By 1899, the catchment area of the CDLA had shrunk to two counties: Galway and Roscommon. However, in July, a patient from Co. Mayo was transported over a very considerable distance to Ballinasloe, despite not being legally entitled to be sent there. An argument was made for transportation on the grounds that the patient had originally been treated at Ballinasloe, and a full recovery was deemed more likely if they were permitted to return to a familiar therapeutic environment. A newly-built asylum had opened in 1866 at Castlebar, Co. Mayo, to serve individuals from Mayo, Leitrim, and

---

<sup>6</sup> To preserve patient anonymity, we use a standard historical methodology of identification in this paper of the individual’s first name, and first letter of their surname. The age of the patient is also provided, as it is important in establishing familial relationships and relative vulnerability to committal.

<sup>7</sup> Grand Juries were the earliest local authorities in Ireland (replaced in 1898 by County and City Councils). Responsible for local expenditure, they were largely staffed by local landowners who were also the key ratepayers, and who therefore tended towards the parsimonious when funding asylums and workhouses.

Sligo. Long-stay patients at the CDLA who had originated from those counties were transferred *en masse*, and despite the overcrowded nature of the Galway asylum, against the Board of Guardian's protests. The dynamics of asylum operations, and the status as well as financial boost that such an institution conferred in an impoverished region such as Co. Galway, meant that each Board argued strenuously for funds to expand their own institutions, and prevent the construction of new asylums that would divert patients to new locales. Using the mapping software to include areas not just beyond the asylum's catchment area, but further afield, will considerably enhance the existing historiography of migration as well as mental illness. For example, on June 20<sup>th</sup>, Catherine R was admitted for breaking furniture in her brother's house. Although recorded on her committal warrant as resident in Liss, Co. Galway, she was in fact newly returned from America, from where she had been 'sent home...as a dangerous lunatic.' The traffic of patients from Scotland, England, Wales, and the United States is a little-explored element in psychiatric history, but a fascinating one.

Examining the archive using locale as a prism alters our view of the admissions process. It also enriches our understanding of the precipitating factors in admissions. Each warrant includes information on the alleged lunatic action that brought the person to official attention, and a close examination brings gendered difference into sharper focus. There were 20 male patients and 13 female patients admitted to the asylum in June 1899. The June admissions show distinct differences between male and female admissions that reflect the limited social spheres women moved in, relative to men. On June 8<sup>th</sup>, for example, Maria M was accused of attempting 'to commit suicide by drowning herself and did assault [her father] by biting him' at their home. Kate L, also admitted on the 8<sup>th</sup>, 'did unlawfully assault [her father] and break the windows and furniture of his house', the 'same being a crime for which she would be liable to be indicted.' On June 12<sup>th</sup>, it was alleged that Catherine G 'occasionally becomes violent and pelts stones at her Father and others; she turned her Father out of his house at Cloonmullen, and she is now living there alone, to the danger of setting herself and the House on fire.' All of the women admitted in June were alleged to have attacked close family members within their shared home. The men, on the contrary, were more often accused of travelling distances from the domestic realm, and offering danger to non-family members as in the case of Michael C, who 'on several occasions leaves his residence and goes about the country, returning at night...[he] threatened violence towards his brothers and the Police.' Similarly, Peter J was admitted in May, discharged, and returned in June because 'he did...cruelly ill-use a mare his own property by running her though soft places and jumping her over walls. He also did at same place strip off all his clothing, except his shirt, went on top of the house, and put his legs down through the chimney of same, and afterwards jumped down off the gable, and [the informant] is in fear of his life.' Pinpointing the location of the alleged acts of insanity provides a nuanced picture of Irish rural society, and the perceptions of when mental illness was perceived as an accepted (if feared) part of the human condition, and when it became a matter of public order.

## 6 Conclusion

The frameworks and methodologies used throughout this study offer templates to conduct an expanded study on asylum admissions in the CDLA as well as other Irish asylums, and has the ability to broaden our current understanding of the position of the asylum system in society in the early twentieth century Ireland. This approach has the potential to significantly enhance the existing body of knowledge in relation to mental health histories, by facilitating a sophisticated understanding of the physical movements of individuals in and out of the institution, and thereby illuminating their engagements with family, community, and systems of state support and control. Most analysis of asylum records to date has focused on the patient experience inside the institution, or the operation of the judico-medical establishment. Combining historical analysis with digital interrogation of the records produces a complex picture of nineteenth-century life, demonstrating the intersectionality of class, gender, migration, and modernity.

There are clear benefits to using digital methods for the analysis of historical asylum records but it must be noted that the figures derived using these tools cannot guarantee complete accuracy due to a number of issues. At first, the original records may appear limited, once these are interrogated, they can reveal a wealth of information pertaining to an individual. A demographic analysis of these records can reveal the social class of a person. When used in conjunction with the census data it can reveal additional information regarding their family and their lives.

The study of asylum admissions can broaden the understanding of the landscape of the institution's expansive influence. This shows how new methods of analysis can be of benefit to historical research, particularly when these are used in conjunction with digital platforms such as GIS. Aside the capability to display data in a clean and accessible format, they can reveal a wealth of other information. Throughout this study it has been key to understand the spatial influence of the CDLA in this period. The use of GIS enabled the mapping of these patients' origins and with further expansion of the methods used in this research these districts could be specified to a broader and more accurate degree. The application of spatial analysis has proven to be beneficial throughout this study, showing how geography and one's mental well-being are evidently deeply intertwined. The use of the newest technology animates the oldest records, allowing us a nuanced insight into the lives and experiences of the asylum inmates and their communities.

## References

1. DeBats, D.A., Gregory, I.N.: Introduction to historical GIS and the study of urban history. *Soc. Sci. Hist.* **35**, 455–463 (2011)
2. Kelly, B.D.: "He Lost Himself Completely": Shell Shock and its Treatment at Dublin's Richmond War Hospital, 1916–1919. Liffey Press (2014)
3. Central Statistics Office, Dublin: Census 1841 and 1951
4. Finnane, M.: Irish psychiatry. Part 1: the formation of a profession. In: Berrios, G.E., Freeman, H. (eds.) *150 Years of British Psychiatry*, vol. 1, pp. 1841–1991. Gaskell/Royal College of Psychiatrists (1991)
5. Foucault, M.: *Madness and Civilisation: A History of Insanity in the Age of Reason*. Tavistock (1965)

6. Gregory, I.N., Healey, R.G.: Historical GIS: structuring, mapping and analysing geographies of the past. *Prog. Hum. Geogr. Hum. Geogr.* **31**(5), 638–653 (2007)
7. Hillier, A.: Invitation to mapping: how GIS can facilitate new discoveries in urban and planning history. *J. Plan. Hist.* **9**(2), 122–134 (2010)
8. House of Commons, Report on District, Local, and Private Lunatic Asylums in Ireland, 1900 Sessional Papers, 1901, Lunatic Asylums (1901)
9. Gregory, I.N., Geddes, A.: *Toward Spatial Humanities*. Bloomington (2014)
10. Kelly, B.D.: Folie á plusieurs: forensic cases from nineteenth-century Ireland. *History of Psychiatry* **20**(1), 47–60 (2009)
11. Kelly, B.: *Hearing Voices: The History of Psychiatry in Ireland*. IAP (2016)
12. Knowles, A.K.: The contested nature of historical GIS. *Int. J. Geogr. Inf. Sci. Geogr. Inf. Sci.* **28**(1), 206–211 (2014)
13. Prior, P.M.: Dangerous lunacy: the misuse of mental health law in nineteenth-century Ireland. *J. Forensic Psychiatry Psychol.* **14**(3), 525–541 (2003)
14. Roy, P., David, W.: *The Confinement of the Insane: International Perspectives, pp. 1800–1965*. CUP (2009)
15. Roy, P.: *A Social History of Madness: Stories of the Insane*. Orion (1989)
16. Edward, S.: *A History of Psychiatry: From the Era of the Asylum to the Age of Prozac*. Wiley (1997)
17. Travis, C., Ludlow, F., Gyuris, F.: *Historical Geography, GIScience and Textual Analysis : Landscapes of Time and Place*. Springer, Cham (2020). <https://doi.org/10.1007/978-3-030-37569-0>
18. Walsh, O.: Cure or custody: therapeutic philosophy at the Connaught District lunatic asylum. In: Preston, M.H., hÓgartaigh, M.Ó. (eds.) *Gender and Medicine in Ireland, pp. 1700–1950*. IAP (2012)
19. Walsh, O.: Gender and insanity in nineteenth century Ireland. In: Jonathan, A., Anne, D. (eds.) *Sex and Seclusion, Class and Custody: Perspectives on Gender and Class in the History of British and Irish Psychiatry*. Rodopi Press (2004)




**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Digitised Historical Sources and Non-digital Humanists: An Interdisciplinary Challenge?

Maelle Le Roux  and Anna Gasperini  

University of Galway, Galway, Ireland

{maelle.leroux,anna.gasperini}@universityofgalway.ie

**Abstract.** The digitisation of sources has opened new perspectives for humanities scholars. Digitisation allowed a larger access to sources, removing some financial and geographical limits, and the use of digital tools provided new perspectives for humanities scholars, who are able to read the sources differently. However, working with digitised sources also created new challenges that humanities scholars are not always equipped to overcome.

The ‘MedIcal Literature and Communication about Child Health’ (MILC) project uses historical medical books for a non-specialist audience to analyse discourses on children’s health in England, France and Italy between 1850 and 1914. Despite being born a non-digital humanities project, with a focus on manual qualitative analysis and a combination of history and literature methods, it took a digital turn when using digitised sources, with issues of digitisation and Optical Character Recognition (OCR) among others. The team working on the project is composed of three humanities scholars, with limited computer science skills. This required us to find digital humanities and in general IT tools adapted to our skillset, and suited to our needs. These tools did not always fit all our needs, and often presented issues in terms of accessibility and compatibility with the general standards of digital humanities.

Using examples from the issues faced by this project, and from the solutions found, this paper will argue that the challenges encountered by humanities scholars are interdisciplinary, not only because they overcome the traditional disciplinary boundaries inside the humanities, but also because they mirror challenges that computer scientists are working to solve. This paper will also argue that collaboration is a necessity which would benefit both humanities scholars and computer scientists in their work on the improvement and development of new tools, with the help of AI for example. Using the work done by a team of non-digital humanities scholars, it will argue that accessibility is a central issue in digital humanities and in the creation of IT tools, which needs to be addressed.

**Keywords:** Digital Humanities · Digitised Sources · Accessibility

## 1 MILC – A Non-digital Digital Humanities Project?

### 1.1 Introduction

MILC - MedIcal Literature and Communication about Child Health is an interdisciplinary Humanities project combining medical history and literature. It performs a transnational comparative analysis of childcare manuals written for a non-specialist

audience in French, English and Italian, focusing specifically on how the texts present the themes of breastfeeding, vaccination and physical education. The project focuses on texts published between 1850 and 1914, with 361 books and pamphlets composing the main corpus.

The project's methodology was originally planned to be non-digital, using qualitative methods traditionally adopted in the field of literary analysis, in which the project was grounded. This analysis was to be contextualised with a catalogue of national French laws collected through archival work, and a study of the translations of these popular medical texts, neither task being considered as requiring digital components. This article analyses the challenges the project presented once it took a more digital turn, changing shape slightly while still including all its original aims and methodologies. It also outlines the solutions that we adopted as the result of collaboration and communication between digital and non-digital scholars.

## 1.2 A Non-digital Methodology

The original, non-digital methodology for MILC, which is still part of the project, envisaged the manual close reading of the texts articulated in three phases. Phase one, recently concluded, was for data gathering, during which we built three "language" corpora – one of texts in English, one in Italian, and one in French – and one corpus of translated texts to help us identify the role of translations in the international circulation of knowledge about child health. The data gathering phase also envisaged two key-intermediate assessment steps (one for the language corpora and one for the translations corpus) to assess the quality and quantity of the material gathered and adjust the literary analysis performed in phase two accordingly.

Phase two, literary analysis, will examine a selection of case studies from each "language" corpus. These texts are not, technically, "literary" in the same way as, for example, novels are: they were handbooks, a series of childcare instructions for parents, guardians, and some professional categories. The project, however, analyses them as literary texts, examining their language and content, as well as the position of the author and that of the reader, against the background of the historical and cultural landscape in which they appeared. Specifically, their content is analysed using a historical social constructivist approach that considers medical knowledge as the product of cultural and social dynamics tied to a specific historical moment, and therefore bound to change over time [1]. The goal is to understand these texts as cultural products that contributed to circulating and creating meaning and knowledge about child health.

The last and final phase envisaged the transnational comparison proper, through case studies. During the project planning phase, it was envisaged that the work should be performed manually, as in, without the aid of digital tools. To an extent, this is still the main method used in MILC. However, as a synergy was created between the digital and non-digital sets of expertise of the different team members, we started realizing the potential digital material has for revealing different types of data than the ones that is possible to gather through manual close reading.

### 1.3 Using Digitized Sources

It was decided early in the project to use digitised sources to overcome the hurdle posed by the geographical distance between the different archives. Using digitised sources, in theory, allowed us to access them without the financial or time limitations that accessing and working on the physical versions required. Using online catalogues from major archival institutions in the countries the project focuses on, we identified a large corpus of sources, with 361 overall. We quickly noted the discrepancies in their digitisation.

While the French sources were mostly digitised and available online through the Bibliothèque Nationale de France (BNF) and its digitised sources database, Gallica [2], few digitised sources were available through the British Library, and even less from the Biblioteca Nazionale Centrale di Firenze (BNCF). We were able to collect digitised versions of some sources in our corpus from other online databases, with the Wellcome Collection especially [3], but the online databases focused mainly on French and English sources, with very few Italian sources available online. These elements caused the first shift towards a more digital methodology, to help us, first, gather digitized copies of the Italian material and, second, explore the possibilities offered by the analysis of what was immediately available to us: the texts' metadata.

### 1.4 A Digital Qualitative Analysis

The distant reading analysis of the metadata, especially the titles of the texts, would allow us to better understand the corpus globally, while also providing some information on the themes and audiences of the texts.

Exploring the options available to the group, we selected the Qualitative Data Analysis Software (QDAS) NVivo. NVivo is a commercial software developed by the Lumivero company [4, 5]. Its last version, NVivo 14, which the project uses, was released in 2023. NVivo is structured as a relational database, although it relies on a software-specific vocabulary, sharing few elements with standard relational databases vocabulary. We selected this software because it combined an excellent fit for the qualitative methodology of the project with immediate free access, training, and technical support for all team members through the University of Galway, where the project is being developed and where NVivo is widely used for qualitative research. The fact that using NVivo does not require specific technical skills further to the training provided made it also especially suited to a team of mixed digital and non-digital scholars, and indeed we were able to proceed immediately, after importing the metadata of the corpus in NVivo from the bibliographical software Zotero, with the 'coding' of the metadata. In this process, the book titles and other elements such as the authors and places of publication were annotated based on pre-determined criteria ('codes') that grouped the texts based on elements such as intended reader, themes, and the vocabulary they used.

Format-wise, NVivo produces projects in a proprietary format, '.nvp', which is not compatible with other databases formats. The results of the analysis are exportable in Excel and CSV formats, which allowed us to conduct some data analysis and create visualisations based on the results of the distant reading of the titles.



## 2 The Digitised Sources

### 2.1 Creating a Corpus from Digital Catalogues

The project intended to create its own corpus to analyse, which was, in itself, a challenge. Indeed, no prior study had catalogued childcare manuals, which required us to identify and collect the sources by researching various online catalogues, namely Worldcat, the BNF, the BNCF, the British Library and the Wellcome Collection [6].

As noted by Blaney et al., the search process is a part of the methodological process that is difficult to document, and rarely reproducible [7]. The reproducibility of searches was an issue we encountered in the project, whereby we met keyword searches problems of the same kind documented by Hitchcock on the topic, with a potential lack of accuracy leading to a larger number of irrelevant texts [8]. This aspect was further complicated by the project's transnational framework. While we aimed for maximum consistency, translating keywords or finding the closest possible alternatives across the three languages, these translations were not accurate in every context.

A further, if opposite, issue emerged when we used the categories available in the online catalogues to identify texts which supposedly were related to the project. Counterintuitively, searches by category returned very few texts, which we could note due to the absence of key-texts we found through other modes of research. Little information exists on the definition of these categories in online catalogues, with limited accessibility to the details behind the search engine and to the original metadata of the texts, making it impossible to verify the accuracy of our search.

The solution we adopted to overcome these challenges was using a manual combing method through the keywords, combined with the use of historiography and secondary readings to identify various other sources we had missed in the first version of the corpus. The result was a corpus composed of 97 sources in the English, 159 in the French and 105 in Italian, which we determined to be balanced and sufficient for the analysis we intended to conduct. This combination of traditional search methods, secondary readings, manual combing of the catalogue, and keyword searches, had the added benefit of allowing us to identify the key-texts in the corpus, as in, the most important specimens of childcare manuals produced in the time span examined.

### 2.2 Digitised Sources and OCR

Besides the matters related to research through digital catalogues, the other main challenge we faced in MILC was related to Optical Character Recognition (OCR). OCR issues, as in, whether or not texts are machine-readable, are common for humanities scholars working with digitized sources. Blaney et al. noted the variety of standards available with digitized texts, depending possibly on the goal of the institutions digitizing the texts, with some focusing on their preservation and their availability to a larger number of readers, and others focusing on the compatibility of these newly digitized texts with new standards in digital texts availability [9]. Since MILC had not been planned as a digital project, we had to work with comparatively limited resources when it came to solving OCR issues. The resource that proved most effective among the ones we could access was the ABBYY PDF FineReader 15 software, on which we heavily relied for

some countries, less for others, based on the differences in digitization policies across the archives.

### **The French Sources – An OCR-Oriented Digitization Policy**

The French sources were mostly pre-OCR'd, as they were made available by Gallica. The institution uses a combination of internal and non-specified external systems to OCR its sources, and has been involved in the research and development of new OCR tools [10]. A sample testing on the texts showed that their OCR was more efficient than other tools we had access to, and we did not modify them in any way.

Gallica, in its OCR policy, indicates aiming for its OCR to have an efficiency rate of 96 per cent on texts, although it acknowledges that this rate does not apply to all the texts due to some being digitised prior to this policy, and due to this policy not being applied to 'numbers, tables, unreadable sections, adverts' [11]. Gallica also states that while a high rate of accuracy might be indicated, this rate might have been calculated excluding pages where the OCR would have encountered difficulties, making the reading of the document less efficient than announced. Furthermore, the quality of the OCR has an impact on the search function and possibly impacted the corpus itself. Indeed, Chiron et al. noted the higher error rates of Gallica's OCR on named entities, and indicated that this impacted the search function for users searching for proper nouns, such as people's names, while being difficult to clean post-OCR due to the absence of some of these names in traditional dictionaries [12]. While searching for specific texts, we encountered similar issues, as the search engine did not retrieve some of the texts we were searching for, despite us knowing of their presence on Gallica through exterior sources. We were able to find the texts through the authors' pages, indicating that while the text had a correct metadata, the search engine seemed to focus on the OCR of the texts which had not read the authors' names correctly.

These issues had a limited impact on the project itself, beyond suggesting the potential existence of more digitised texts that could have been added to the corpus; had we proceeded to scale up the project with a social network analysis, which we considered at one stage of the project before discarding the idea due to time constraints, it would have been a different matter. Overall, while noted, these OCR issues did not impact the quality of the corpus, which we deemed sufficient for the distant and close reading work we intended to do.

### **The English Sources – Dealing with Limited OCR Within an Image Preservation-Oriented Digitization Policy**

Most of the English texts we collected were digitised by the Wellcome Collection, a private library and museum focusing on the theme of medicine which is part of the Wellcome Trust charity. Unlike the texts available through Gallica, the texts we collected through this database were not pre-OCR'd. Even though it does mention OCR a few times in their digitization policy [13], from the document emerges that the main goal of digitisation at the Wellcome Collection is to preserve a copy of the historical text and to make it accessible online to users who cannot access the physical library. Both the policy and the presentation of the digitised sources would indicate that each page of the texts is perceived as an image, rather than as text.

To overcome this obstacle, we used ABBYY PDF FineReader 15 to OCR the sources available to us from this collection. The software, and this specific version, was chosen because it is recognised as a standard in the field of non-professional OCRing historical sources [14, 15], but also for its convenience as we had access to a licence of the software and we could rely on previous experience in using it to OCR digitised historical sources. Our choice of version was also guided by availability of funds. Since ABBYY PDF FineReader is a commercial proprietary software and the latest version, FineReader 16, was released in 2023 on a subscription model, we decided to use version 15, released in 2020. The OCR model of the software is not accessible due to its proprietary nature, and few details are available on it. The software itself emphasises its use of Artificial Intelligence (AI), with Machine Learning especially, and indicates the advances made by each new version using the latest research in the field [16].

The use of a previous version of the software to OCR the sources in this project can be considered an issue, as it means that we did not use the latest technology available and that potentially we could have produced better OCRed documents using the latest version. However, there were multiple reasons for choosing this version: first, our analytical approach did not require perfect accuracy in words recognition by the software, as would have had a qualitative approach such as corpus linguistics; second, we did not intend to share the texts of the full digitised corpus for copyright reasons. These first two elements allowed us a certain degree of flexibility in opting for a lower degree of accuracy in the OCR. Finally, as with NVivo, we had to account for the fact that the digital element was integrated in the project at a later stage, which made budget a key-element to take into account in our choice. As with NVivo, ABBYY PDF FineReader combined cost efficiency with suitability to our purposes, as one of the team member had access to a licence. Therefore, we OCRed the texts using mainly automatic settings in ABBYY PDF FineReader with limited manipulation.

### **The Italian Sources – Dealing with Limited Digitization**

The Italian sources presented the greatest challenge regarding OCR. The BNCf, like other national Italian libraries, has a digitization policy and regularly collaborates with institutions and projects to digitize and make accessible some of its collection [17, 18]. A governmental report in 2016 indicates that the BNCf intends to pursue this digitization work to improve accessibility of its collection, while following the current standards of the field [18]. Due to the quantity of material in their collection, they had not been able to digitize most of the texts in our corpus, possibly due to their lack of popularity amongst researchers and readers, compared to other documents. We started the process of having them professionally digitized and, in the meanwhile, we proceeded with manual digitization to be able to start the close reading work.

The documents were photographed page by page by members of our team visiting the BNCf using a phone camera. Blaney et al. notes that this is a common method of digitization amongst historians, as due to the necessary selectivity of institutions in digitizing sources, as well as financial constraints from these same institutions and from researchers in the case of digitized documents being behind a paywall, it is common for historians not to have access to a digitized version of the documents they intend to use [18]. The photographs were directly transformed into a PDF using OneDrive, as the program organized and backed up the documents. This method, while cost-effective,

was not without its pitfalls. First, the manual photography method for digitization is time consuming for the researcher, and with the lack of proper photography equipment, can represent a physically difficult task. Secondly, this digitisation method also presents problems in terms of long-term preservation of the data, as the most popular photographs formats are formats such as JPEG or PDF, and not TIFF, which is the recommended format for long-term preservation of digitized sources. Finally, and most importantly, this method of digitization produces images whose quality may vary significantly.

A researcher without specialist skills in photography will hardly perceive this difference in quality while taking the pictures; however, it will emerge and potentially present challenges during the OCRing process. These difficulties are of three kinds: blurriness; variety of frames and light; and quality of the original material. Images captured during our research trips were at times blurry and, while still readable by human eyes, they were partially unreadable by the OCRing software. As Taş and Müngen noted, pre-processing the images can improve the OCR results for historical sources [19]. Image processing through OCRing process in ABBYY PDF FineReader 15 allowed us to an extent to correct the blurriness of the photographs and to improve their overall quality. In our case, pre-processing was further complicated because, compared with professionally digitized sources, the photographs taken manually had variable frames, orientations and lighting.

After processing the PDF in the software and conducting a first OCRing of the document, we proceeded to more targeted interventions on pages in which the software had been unable to read the text. The software allowed us to process images individually or per document; the variable quality of our pictures made us opt for processing per individual image. The most common issues encountered were defocus blurs, especially side-ways due to the book format making the surface uneven, and making lines of text askew, which was often fixable with the software's tool to align the text lines, or to reorientate the page in the photograph. In instances where blurriness of the picture affected only a small portion of text, we simply corrected the OCR text by hand. There were cases instead in which some or all of the text was fully unreadable, either because of the blurriness of the picture, or because of the quality of preservation of the original document. Indeed, a major issue for OCRing is the quality of the original paper and ink, associated with potential preservation issues creating stains in the paper [8, 9]. While this is usually an issue associated with historical newspapers, we can assume that childcare texts, which aimed to be financially accessible to a relatively large public, were sometimes made with lower quality materials, resulting in these issues. Furthermore, accidental flooding of the archives in the middle of the twentieth century caused damages to some of the books in our corpus, causing further readability issues.

Overall, while this manual version of the Italian corpus is digitized and OCRed, it is only partially, and while we were able to solve some of the issues, we faced limitations that we did not have the resources to overcome.

The aim of MILC is to analyse these texts comparatively, with a transnational perspective, to understand the nuances in child health discourses they present based on context. Since the methodology for MILC always intended to focus on case studies and did not necessarily require using all the texts in a full comparative approach, the issues

we encountered with the digitized sources had a limited impact on the project. However, the potential impact of the issues encountered when working with digitized sources on a project should be noted, especially in a context of comparative approaches, in a transnational or global history perspective.

### 2.3 Open-Source Alternatives for OCR and Non-digital Humanists

As discussed in the previous section, we decided to use ABBYY PDF FineReader 15 to OCR the texts in our corpus for multiple reasons, one of them being that we had access to a license through a member of the team. This caused some concerns regarding the long-term access to OCR methods, and overall the reproducibility of the research. Indeed, since ABBYY FineReader is a commercial proprietary software, we do not have the details of its OCR, making it impossible for other researchers to reproduce the methodology unless they themselves had access to the same version of the software. Furthermore, the use of a license associated to an individual member of the team could cause some issues in the long term, if the team member were to stop working for the project and the project required other texts to OCR. This would therefore require finding a different OCRing tool and defining a new process. With these issues in mind, we considered other OCRing tools that could be used by all the team members and by other researchers without constraints of cost or technical skills.

As we looked into open-source OCR software, we noted that studies indicated Tesseract to be the main open-source alternative in the field and decided to experiment with it, to see if it would be a suitable alternative [15]. Despite fitting our requirements for being open-source and free, it did require technical skills, rendering it difficult to use by our team of non-digital humanists. Tesseract was developed at HP before being released under an Apache license [20]. The software can be installed directly on a computer or run through an Application Programming Interface (API). We did not experiment with API in this project and worked exclusively with Windows OS. The installation of the software is possible through files in Github or through an installer developed by the Mannheim University Library (UB Mannheim), which is the process we decided to follow as an installation through GitHub files required more technical skills [21]. As noted in the documentation, Tesseract does not have a graphical user interface (GUI), which meant that there was no front-end visual as support for the user. This required us to use a command line interface, which is a complex tool to use for non-digitally trained researchers. Indeed, command line interface requires a good understanding of the logic involved with computer languages, which is rarely ever part of non-digital humanists training. While it is possible to find documentation and tutorials online which explain how to use some basic functions, use of this software remains, mostly, entirely inaccessible without extensive specialist training. We attempted to find a suitable GUI through the ones developed by third parties [22], but this attempt had limited success, with multiple GUIs being difficult to install or difficult to use without specific technical skills.

Overall, the OCRing of the texts was the biggest challenge we encountered in the digital aspect of the project, as it had multiple ramifications and our methodology depended on having access to a corpus of OCRed texts in three various languages, digitized in different contexts and with different methods. While we aimed to be consistent in the

OCRing process, and we aimed to find a solution to the non-reproducibility associated with the use of commercial proprietary software, there were none suitable for a team of non-digital humanists.

### **3 Integrating the FAIR Principles into a Non-digital Native Project**

#### **3.1 From Corpus to Dataset**

The Findable Accessible Interoperable and Reusable (FAIR) data principles are central in research, and especially in digital humanities and computer science due to the creation and use of datasets [23]. The MILC project was planned with these principles in mind, and the introduction of the digital element gave a different inflection to its FAIR approach. Indeed, while the corpus originally intended to be a non-digital object, the use of various digital tools to catalogue it and analyse it transformed it into a potential dataset, which led us to reflect on how to integrate this new component in the project.

In the original planning of MILC, the catalogues were to be made accessible at the end of the project. These catalogues would have been tables with the metadata of the texts composing the corpus, but also of the French laws and their relevant texts. This took a different turn when we started using NVivo to use the metadata of the corpus into the distant analysis, but also when the catalogue of the French laws became more complex and we decided to organise it as a relational database using Access. With the change in methodology for both catalogues came the issue of the sources themselves, either institutionally digitised or privately digitised, which were directly associated with the items in the catalogues. Indeed, to avoid issues of copyright and rights of use of the digitised sources, it was decided early on not to release the sources themselves. As these sources have various origins, with various legislative frameworks, not releasing any source was easier than releasing only a portion of the corpus. Therefore, while the metadata of the dataset is releasable, the full corpus will not be released.

#### **3.2 From NVivo to FAIR-Compatible Data?**

Since the work conducted on the corpus through NVivo became an important part of the analysis, it was necessary to consider how we could make this data potentially available to researchers in a FAIR-compatible format. FAIR-compatible format here refers to interoperable formats usable by most software and Operating Systems, such as XML here in the case of the corpus. NVivo, as a commercial software, has its own proprietary formats, which are not interoperable. This caused us some issues when we attempted to extract the full dataset with its encodings from the software for data analysis, as this was not an option it offered.

This led us to find that NVivo projects could be exported in an interoperable format compatible with other QDAS, the REFI-QDA format, which is based on, and easily convertible to, the XML format [24]. This format is mainly used in transferring a project from a QDAS to another similar software, and so far the tests we conducted with open source alternatives did not bring satisfactory results. Another issue is that the format, and the XML file resulting, are structured around the texts used to create the project.

Therefore, making available this data in XML or other formats would necessitate making the full corpus available, which is not an option due to copyrights issues.

This issue shows the limits encountered by the project in fitting with the FAIR principles, since the corpus and part of the analysis were not expected to become potentially sharable data.

## **4 Conclusion – Learning from Non-native Digital Humanities Project for the Future of the Field**

### **4.1 Learning from MILC – Is There Such a Thing as a Non-digital Humanities Project? Digital Humanities and the Accessibility Question**

The issues the MILC team encountered are fairly normal for a Digital Humanities project. The real challenge was that they were not anticipated, because the project was perceived as non-digital, leading to limitations in the resources available to face these challenges. Despite these challenges, we were able to create datasets in the project which fit the standards of the field, and partly fit the FAIR principles.

In the early phases of the project, as a team we reflected on whether MILC should be called a Digital Humanities project, or a project using Digital Humanities tools, which led us to question the difference between these two definitions. We concluded that, even though it does not lead to any significant innovation in Digital Humanities, MILC is a Digital Humanities project because it uses digital sources, and uses digital tools to clean, structure and analyse them.

By adopting this broad definition of Digital Humanities, and considering the fast-paced development of new tools and the advancement of the digitisation process, we can envisage that most, if not all, humanities project will take a digital turn, similarly to MILC. Learning from this non-native Digital Humanities project, and the challenges we encountered in its transformation to the digital, Humanities researchers must plan for projects that might need higher digital skills than anticipated, and therefore account for the necessary interdisciplinary collaborations.

The question of accessibility was central in our methodological process, especially when it came to the question of selecting software. As classically-trained humanities scholars, we had to select software and tools that were easily accessible to us. While training was an option, we relied on training that was easily available to us, and that would allow us to quickly become proficient in using the tool. Since this was a minor aspect of the methodology, we did not have the resources and time necessary to focus on general computer science training.

These accessibility and training matters necessarily limited our choice, resulting in the use of tools that would complicate the process to make the data FAIR. The open-source or FAIR compatible alternatives did not fit our skills or the project's needs. Therefore, we would say that there is a need for more accessible Digital Humanities software, as an increasing number of researchers in the field come from non-Digital native background and do not have the technical skills that most open-source software require at the present.

## 4.2 Creating Accessible Tools - The Importance of Collaboration

MILC, while interdisciplinary, was born as a Humanities project, without a digital aspect. Its conception as a non-Digital Humanities project made collaboration with the field of computer sciences more difficult, as no computer scientists were involved, and no resources were allocated to such a collaboration, which was not envisaged in the initial layout of the project. However, MILC was thought as a stepping stone project, leaving the possibility for future researchers to scale it up as a Digital Humanities project. By using the latest advances in Artificial Intelligence and Machine Learning applied to OCR the texts, therefore creating a more accurate dataset, researchers could then apply various tools to the corpus in order to provide a broader analysis using both qualitative and quantitative methods. The data we have accumulated would benefit from applying a social network analysis method, and the multilingual corpus would be an excellent source for language analysis through corpus linguistics and Natural Language Processing, with named entities analysis especially.

Digital Humanities is by its nature an interdisciplinary field, and collaboration is central. First and foremost, it is central in the production of more accessible tools, as this accessibility can only happen by a communication process between non-digital native humanists and computer scientists. These non-digital researchers are part of the future of Digital Humanities, and they are the potential users of the tools that computer scientist researchers are developing. Making accessibility a central part of this development and facilitating a dialogue between computer scientists and non-digital scholars would ensure that the innovative tools supporting Digital Humanities research are strategically tailored to its needs.

Another important step would see archival institutions, humanities scholars and computer scientists collaborating to establish standards in the process of digitising sources and making them accessible, including through retrospective work on sources digitised using outdated methods and tools. As AI and Machine Learning progress, computer scientists are creating tools that are used by some archival institutions in their digitisation process, but not by all, and archives tend not to provide information about this aspect. Consequently, humanities scholars do not have sufficient data to ascertain whether the quality of the digitised source they intend to use is sufficient for their purpose.

The development of standards for Digital Humanities must be done collaboratively, to ensure that they are both sustainable and accessible by researchers without formal computer science training, while still promoting good practices for the long-term contribution to knowledge and research. For example, the achievement of the FAIR principles, amongst other standards promoting the release of a fully reusable dataset to other researchers, may be out of reach for scholars with limited computer science training, due to technical difficulties, legal complexities, or the absence of accessible compatible tools. A reflection on these issues, in collaboration with archival institutions and researchers in computer science and Humanities is necessary to promote these standards of open-data in the Humanities and help the field of Digital Humanities navigate this turn of non-native Digital Humanities project.

**Acknowledgments.** The research conducted in this publication was funded by the Irish Research Council under award number IRCLA/2022/3921.



**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Lupton, D.: *Medicine as Culture. Illness, Disease and the Body*. 2nd edn. Sage, New York (2012)
2. Gallica Homepage. <https://gallica.bnf.fr/>. Accessed 09 Dec 2023
3. Wellcome Collection Homepage. <https://wellcomecollection.org/>. Accessed 09 Dec 2023
4. NVivo product page. <https://lumivero.com/products/nvivo/>. Accessed 09 Dec 2023
5. Jackson, K., Bazeley, P., Bazeley, P.: *Qualitative Data Analysis with NVivo*, 2nd edn. Sage, New York (2018)
6. WorldCat Homepage. <https://search.worldcat.org/>. Accessed 09 Dec 2023
7. Blaney, J., Winters, J., Milligan, S., Steer, M.: *Doing Digital History*. Manchester University Press, Manchester (2021)
8. Hitchcock, T.: Confronting the digital: or how academic history writing lost the plot. *Cult. Soc. Hist.* **10**(1), 9–23 (2013)
9. Pardé, T., Jacquot, O.: *Les humanités numériques à la Bibliothèque nationale de France*. Paris, Culture et recherche (2016), hal-01379908
10. Gallica: Mode texte et OCR. <https://gallica.bnf.fr/edit/und/consulter-les-documents#Mode%20texte%20et%20OCR>. Accessed 09 Dec 2023
11. Chiron, G., Doucet, A., Coustaty, M., Visani, M., Moreux, J.-P.: Impact of OCR errors on the use of digital libraries: towards a better access to information. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada, pp. 1–4 (2017). <https://doi.org/10.1109/JCDL.2017.7991582>
12. Wellcome Collection Digitisation Strategy, 2020–2025. [https://wellcomecollection.cdn.prismic.io/wellcomecollection/0047856d-bba9-4ab2-81b6-a270f887a8fb\\_WC+Digitisation+Strategy+2020-2025.pdf](https://wellcomecollection.cdn.prismic.io/wellcomecollection/0047856d-bba9-4ab2-81b6-a270f887a8fb_WC+Digitisation+Strategy+2020-2025.pdf). Accessed 09 Dec 2023
13. Volk, M., Furrer, L., Sennrich, R.: Strategies for reducing and correcting OCR errors. In: Sporleder, C., van den Bosch, A., Zervanou, K. (eds.) *Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing*, pp. 3–22. Springer, Berlin, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-20227-8\\_1](https://doi.org/10.1007/978-3-642-20227-8_1)
14. Tafti, A.P., Baghaie, A., Assefi, M., Arabnia, H.R., Yu, Z., Peissig, P.: OCR as a service: an experimental evaluation of google docs OCR, tesseract, ABBYY finereader, and transym. In: Bebis, G., et al. (eds.) *Advances in Visual Computing*. ISVC 2016. Lecture Notes in Computer Science(), vol. 10072, pp. 735–746. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-50835-1\\_66](https://doi.org/10.1007/978-3-319-50835-1_66)
15. ABBYY FineReader PDF: Powered by AI. <https://pdf.abbyy.com/blog/finereader-powered-by-ai/>. Accessed 09 Dec 2023
16. BNCF website: Categorie delle Risorse: Collezioni digitalizzate. [https://www.bncf.firenze.sbn.it/categoria\\_risorse/collezioni-digitalizzate/](https://www.bncf.firenze.sbn.it/categoria_risorse/collezioni-digitalizzate/). Accessed 09 Dec 2023
17. Lucarelli, A.: Web dei dati alla biblioteca nazionale centrale di Firenze. *DigItalia*, **10**(1/2), 30–39 (2016). <https://digitalia.cultura.gov.it/article/view/1471>
18. Taş, İ.Ç., Müngen, A.A.: Using pre-processing methods to improve OCR performances of digital historical documents. In: 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, pp. 1–5 (2021) <https://doi.org/10.1109/ASYU52992.2021.9598972>
19. Tesseract Documentation. <https://tesseract-ocr.github.io/tessdoc/Installation.html>. Accessed 09 Dec 2023

20. Tesseract at UB Mannheim. <https://github.com/UB-Mannheim/tesseract/wiki>. Accessed 09 Dec 2023
21. GUIs and Other Projects Using Tesseract OCR. <https://tesseract-ocr.github.io/tessdoc/User-Projects-3rdParty.html>. Accessed 09 Dec 2023
22. Wilkinson, M., et al.: The FAIR guiding principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
23. Müller, A.: From QDA to XML: The REFI-QDA project exchange standard. <https://methodos.hypotheses.org/1707>. Accessed 09 Dec 2023






**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Using Passive Sensing to Identify Depression

Evi Zafeiridi<sup>1</sup>  , Malik Muhammad Qirtas<sup>1</sup> , Eleanor Bantry White<sup>2</sup> ,  
and Dirk Pesch<sup>1</sup> 

<sup>1</sup> School of Computer Science and Information Technology, University College Cork, Cork, Ireland

EZafeiridi@ucc.ie

<sup>2</sup> School of Applied Social Studies, University College Cork, Cork, Ireland

**Abstract.** Depression is a common mental health issue that affects people's thoughts, behaviours, and feelings. However, depression can often be under-diagnosed or under-treated. Early identification of depression can help to reduce the severity of the condition. Passive sensing, which captures data through mobile applications and wearable devices, has been shown effective in monitoring and identifying mental health problems, including depression. In line with the scope of AISoLA for Digital Humanities to explore the challenges and opportunities of interdisciplinary action to develop better practices in research, this paper explores the efficacy of passive sensing through mobile applications and fitness trackers to identify signs of depression among 52 adults in a three-week study. Sensing data captures calls, text messages, locations, nearby devices, usage of social applications, physical activity, sleep duration and quality through the AWARE and FitBit applications. The paper also investigates differences in the behaviour between people without depression and people with symptoms of depression, and it explores which sensor data can help to accurately identify depression. The results show high accuracy of certain sensing data to identify symptoms of depression. Depression is associated with reduced physical activity, higher sleep efficiency, increased number of incoming calls, increased number of visited places and reduced application use. Differences between behaviours show that people with symptoms of depression are less active, have a higher sleep score and receive more calls compared to people without symptoms. These findings should be interpreted within the methodological issues that are discussed in this paper in relation to wider research in sensing technology that aims to identify and monitor depression, including small sample sizes and lack of information about participants.

**Keywords:** Passive sensing · smartphones · fitness trackers · depression

## 1 Introduction

### 1.1 Symptoms of Depression

Depression is a common mental health condition that affects people's thoughts, feelings and behaviour [1]. It is estimated that approximately 280 million people across the world have depression [2]. Common symptoms of depression include sleep and appetite

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 132–143, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_9](https://doi.org/10.1007/978-3-031-73741-1_9)

disturbance, and lack of interest or pleasure [3]. Depression is often not diagnosed in a timely manner [4], but early detection and intervention of depression can help to avoid many negative impacts of depression [5].

According to the Diagnostic and Statistical Manual of Mental Disorders [DSM5], depression is associated with lower physical activity, and mood and sleep disturbance. A study by Kaneita et al. [6] showed that people who sleep less than 6 h or 8 or more hours per day are more likely to be depressed compared to people who sleep between 6 and 8 h. This study also showed that sleep sufficiency was decreased in people with more symptoms of depression. Changes in physical activity are another behaviour that is associated with depression. A review of the long-term effects of depression on physical activity [7] showed that depression was associated with reduced physical exercise. Social disconnectedness and perceived isolation are also predictors of depression and have a bi-directional relationship with their influences [8]. Passive sensing has been shown to be a promising approach to identifying mental health problems, including depression [9–11]. This paper focuses on the identification of symptoms of depression through passive sensing.

## 1.2 Passive Sensing to Identify Depression

A common way to passively capture data to identify mental health issues is through mobile phone applications and wearable devices, such as fitness trackers. This approach requires minimum effort from participants. Some of the behaviours that are associated with depression, such as sleep disturbance, can be captured through the sleep assessment algorithm integrated into the smartphone applications of wearable devices such as fitness trackers or smart watches. An indicative example is the study by Narziev et al. study [9], who used smartphone and smartwatch applications to identify depression in 20 participants through behaviours associated with depression, including physical activity, sleep levels, mood and food intake. The results from this study showed that behavioural data from passive sensing through smartphones and wearable devices was correlated with participants' activity, sleep and mood as they were assessed through self-reported data. Another study by Wang et al. [10] used the StudentLife dataset to explore whether a smartphone application can identify different outcomes for 48 university students. The authors found correlations between depression and sensing data, especially for sleep, communication and location data, meaning the number of places that people visited during the ten-week study. A two-week study with data from 28 people showed a relationship between symptoms of depression, phone use and visited locations [11].

Accelerometers and global positioning systems sensors are some of the most widely used smartphone sensors in passive sensing for health and well-being; however, a systematic review by Trifan et al. [12] highlighted a possible gap between smartphone outcomes and clinical knowledge in this field, and the need for higher user engagement and further validation of the technology that is under investigation, including larger sample sizes. These methodological issues in research may explain varying results in studies. For example, in contrast to the study by Wang et al. [10], other research evidence suggests that mobility that is measured through GPS location tracking is not associated with depression [13]. Issues arise from converting raw data into summaries, known as features, especially when using existing technology in research studies that provide only

featured data instead of raw data. The frequency with which data is collected and sparsity because of missing data are also possible explanations for different research findings [13]. Despite those issues, passive sensing through smartphone applications and wearable devices has the potential to improve the monitoring and management of mental health conditions [14].

The present paper analyses data from our study at University College Cork to explore the accuracy of passive sensing. We collected behavioural data for people's physical activity, sleep, and communication patterns, to identify depression. Another aim of the paper is to assess which behavioural features can accurately identify depression. The paper also investigates behavioural differences between people who experience symptoms of depression with people without depression, such as differences in sleep quality. While most of the past research recruited young university students as participants, this paper uses participants from a broader age group and collects data using numerous sensors that have been shown to effectively identify symptoms of depression. Section 2 in this paper introduces the methodology of our study. Section 3 presents the results and Sect. 4 discusses the findings in light of past research while considering limitations and opportunities for further research in line with the scope of AISoLA to contribute to better research practices.

## 2 Methods

### 2.1 Data Collection

We collected data for three weeks using the AWARE smartphone app [15] and a FitBit fitness tracker and associated smartphone app (Table 1). Both applications have been used in research on passive sensing for monitoring different social and health issues [16]. The authors used both applications in order to collect both physical sensor data and communication and social connection data from participants. Data from both applications were collected for three weeks to allow sufficient time to collect enough data. Past research collected data between two to several weeks. In a study by Saeb et al. [11], the researchers collected data from 40 participants for two weeks, 28 of which provided enough data from sensors for analysis. The three-week duration of the present study was decided because of this possible lack of sufficient data and to avoid issues with recruiting enough participants for a longer study.

Data for participants' calls, text messages, changes in location, nearby Bluetooth devices and use of social applications was collected through AWARE. A FitBit Inspire 2 fitness tracker was used to collect data on participants' physical activity and sleep. Participants were asked to keep their Bluetooth and Wi-Fi activated for the duration of the study. Data was automatically collected without effort from participants. We installed the two applications on participants' smartphones, and they were asked to wear the fitness tracker for three weeks. Participants were provided with instructions and a user guide showing them how to charge their fitness tracker.

### 2.2 Participants

Data was collected from 52 adults, 18 years old or older using a random and non-clinical sampling strategy. Table 1 shows the minimum and maximum values of sensor data.

Participants' demographic information is presented in Table 2. Data on participants' demographic characteristics was also collected, including their age, gender, employment, level of education, and marital status. At the beginning of the study, we asked participants to complete the short version of the Geriatric Depression Scale (GDS) [17] to assess feelings of depression. The short form of GDS includes 15 questions about people's moods, activities, and feelings to be answered with Yes/No responses. GDS scores vary from 0 to 4, indicating no depression symptoms; 5–8, indicating mild symptoms of depression; 9–11, indicating moderate depression; and 12–15, indicating severe depression. For the purposes of this study, we grouped people with GDS scores from 0 to 4 in a group of people without depression, and people with GDS scores from 5 to 15 in a group of people who experience depression symptoms. This categorization was decided because we are interested in identifying depression at an early/mild stage. This study gained ethical approval from the Social Research Ethics Committee of University College Cork (number 2021–249) and complied with the General Data Protection Regulation.

**Table 1.** Sensor data from AWARE app and FitBit tracker

Application	Sensor data	Depression symptoms	No symptoms
		Min-Max values in a day	Min-Max values in a day
AWARE	Incoming calls (number)	0–9	0–7
	Incoming calls (duration)	0–4910 s	0–6002 s
	Outgoing calls (number)	0–19	0–21
	Outgoing calls (duration)	0–6425 s	0–6358 s
	Received texts	0–19	0–53
	Sent texts	0–9	0–9
	Locations (number of changes in location)	0–318	0–508
	Bluetooth (detecting nearby devices)	0–4502	0–4180
	Application usage	0–343	0–300
FitBit	Physical activity (duration)	0–793 min	0–931 min
	Sleep duration	119–750 min	46–714 min
	Sleep quality score	82–100	15–100

### 2.3 Data Processing and Analysis

For sensor data from the AWARE application, the total performance of each behavioural feature was calculated for the three-week study duration. We calculated the sum of incoming and outgoing calls and their duration in seconds, the sum of received and sent text messages, and the number of Bluetooth devices that were detected near each participant to provide a measure of how many other people our participants encountered during the day. The number of changes in location was calculated using the longitude and latitude of each location. Changes in locations were identified if the distance between them was at least 400 m and people spent at least 300 s in each location to avoid traffic delays being recorded as transitions [18, 19].

Application usage was measured based on the number of times that participants used a social networking application, including Whatsapp, Skype, Messenger, Instagram, Viber, MS Teams and Zoom. For data from the FitBit application, we calculated the duration of physical activity and the sleep duration across the three weeks of the study. The sleep efficiency score was generated by the FitBit app based on participants' heart rate and movement during sleep utilising the Fitbit sleep score algorithm. Missing values were replaced with the mean of each behavioural feature for both groups of people with and without symptoms of depression.

A logistic regression was used to assess the accuracy of the overall sensing data to identify depression as a binary outcome and to show the correctness of the results in the model. Behaviour differences between the group of people without depression and the group of people who experience symptoms of depression were assessed using independent samples t-tests that assess mean differences between two independent groups. Finally, binomial regression models were used to explore which of the behavioural features could accurately identify depression. Such models predict the odds of a variable falling into one of the two categories of the outcome (symptoms of depression or no symptoms). Confidence intervals (CI) show the range of possible values for each variable.

**Table 2.** Demographic characteristics for participants per group

	People with symptoms of depression	People with no symptoms	P-value
N	12	40	
Age [mean, SD]	51 (24)	47 (19)	0.073 <sup>a</sup>
Gender [number]			0.538 <sup>b</sup>
Males	4	16	
Females	8	24	
Employment [number]			0.138 <sup>b</sup>
Retired	6	13	

(continued)

**Table 2.** (continued)

	People with symptoms of depression	People with no symptoms	P-value
Undergraduate student	2	3	
Postgraduate student	0	2	
Full-time employed	2	16	
Part-time employed	0	5	
Unemployed	2	1	
Education [number]			0.307 <sup>b</sup>
Secondary	3	5	
Undergraduate	5	12	
Postgraduate	4	23	
Marital status [number]			0.056 <sup>b</sup>
Married	1	13	
Single	6	23	
Divorced/separated	3	2	
Widowed	2	2	

<sup>a</sup>Independent samples t-test

<sup>b</sup>Chi-square test

### 3 Results

Initial analysis with raw data showed that reduced activity could identify symptoms of depression ( $p = 0.043$ ) (model 1 in Table 3). In model 2, missing values were imputed with the mean score in each sensor. Because of the small number of people with symptoms of depression in this study ( $N = 12$ ), we used the Synthetic Minority Oversampling Method (SMOTE) [20]. The logistic regression model achieved 82.8% overall, with 78.1% accuracy in identifying no symptoms, and 87.5% accuracy in identifying symptoms of depression. Binomial regression models showed that reduced time of physical activity ( $p = 0.002$ ) and reduced application use ( $p = 0.032$ ) could identify the presence of symptoms of depression (Table 3). Higher sleep efficiency score ( $p = 0.046$ ), visiting more locations ( $p = 0.005$ ) and increased number of incoming calls ( $p = 0.002$ ) could also identify depression.

**Table 3.** Individual sensors for identifying depression

Sensors	Model 1			Model 2		
	Odds ratio	CI	Sig.	Odds ratio	CI	Sig.
Activity duration	0.99	0.99–1.00	<b>0.043</b>	0.99	0.99–1.00	<b>0.002</b>

(continued)

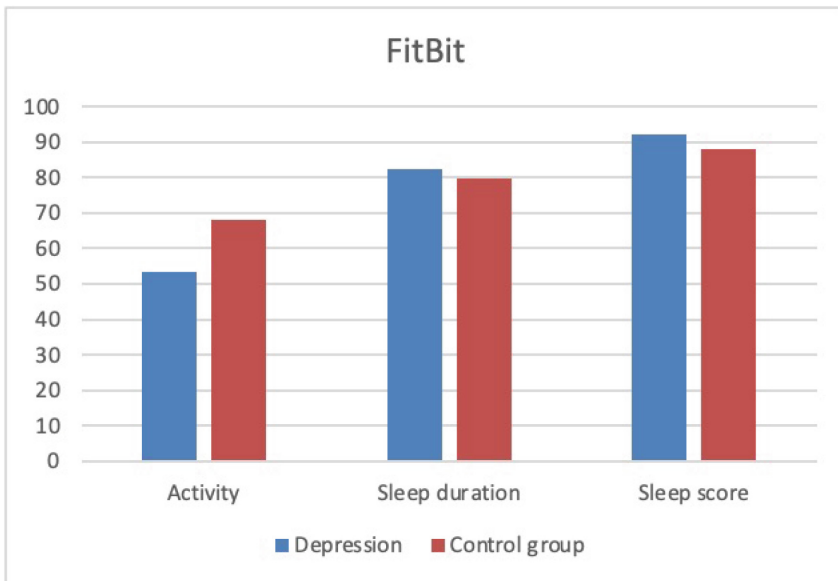


**Table 3.** (continued)

Sensors	Model 1			Model 2		
	Odds ratio	CI	Sig.	Odds ratio	CI	Sig.
Sleep duration	1.00	1.00–1.01	0.462	1.00	1.00–1.00	0.487
Sleep score	1.02	0.96–1.10	0.521	1.10	1.00–1.20	<b>0.046</b>
Application usage	1.00	1.00–1.00	0.935	0.99	0.99–1.00	<b>0.032</b>
Received texts	1.00	0.96–1.04	1.000	0.95	0.90–1.01	0.072
Sent texts	0.98	0.91–1.04	0.447	1.02	0.95–1.10	0.523
Number of outgoing calls	0.98	0.94–1.02	0.282	1.00	0.96–1.04	0.880
Outgoing calls (duration)	1.00	1.00–1.00	0.541	1.00	1.00–1.00	0.337
Number of incoming calls	1.05	0.97–1.12	0.227	1.17	0.95–1.17	<b>0.002</b>
Incoming calls (duration)	1.00	1.00–1.00	0.467	1.00	1.00–1.00	0.149
Changes in location	1.01	0.99–1.03	0.176	1.04	1.01–1.06	<b>0.005</b>
Bluetooth	1.00	1.00–1.00	0.286	1.00	1.00–1.00	0.057

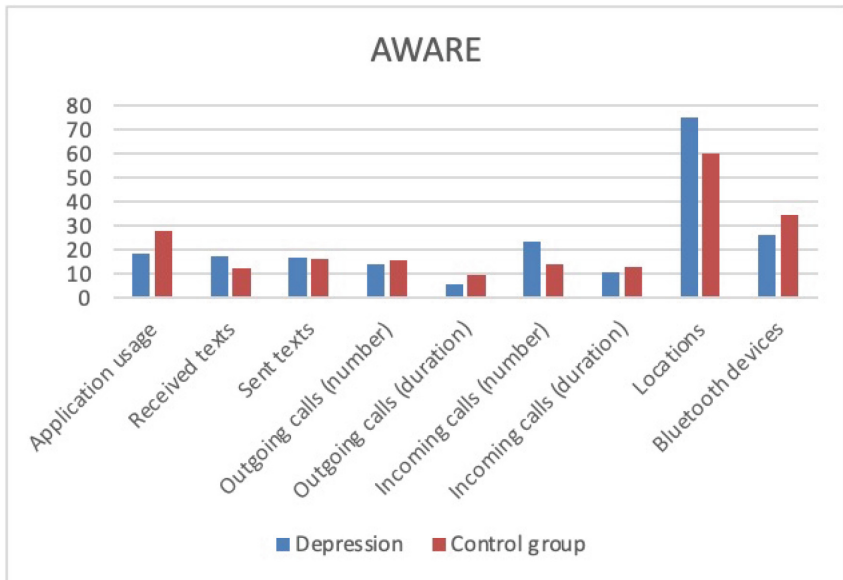
Model 1: results with raw data, Model 2: results after applying SMOTE and missing values calculations

Independent samples t-tests showed that people who experienced symptoms of depression had significantly more incoming calls [ $t(62) = -1.73, p = 0.045$ ], more



**Fig. 1.** Differences in behaviours between people with and without depression in activity and sleep.

changes in location [ $t(62) = -2.58, p = 0.006$ ], and lower duration of physical activity [ $t(62) = 3.00, p = 0.002$ ]. Figures 1 and 2 present the sensor data of people with symptoms of depression and the control group without symptoms in percentages using the highest score in each sensor as 100%. No significant difference was found between the two groups in the duration of sleep [ $t(62) = -0.61, p = 0.274$ ], the sleep efficacy score [ $t(62) = -1.00, p = 0.161$ ], their application usage [ $t(62) = 1.61, p = 0.057$ ], the number of received and sent texts [ $t(62) = -1.11, p = 0.135, t(62) = -0.09, p = 0.466$ , respectively], the number and duration of outgoing calls [ $t(62) = 0.36, p = 0.360, t(62) = 1.16, p = 0.125$ , respectively], the duration of incoming calls [ $t(62) = 0.48, p = 0.316$ ], and in the number of Bluetooth contacts that were detected near them [ $t(62) = 1.38, p = 0.086$ ].



**Fig. 2.** Differences in behaviours between people with and without depression in communication, changes in location and detected Bluetooth devices nearby people.

## 4 Discussion

This study explored the accuracy of passive sensing to identify symptoms of depression. Moreover, the study investigated which sensors from mobile applications and fitness trackers are accurate in identifying depression and whether there are behavioural differences between people with symptoms of depression and people without symptoms. The results from this study confirmed past findings about the accuracy of passive sensing to identify symptoms of depression. Depression was associated with reduced physical activity, higher sleep efficiency, an increased number of incoming calls and locations, and reduced use of social applications.

The results of reduced physical activity are in line with past research [7]. Incoming calls can accurately identify symptoms of depression [21]. Past research showed that reactive phone users answer more phone calls than make calls themselves and are more likely to experience symptoms of depression [22]. Our finding that people with depression visit more places is in contrast to existing evidence [10, 11]. This increased number of visited places could reflect the need of people with symptoms of depression to meet people. Our finding that people with symptoms of depression used fewer social applications is consistent with past research and can be explained by social withdrawal, which is often observed in depression [23]. Our results about improved quality of sleep in people with symptoms of depression are not in line with past research [6]. However, FitBit calculates the sleep score based on the duration of sleep, the time spent in different sleep stages and how relaxed people are during sleep. People with depression often present different sleeping patterns varying from oversleeping to sleep deprivation. Our behavioural results did not show a difference in the sleep duration between people with depression symptoms and those without symptoms. This may be because of the different sleep disturbance issues across different people with depression, and this result may affect their efficiency score.

This study showed that other types of communication, including outgoing calls and text messages, could not identify depression or its absence and thus did not confirm past research that showed reduced social activity and communication [8] in people with depression symptoms. This failure of our sensing data to reproduce past results may be because of the small sample size, especially in the group of people with symptoms of depression. This is a limitation of the study compared to some past studies. A systematic review of methods for passive monitoring of depression [24] has shown past studies' pitfalls in this field, including the lack of necessary information, such as information about participants' recruitment strategies, and information about participants and the features. Other issues that were observed in this review about studies using passive sensing to identify depression include their small sample sizes and short follow-up duration [14, 24]. This study was part of a larger project on ageing that recruited people between 18 and 50 years old, and people 65 years old or older. Thus, data from people between 51 and 64 years old was not collected. Another limitation of our study is the inability to group people with depression symptoms in more groups depending on the severity of symptoms because of the small number of people with symptoms and the imbalanced gender ratio with female participants being more than male participants. Similar to past studies, the issue with accuracy and amount of data that is often missing affects data analysis. For example, Table 1 shows 0 as the minimum value of changes in location. This value is inaccurate because participants visited University College Cork at the beginning and the end of the study. Improvements in technology for monitoring mental health and well-being should be taken into account in future research. In addition, future research with larger sample size could employ more personalized machine learning models to identify symptoms of depression of the severity of the symptoms, for example by identifying subgroups within the sample with similar behavioural patterns. Group-based prediction models have been found to be more accurate in identifying other outcomes compared to generalized models [25].

## 5 Conclusions

Concluding, this study extends past research to explore the identification of symptoms of depression through passive sensing by including people from a broader age range and by combining data from both smartphones and fitness trackers. The results showed the accuracy of certain sensors for activity, sleep, calls, application use, and changes in location to identify symptoms of depression; however, this finding should be interpreted in relation to the methodological issues that have been identified in this field, including the sample size and the information that is provided from each study to enable comparisons. Future research could address those limitations to improve the generalizability of results and the reproducibility of research, for example, through group-based or personalized predictions models. The challenges in this field that are discussed in Sect. 4 could be considered by researchers who can contribute to the development of interdisciplinary teams to design technological solutions based on the needs of users. Research in this field of technology, mental health, and wellbeing can result in more accurate tools to assess ourselves in our everyday life and prevent the development or more serious symptoms that require professional help.

The results and evaluation of the present study contribute towards adopting more effective research practices. Similar to the Digital Humanities track of AISoLA, the authors of this paper form an interdisciplinary team with backgrounds in computer science, psychology and social work who explore the challenges and opportunities of technology to support people in their everyday life through different research studies.

**Acknowledgments.** This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 13/RC/2077-P2 (CONNECT Centre for Future Networks).

## References

1. American Psychiatric Association. <https://www.apa.org/topics/depression>. Accessed 10 Dec 2023
2. Institute of Health Metrics and Evaluation. <https://vizhub.healthdata.org/gbd-results/>. Accessed 10 Dec 2023
3. World Health Organization. [https://www.who.int/health-topics/depression#tab=tab\\_1](https://www.who.int/health-topics/depression#tab=tab_1). Accessed 10 Dec 2023
4. Halfin, A.: Depression: the benefits of early and appropriate treatment. *Am. J. Managed Care* **13**(4), S92–S97 (2007)
5. Cacheda, F., Fernandez, D., Novoa, F.J., Carneiro, V.: Early detection of depression: social network analysis and random forest techniques. *J. Med. Internet Res.* **21**(6), e12554 (2019)
6. Kaneita, Y., et al.: The relationship between depression and sleep disturbances: a Japanese nationwide general population survey. *J. Clin. Psychiatry* **67**(2), 196–203 (2006)
7. Roshanaei, B., Katon, W.J., Russo, J.: The longitudinal effects of depression on physical activity. *Gen. Hosp. Psychiatry* **31**, 306–315 (2009)
8. Santini, Z.I., et al.: Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NSHAP): a longitudinal mediation analysis. *Lancet Public Health* **5**(1), e60–e70 (2020)

9. Narziev, N., Hwarang, G., Toshnazarov, K., Lee, S.A., Chung, K.M., Noh, Y.: STDD: short-term depression detection with passive sensing. *Sensors* **20**(5), 1396 (2020)
10. Wang, R., et al.: StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: *ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2014*, pp. 3–14 (2014)
11. Saeb, S., et al.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J. Med. Internet Res.* **17**(7), e175 (2015)
12. Trifan, A., Oliveira, M., Oliveria, J.L.: Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. *JMIR Mhealth Uhealth* **7**(8), e12649 (2019)
13. Renn, B.A., Pratap, A., Atkins, D.C., Mooney, S.D., Areán, P.A.: Smartphone-based passive assessment of mobility in depression: challenges and opportunities. *Ment. Health Phys. Act.* **14**, 136–139 (2018)
14. Sheikh, M., Qassem, M., Kyriacou, P.A.: Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Front. Digit. Health* **3**, 662811 (2021)
15. Ferreira, D., Kostakos, V., Dey, A.K.: AWARE: mobile context instrumentation framework. *Front. ICT* **2**(6) (2015)
16. Kumar, D., Jeuris, S., Bardram, J.E., Dragoni, N.: Mobile and wearable sensing frameworks for mHealth studies and applications: a systematic review. *ACM Trans. Comput. Healthc.* **2**(1) (2020)
17. Sheikh, J.I., Yesavage, J.A.: Geriatric depression scale (GDS): recent evidence and development of a shorter version. *Clinical Gerontologist. J. Aging Mental Health* **5**(1–2), 165–173 (1986)
18. Chow, P.I., et al.: Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *J. Med. Internet Res.* **19**(3), e62 (2017)
19. Goodspeed, R., et al.: Comparing the data quality of global positioning system devices and mobile phones for assessing relationships between place, mobility, and health: field study. *JMIR Mhealth Uhealth* **6**(8), e168 (2018)
20. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
21. Tlachac, M., Melican, V., Reisch, M., Rundensteiner, E.: Mobile depression screening with time series of text logs and call logs. In: *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens*, pp. 1–4 (2021)
22. Aubourg, T., Demongeot, J., Renard, F., Provost, H., Vuillerme, N.: Association between social asymmetry and depression in older adults: a phone call detail records analysis. *Sci. Rep.* **9**, 13524 (2019)
23. Elhai, J.D., Levine, J.C., Dvorak, R.D., Hall, B.J.: Non-social features of smartphone use are most related to depression, anxiety and problematic smartphone use. *Comput. Hum. Behav.* **69**, 75–82 (2017)
24. De Angel, V., Lewis, S., White, K., et al.: Digital health tools for the passive monitoring of depression: a systematic review of methods. *NPJ Digit. Med.* **5**(3) (2022)
25. Malik, M.Q., Bantry-White, E., Zafeiridi, E., Pesch, D.: Personalizing loneliness detection through behavioral grouping of passive sensing data from college students. *IEEE Access* **11**, 88841–88851 (2023)




**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# The GraphBRAIN Framework for Knowledge Graph Management and Its Applications to Cultural Heritage

Stefano Ferilli<sup>(✉)</sup>, Eleonora Bernasconi, Davide Di Pierro,  
and Domenico Redavid

University of Bari, Bari, Italy

{stefano.ferilli,eleonora.bernasconi,davide.dipierro,  
domenico.redavid}@uniba.it

**Abstract.** The traditional record-based approach to the description of Cultural Heritage is nowadays obsolete. It is unable to properly handle complex descriptions and it cannot support advanced functions provided by Artificial Intelligence techniques for helping practitioners, scholars, researchers and end-users in carrying out their tasks. A graph-based, semantic approach is needed, such as that provided by Semantic Web solutions. Also, a ‘holistic’ description approach is needed, that includes and inter-connects all branches and types of Cultural Heritage, and that is not limited to describing just the formal metadata of cultural objects, but can deal with their content, physicality, context and life-cycle, as well. The GraphBRAIN framework and technology for Knowledge Graph management enforces all these ideas and enjoys improved efficiency, expressiveness, and flexibility thanks to the use of the LPG model for knowledge representation. This paper describes GraphBRAIN and its application to several Cultural Heritage-related fields, including digital libraries, archives and museums, history of computing, and tourism as a way to boost fruition of these items.

**Keywords:** Semantic Web · Knowledge graph · Knowledge representation · Cultural Heritage

## 1 Introduction and Motivations

For many good reasons, description of Cultural Heritage (CH) items has traditionally been organized in the form of records with a fixed number of pre-defined fields. These fields were organized around a set of metadata mostly oriented to describing the formal aspects of the items. This choice made perfect sense based on the technology available in the past, and to the descriptions being thought specifically for use by experts (practitioners and researchers). The landscape has now changed in both perspectives. Digital technologies provide representational and computational support much more advanced and complex<sup>1</sup> than what was

<sup>1</sup> By ‘complex’ we mean with a very large number of fields, with variable or repeated fields, and with fields which are in turn record themselves.

possible when using paper cards. Also, there is a trend in the last decades toward opening CH to the wider public, which are often enthusiasts or curious, without any technical skill in cataloging and with an immensely broader range of motivations, interests, goals, backgrounds, preferences, etc.

One of the drawbacks of legacy approaches to cataloging and description of CH is that a different record structure, with different sets of fields, was defined for different kinds of CH. Just to mention the most prominent ones, we may think of Galleries, Libraries, Archives and Museums (the so-called GLAM). First of all, this specialization requires new description standards when new kinds of CH items are approached (e.g., new standards have been issued for the description of scientific instruments, or for electronic instruments, after realizing that the standard records for museum items were totally unable to capture their fundamental peculiarities). The Central Institute for Cataloging and Documentation (ICCU) of the Italian Ministry of Culture currently provides for 4 areas of protection (archaeological, architectural and landscape, demo-ethno-anthropological and historical-artistic), and defined 9 different catalog cards organized according to the different subject areas: archaeological heritage, architectural and landscape goods, demo-ethno-anthropological goods, photographic heritage, musical goods, naturalistic goods, numismatic goods, scientific and technological goods, historical and artistic goods<sup>2</sup>. Still, this is insufficient: the card for scientific and technological goods, while fitting standard equipment, is totally unable to capture the complexity of, e.g., the computing domain, especially from its historical perspective, that is of utmost interest for CH [13]. The landscape, and other kinds of immaterial CH, are progressively attracting attention, both for preservation and for economic exploitation purposes. And new branches and needs will continuously arise as technology and society develop.

Secondly, a consequence of having different description standards means that the various subject areas follow their own way, making it difficult to make them converge, while the CH field is obviously a coherent whole (e.g., archive documents describe the history of museum items, whose interpretation is reported in scholarly books). A solution is needed that can be applied to CH as a whole, while still keeping the specificities of each of its branches, and that is flexible, so as to allow easy extension and variation to cover new or changed needs of the field. Even more, we think that this solution should go beyond the pure and traditional focus on formal CH metadata, also including all knowledge that is non-strictly related to the single items, or even to CH in general, but that can provide precious ‘glue’ to connect and inter-relate them, opening new possibilities for their understanding, management and exploitation. We strongly believe that such a unified framework will boost the field, providing immense expansion opportunities and unprecedented support to all the stakeholders (practitioners, researchers, scholars, enthusiasts, tourists, end-users), also thanks to the use of state-of-the-art Artificial Intelligence (AI) solutions.

Initially, the legacy approach to description and cataloging has been simply ported to digital as-is. Then, after realizing the flexibility that digital solutions

<sup>2</sup> <http://iccd.beniculturali.it/it/settoridisciplinari>.



could provide, it was expanded, but still centered on the record-based approach, just providing for many more fields, structuring them, and making most of them optional, as in the MARC series of standards. Instead, a much better representational option to overcome the limitations of record-based descriptions is the graph model. Basically, it consists of nodes, typically representing objects, and arcs, representing associations between pairs of objects.

The most outstanding advantage of graphs is their very intuitive interpretation by humans, and ubiquitous use in everyday life, paired with their having a mathematical definition that allows to build a whole formal theory on them. In fact, graphs have been chosen as the basic structure for research in the Knowledge Representation and Reasoning (KRR) branch of AI. KRR investigates how to represent, store and manipulate knowledge, in so-called Knowledge Bases (KBs). More specifically, when KBs are based on graph representations, they are called Knowledge Graphs (KGs). KBs/KGs typically include two components: the ontology (defining what can be represented, how, and what are its properties and behavior) and the instances (i.e., the actual data). The former provides meaning to the latter, and allows different entities or systems to interoperate, by assigning the very same meaning to the same concepts and objects. This is one of the main objectives of the Semantic Web, that indeed adopted this approach and developed its own standards, formalisms and storage solutions for it.

In this paper we propose a KG-based approach to CH description and manipulation. Still, we propose to depart from the standard Semantic Web practices, and pursue its same objectives but starting from a database (DB) perspective, for several reasons. First, this is the traditional setting in the field of CH description. Second, this allows us to take advantage of the efficiency and scalability of the latter, while ensuring semantics and interoperability as in the former. In fact, modern DB solutions are available that rely on the graph model. Third, not being bound to the SW representations, we may apply a wider range of tools to our knowledge, and in particular many AI techniques that can support advanced tasks for the final users of the knowledge. To support our vision we developed the GraphBRAIN framework and platform. It uses technology from the DB community for storing instances, and superimposes ontologies that, on the DB side, are interpreted as data schemas, while, on the outer world, enable semantic-based interoperability. Among others, GraphBRAIN was applied to the field of CH, and to several branches thereof. Here we will provide an overall account of these applications, to show the power of our solution and its potential.

The rest of this paper is organized as follows. After discussing related work in the next section, we summarize the GraphBRAIN framework in Sect. 3, and review its applications to the CH domain in Sect. 4. Finally, Sect. 5 concludes the work and outlines future work issues.

## 2 Related Work

In this section we will explore two different aspects of related work on KGs for CH, namely the existing KBs and interfaces.

## 2.1 Knowledge Bases for Cultural Heritage

The development and curation of domain-specific knowledge structures have traditionally been essential in the humanities [18]. Some efforts focused on porting existing information to digital: Pleiades [15], a repository of data pertaining to geographical locations, with relevance to the examination of ancient literature and history; Papyri.info [5], a search engine that seamlessly integrates multiple DBs containing ancient documents; MANTIS [17], the semantically enriched DB maintained by the American Numismatic Society, focused on the comprehensive study of coins from various historical periods and cultures; Open Context [20], a repository encompassing diverse resources, including archaeological reports; Trismegistos [6], a metadata platform catering to the study of texts from the Ancient World, housing data related to ancient documents, individuals, and locations; EDH [16], the Epigraphic Database Heidelberg, a search tool dedicated to Latin epigraphic data.

In this direction, the interest in semantic annotation through formal languages has also been active in the Semantic Web since its beginning. In [7] the characteristics of CH are analyzed to identify how ontologies could be used to improve CH information management. In particular, this paper analyzes the integration of different schemas using the CIDOC-CRM<sup>3</sup> ontology (ISO 21127 standard) as a reference. This ontology, used by various cultural organizations worldwide, was developed from the bottom up by integrating semantic contents of various DB schemas and documentation structures from all kinds of museum disciplines, archives and libraries. In order to keep it compact, only a part of the initial concepts and properties were used, thus making it unsuitable for handling complex scenarios that combine descriptive and management aspects of CH. To enable semantic interoperability, an OWL version of this ontology<sup>4</sup> is used by the British Museum. It remains, however, difficult to use in scenarios involving the combination with additional aspects that might aid fruition of the annotated material.

Worth mentioning for our purposes are some National projects. A relevant effort was carried out in the Netherlands and Finland for the Dutch national project CLARIAH [25], concerned with Digital Heritage from 3 perspectives: economic and social history, language, and media studies. To tackle the variety of facets, the project used 26 different ontologies, taxonomies, classification systems and lexicons, of which many applicable to broader context (such as SKOS, PROV and Schema.org), and 6 developed from scratch to model new humanities domains. The knowledge representation and storage was based on Semantic Web technology, focusing on Linked Open Data. However, the authors note that this setting was not comfortable for users not familiar with these technologies. In Italy, Cultural-ON (Cultural ONtology) [21] was developed, an ontology aimed at modeling the data regarding cultural institutes or sites, their contact points, all multimedia files which describe them, the agents that play a

<sup>3</sup> <https://cidoc-crm.org/>.

<sup>4</sup> <http://erlangen-crm.org/current>.

specific role in them, events that can take place in them, and any other information useful to the public in order to access them. It is aligned with external ontologies (FOAF, PROV, schema.org, Dublin Core, etc.). ArCo (Architecture of Knowledge) [3], an ontology for, and a KG of, Italian Cultural Heritage, models many types of cultural properties (including technological heritage), for which it allows to capture details such as elements affixed on cultural properties, copies, forgeries and other works related to a cultural property, specific surveys, cadastral information, historical locations, the communication medium of intangible demo-ethno-anthropological heritage, etc. It currently reuses, and is aligned to, CIDOC-CRM, EDM, Cultural-ON, and OntoPiA. These projects were sponsored by the Italian Ministry of Culture. We take inspiration from, and aim at being aligned with, these works, but we propose to use a different technology for handling this information, and envision a much broader context to be described in CH ontologies.

Still from the point of view of using annotations in complex scenarios, [4] presents an ontology-based approach to improve data retrieval by expert users in the CH field, e.g., archaeologists, art historians, geologists, etc. It proposes an advanced architecture with semantic search capability that can transform a vast amount of data into linked concepts for easy information comparison. All concepts are mapped onto elements of the ArCo ontology, and a semantic query layer allows to execute SPARQL queries (based on simple pattern matching, not on Description Logic reasoning). A mapping between ICCU (the Italian Central Institute for a Unique Catalog) and ArCo is proposed so that SPARQL queries can be run on ArCo to search information in ICCU records. However, as reported in the article, the mapping is partial, and therefore some information would not be found unless ArCo is extended. In contrast, our approach can generate the OWL classes and properties needed for a more complete coverage of SPARQL queries, and can also use several types of reasoning, including those provided by standard Semantic Web reasoners.

Linked datasets of libraries or museums have been gaining traction as an interconnecting spine through which community-specific datasets can build out-bound links to contribute to a global graph (e.g. the Virtual International Authority File<sup>5</sup>; the Getty Thesaurus of Geographic Names<sup>6</sup>; the Getty Art and Architecture Thesaurus<sup>7</sup>).

## 2.2 Creating and Managing Graph Data Model

In this section, we will explore the current landscape of platforms dedicated to creating, managing, and visualizing the models that make up Knowledge Graphs. These tools, including those for data visualization, schemas, and ontologies, provide users with a comprehensive overview of the underlying data structure [1].

---

<sup>5</sup> <https://viaf.org>.

<sup>6</sup> <https://www.getty.edu/research/tools/vocabularies/tgn/>.

<sup>7</sup> <https://www.getty.edu/research/tools/vocabularies/aat/>.

TopBraid Composer<sup>8</sup> serves as an ontology editing tool with visualization capabilities as an additional feature. The visualization approach draws inspiration from UML and offers horizontal and vertical tree layouts, accompanied by a traditional indented list view. This visualization represents classes and properties as nodes connected by directional edges labelled with their corresponding predicate names. Notably, this visualization operates at the RDF level, treating owl:Class as a distinct node and linking each class to it through an rdf:type edge.

WebVOWL [22] is an online application designed to offer user-friendly visual representations of ontologies, supporting exploration and allowing its users to engage with and personalize ontology visualizations. It must strictly adhere to the Visual Notation for OWL Ontologies (VOWL) to construct graphical depictions of OWL components. VOWL visualizations are automatically generated from JSON files, requiring the conversion of ontologies into JSON format, carried out by the provided Java-based OWL2VOWL converter. The force-directed graph layout relies on a physics simulation, resulting in dynamic animations that continually adjust node positions.

Ontodia<sup>9</sup> [24] is a web-based tool tailored for visualizing ontologies and semantic datasets. It employs a 2D node-link visualization approach and incorporates UML-inspired techniques to convey supplementary information about nodes. Users can choose between force-directed and grid layouts, and there's a hierarchical relationships view for displaying parent-child connections in a tree format. This tool provides flexibility through drag-and-drop functionality, allowing users to customize views by rearranging elements, removing nodes, and toggling links.

The Arca system [2] seamlessly links unstructured content with concepts within a knowledge graph (KG). This integration empowers users to perform intricate data queries and visualize the rich web of semantic connections that bridge concepts and documents.

### 2.3 Our Proposal

In this context, we introduce an approach that addresses the challenges associated with creating and managing ontologies in diverse domains, ultimately enriching the field of CH. Our proposal revolves around harnessing the power of GraphBRAIN, a versatile framework that transcends domain boundaries. GraphBRAIN is poised to revolutionize the ontology development and maintenance process, offering a robust solution for establishing and nurturing knowledge structures that amplify the preservation and exploration of CH. SKATEBOARD, serving as a GraphBRAIN interface, represents a pivotal advancement in the realm of semantic data visualization and exploration. With its multifaceted visualizations, user-friendly interface, and collaborative features, SKATEBOARD provides added value to researchers, data scientists, and knowledge professionals aiming to unlock the full potential of semantic knowledge graphs and derive actionable insights from complex data structures.

<sup>8</sup> <https://www.w3.org/wiki/TopBraid>.

<sup>9</sup> <https://github.com/metaphacts/ontodia>.

### 3 The GraphBRAIN Framework

GraphBRAIN [11, 14] is a framework developed to cover all tasks in KG management and exploitation based on the combination of leading graph DB technology for instance storage and ontologies for schema description. From the former it draws efficiency and a wide library of data analysis tools; from the latter it draws semantic power, interoperability and the possibility of plugging automated reasoning facilities. Differently from standard Semantic Web approaches, based on the simple atomic triples  $\langle \textit{subject}, \textit{predicate}, \textit{object} \rangle$  provided by the RDF model, it is based on the Labeled Property Graph (LPG) model. LPGs allow to add labels to nodes and arcs, and to specify attributes with their value for both nodes and arcs. Moreover, each node and arc gets a unique identifier, allowing to have different nodes with the same content and different arcs of the same type between the same pair of nodes. This enhances their expressiveness, readability and compactness (it is estimated that an LPG takes one order of magnitude less nodes to store the same information as an RDF graph). As typical in traditional relational DBs, and differently from the Semantic Web approach, GraphBRAIN keeps apart the schema/ontology, described in a GBS file, from the data/instances, stored in the DB.

GraphBRAIN ontologies can be defined using an XML-based formalism specifically designed to match the features of LPGs. It is organized in different sections that allow to: import existing ontologies in order to expand them; define new datatypes in the form of lists or trees of values; define a hierarchy of entities with their attributes; define a hierarchy of relationships with properties (symmetry, transitivity, functionality, etc.) and their attributes; define axioms in the form of logic formulas (typically rules or constraints) that must be verified by the instances in the KG. The basic datatypes provided by GraphBRAIN are: boolean, integer, real, string, text. Ontologies can be combined using the import section provided that they are compliant to each other, i.e., basically, that their hierarchies of entities are not inconsistent (a class  $C'$  is a superclass of class  $C''$  in one ontology, while class  $C'''$  is a superclass of  $C'$  in the other) and that their attributes are, too (the same attribute in different ontologies must be of the same type). Two ontological components are considered as the same if they have the same name.

The instances handled by one GraphBRAIN installation are stored in a single graph, using the Neo4j graph DB [27]. In GraphBRAIN's use of Neo4j, nodes are used to represent class instances, arcs are used to represent relationship instances (i.e., object properties in the Semantic Web); node labels are used to specify the specific class of the instance represented by the node, and all domains that are relevant to that instance; arc labels are used to specify the relationship expressed by the arc; node attributes are used to represent datatype properties in the Semantic Web; arc attributes have no counterpart in the Semantic Web (they cannot be expressed). Since Neo4j is schemaless, the ontology acts as the schema to determine what information can be stored in the graph, and how. Still, different ontologies may be applied to the same graph, providing different views on the data. The single-graph approach is fundamental for our purposes: even

if not visible when using an ontology for accessing the graph, the information associated to other ontologies is still there and may allow indirect connections among the items of the current ontology, that can be explored by the end user while browsing the graph or might be used by graph-based algorithm during their execution.

While mainly designed to allow semantic-based processing on a single KG, GraphBRAIN is open to integration with other resources, especially those available in Semantic Web repositories. In fact, a mapping between the GraphBRAIN formalism and standard Semantic Web is available, allowing interconnection of ontologies and instances alike, and interoperability of systems. As a first advantage, this allows to immediately use ontological reasoners on the knowledge handled by GraphBRAIN. On the other side, a large set of network analysis and graph mining functions can be applied on the data, inherited by the Neo4j libraries and tools. Additionally, not being tightly bound to the standard RDF format, the information in the KG can also be sent to other AI tools, such as rule-based or constraint-based reasoners. We are currently working on the MultiStrategy Reasoning engine GEAR [12], providing a combination of deduction, abduction, abstraction, induction, argumentation, probabilistic reasoning, and abstraction.

A GraphBRAIN API is provided, ensuring that all interactions with the DB happen according to the schema. Given an ontology and a DB, the API provides both basic and advanced functionality on the KG. Basic functionality includes standard CRUD (Create, Read, Update, Delete) operations. For queries, it wraps the Neo4j language Cypher, checking that the specified information is compliant to the ontology before running the query. Advanced functionality include analysis, mining and reasoning functions. E.g.: computing the centrality of an entity instance in the graph according to different algorithms; extracting a relevant portion of the graph starting from given nodes, possibly considering the user profile to obtain a personalized result; finding all possible paths in the graph between given pairs of nodes; checking consistency of the available knowledge; deducing or abducting knowledge that is not explicitly present in the graph; etc.

The API can be used by any third-party application. GraphBRAIN natively provides a Java-based Web Application implemented in JSF technology that allows ontology browsing and development, form-based CRUD operations on the single nodes (entity instances) or arcs (relationship instances), management of attachments and of collaborative interactions to populate the knowledge, etc. A graph-based visualization is also provided, where the user can browse the knowledge, reshaping and expanding the visible portion of the graph, and can apply the various advanced tools provided by the API. Through this interface, ontologies and instances can also be exported or imported to or from other formalisms, including the Semantic Web standard OWL.

Figure 1 shows an ontology as seen in the GraphBRAIN Web Application interface. On the left, the hierarchy of classes can be browsed; selecting a class its attributes are displayed; selecting an enumeration attribute its values are displayed. Classes, attributes, and their values can be added, deleted or renamed

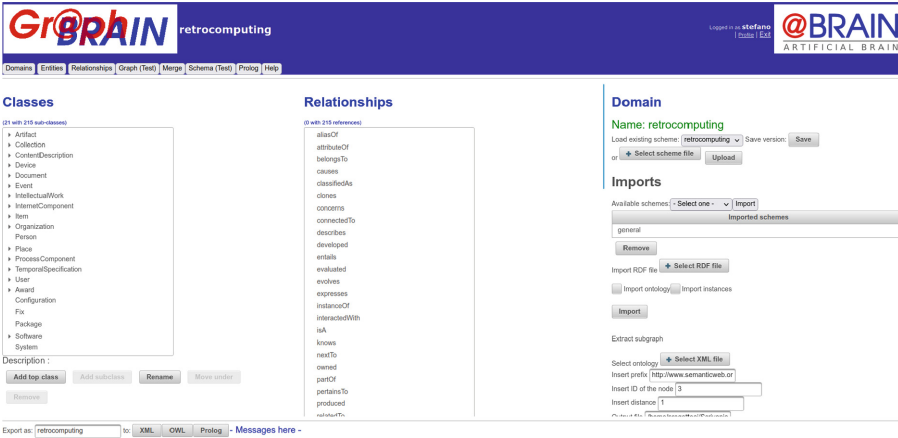


Fig. 1. Ontology in GraphBR@IN KG

from this section. In the center section, the same information and controls as for the classes are available, plus information and tools concerning the inverse relationships and the subject and object classes available for the relationships. On the right, controls to upload ontologies, or to import ontologies in the current one, are provided. On the bottom, controls to save local copies of the current ontology in several formats, including standard OWL, are available.

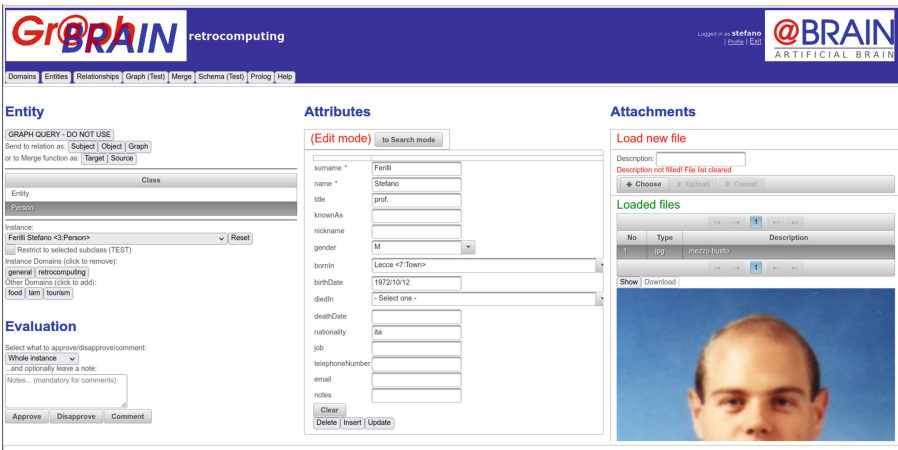


Fig. 2. GraphBR@IN Web Application: Entities form-based management tab

Figures 2 and 3 show the form-based visualization tabs for entities and relationships, respectively, in GraphBR@IN’s Web Application. They allow to select a type of class or relationship and to perform CRUD operations on their

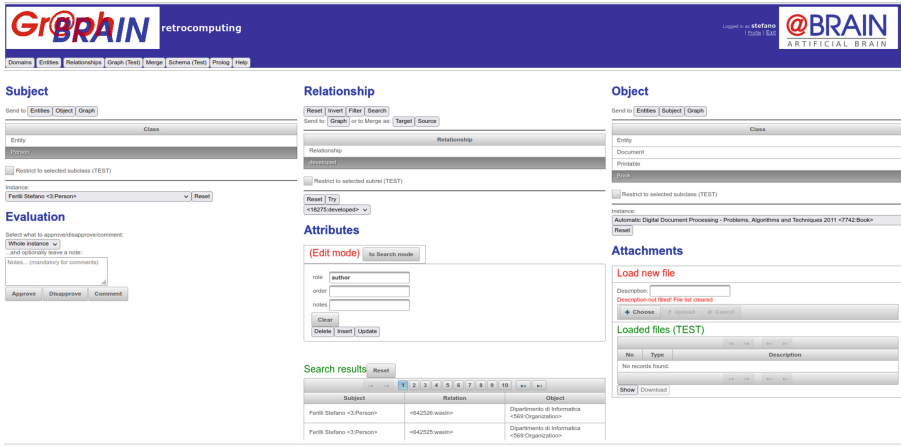


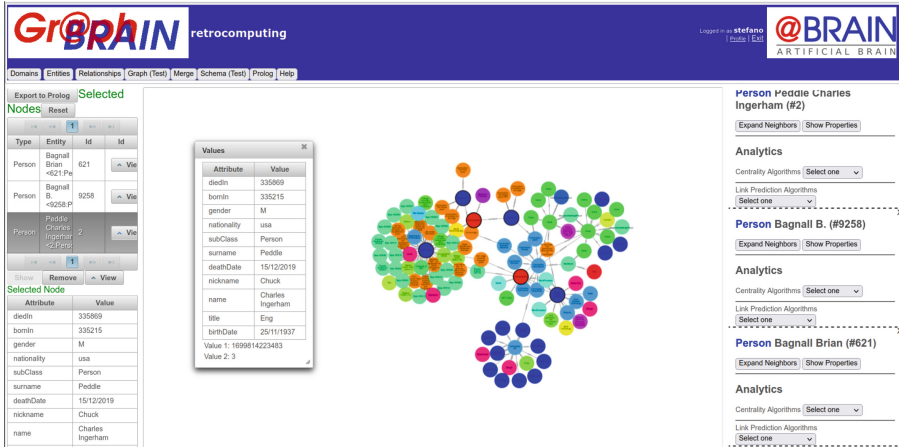
Fig. 3. GraphBRAIN Web Application: Relationships form-based management tab

instances. The entity or relationship attributes are shown in the middle. Attachments can also be managed on the right, and various kinds of filters and of tools for moving information across different sections are provided. Users can also provide feedback on the available knowledge items using the section on the bottom-left.

Figures 4 and 5 show a portion of the instances in the KG at various levels of zoom, as rendered by the graph-based visualization section of GraphBRAIN’s Web Application. The starting nodes to extract the subgraph are those listed in the table on the top-left of the interface and highlighted with a thick border in the graph on the right. Different node colors denote different classes. In Fig. 5 it is better visible that the node content is a summary of the entity instance’s data, and that arcs are labeled with the corresponding relationship name. Automatically formed aggregates of strictly related nodes are clearly visible, especially in Fig. 4. For each selected node, its detailed information can be displayed in a table on the bottom of the left column, in a pop-up window, or in a dedicated area on the right column (see Fig. 4). Controls to compute indicators about the node (centrality, etc.), or to run several node-based analysis and mining algorithms on the graph, are available on the node-related areas in the right column or by left clicking on any node in the central area. By right clicking on a blank zone in the central area, other controls are available to run various graph-based algorithms (clustering, path finding, personalized subgraph extraction, etc.) and for rendering in natural language selected portions of the graph.

For end-users, a separate interface called SKATEBOARD is also provided as a Web Application (shown in Fig. 6). It is mainly based on knowledge browsing and exploration, allowing to visualize, expand or compress a portion of the graph, to look into the single nodes or arcs, and to apply a number of semantic filters





**Fig. 4.** GraphBRAIN Web Application: Graph-based visualization and management tab (zoomed out)

that can support the needs of the different users. Since these functions can be applied also to standard Semantic Web KGs, this interface is separate from the previous one, and designed to work also with standard SPARQL endpoints.

## 4 Cultural Heritage Applications of GraphBRAIN

The GraphBRAIN framework and technology have been used to power several projects, many of which in the field of Cultural Heritage. Here we mention LAM (Libraries/Archives/Museums) [8,9], Open Science [10], Linguistics [23] and Retrocomputing [13].

As to the LAM domain, it was investigated after the consideration that traditional record-based approaches are obsolete and insufficient to support modern exploitation of, and research on, library, archive and museum items. We termed our graph-based approach a ‘holistic’ one, since it aimed at representing all possible aspects of LAM, not just those related to the formal metadata traditionally used to describe Cultural Heritage items. While starting from a core ontology that is fully aligned with the IFLA proposals for library description, FRBR [19] and LRM [26,28], we expanded it to make it able to capture the *content* of cultural objects (text, images, concepts expressed therein), their *physicity* (materials, manufacture, shape, structure), their *context* (the periods, people, organizations, events, places, etc. to which it has some relationships, and the relationships among these contextual items, even independently of the reference cultural object, that are crucial to find unknown and/or unexpected indirect connections that may suggest, prove or support investigation hypotheses and directions), and even their *lifecycle* (involving all the history of the objects, of its uses and of its users). For specific educational applications, the ontology was



**Fig. 5.** GraphBRAIN Web Application: Graph-based visualization and management tab (zoomed in)

also extended and aligned with IEEE’s LOM schema for describing Learning Object Metadata<sup>10</sup> and with OERSchema for Open Educational Resources<sup>11</sup>.

The Open Science domain was an almost straightforward extension of the LAM one. In fact, Digital Libraries are the obvious candidate infrastructure to support it, given that scientific publications are the core of open science. Still, it had to be expanded to describe the context and environment in which scientific development takes place. This involves processes and projects, datasets and corpora, scientific groups and communities, hardware and tools, software and storage facilities, etc. The ontology is currently compliant with OpenAIRE<sup>12</sup> and OAI-ORE<sup>13</sup>.

Connected to the LAM domain are also the investigations carried out on linguistics, due to their being based on the sources available in library and archive documents, or on museum items such as epigraphs. Here, the integration of semantic information into language resources was key to open up new avenues of enquiry into the mechanisms of language change. Experiments were run in integrating data from Latin textual corpora and language resources, and showed the potential of the GraphBRAIN framework for research into the mechanisms of semantic change in Latin.

While being one of the oldest applications of GraphBRAIN, the Retrocomputing domain (concerning the history of computing) took great advantage from the subsequent development of the LAM and Open Science perspectives. In fact, the history of computing heavily relies on the LAM perspective because

<sup>10</sup> <https://ieeexplore.ieee.org/document/9262118>.

<sup>11</sup> <https://oerschema.org/docs/schema.html>.

<sup>12</sup> <https://www.openaire.eu/>.

<sup>13</sup> <https://www.openarchives.org/ore/>.

The screenshot displays the GraphBRAIN KG interface. On the left, a search bar contains the URL `http://barbari-https://purl.archive.org/ontology/barbari_owl_pook_24823`. Below it, a list of entities is shown, including 'Abani Bari Achho', 'Barbarian Odes', 'Takhliis al-Ibriz fi Talkhiis Bariz', and 'Waiting for the Barbarians'. The central area features a graph visualization with nodes and edges. Nodes include 'E Flanders Road' (Expression), 'P London' (Place), 'E The Sinking Of The Scharnhorst' (Expression), 'Person Philippa Gregory', 'Person Nanni Balestrini', 'Person Etel Adnan', and 'C Person' (Class). Edges are labeled 'publishing place' and 'type'. The right panel, titled 'INFO', contains a welcome message and instructions on how to use the resource, such as searching for authors, works, or places, and highlighting authors' minority.

Fig. 6. Instance browsing in GraphBRAIN KG

it encompasses books and manuals for the machines and software (concerning libraries), archival documents concerning the persons, organizations and events that took place (concerning archives), and the hardware components (concerning museums). On the other hand, it also relies on the Open Science perspective because the scientific research in Computer Science is primarily concerned with scientific papers, experiments, datasets, tools and, as a consequence, with the hardware and software used in the research and experiments.

Even on its own, the Retrocomputing domain is representative of an extremely complex domain to represent. It involves, and inextricably interconnects, documentation, hardware, software and even immaterial heritage (e.g., the anecdotes that can still be known from the pioneers and central players in the history of computing, that are in many cases still alive and willing to tell behind-the-scenes information that is lost for most of the other, much older disciplines). Hardware and software cannot be understood without their associated documentation; the software is nothing without the hardware to run it; the hardware is dead without software to run; the archival documents and immaterial knowledge are often key to properly understanding all the other items. Also, the traditional fields defined for other types of CH (even those for scientific instruments and electronic equipments) do not fit at all the needs for the description of computing hardware, where nearly each single unit is unique, for several reasons: several versions of the apparently same model may exist, units are highly configurable and expandable, many components are interchangeable while apparently not changing the perceived behavior, several tweaks or modifications can be needed to restore and fix some units, especially old ones for which the original parts are not available anymore.

We include in this list also the Food and Tourism domains. While not CH domains by themselves (albeit, if considering traditional dishes and recipes, or landscapes and folklore, they may well be considered immaterial CH), together with the CH proper sections of the KG, they contribute to make up an ecosystem aimed at enhancement and exploitation of the CH items by final users. This again falls in our holistic perspective, and provides a clear example of how it can open up new possibilities with respect to traditional approaches to CH.

While these ontologies can be connected to each other via a few common entities, that act as bridges between the different domains and allow reuse of knowledge across them, the most relevant opportunity for their interconnection comes from a *general* top-level ontology, defined in GraphBRAIN independently of the various specific domains, and including ubiquitous and highly reusable concepts that can be reused (and specialized, if needed) by the domain-specific ontologies: Person, Organization, Event, Place, Collection, IntellectualWork, Item are just a few prominent examples.

Table 1 reports figures on the current content of the KG that can be freely consulted from the GraphBRAIN’s demo prototype available (upon registration) at <http://digitalmind.di.uniba.it:8088/GraphBRAIN/>. Note that the overall number of items is much larger than the number of items labeled with a domain. This is important, because unlabeled items are not part of any specific domain, but allow to indirectly link and inter-relate items otherwise disconnected across domains or even within single domains. Not all domains have figures in Table 1, because some of them are still under investigations and their data were not yet uploaded in the prototype. The most populated domain is Retrocomputing, which could be expected since it is the oldest one. The less populated are Food and Tourism, which were most recently introduced. The number of entity attribute values is much larger than that of relationship attribute values, which again could be expected, since relationships are meaningful by themselves, while entity instances can be identified and distinguished only based on their attributes. On the other hand, the number of relationship instances is typically larger than that of entity instances, because many different relationships can be established among the same set of objects<sup>14</sup>.

## 5 Conclusions and Future Work

Since the traditional record-based approach to the description of Cultural Heritage is nowadays unable to properly handle complex descriptions, or to support advanced functions provided by Artificial Intelligence techniques for helping practitioners, scholars, researchers and end-users in carrying out their tasks, this paper focused on a graph-based, semantic approach, such as that provided

<sup>14</sup> For each type of relationship, given  $n$  objects the number of possible relationship instances is  $2^n$  if only one such instance can be set between a given pair of objects. Since in LPGs many instances of the same relationship may be set between the same pair of objects (distinguished by unique ids associated to each instance), in our case this number is theoretically unbound.

**Table 1.** Statistics on the content of the current GraphBRAIN prototype’s KG

Domain	entity inst.	entity attr.	relationship inst.	relationship attr.
Overall (unlabeled)	337287	2089580	496839	41594
Overall (labeled)	2038	8069	2512	1958
General	102	573	222	132
LAM	63	294	93	69
OpenScience	—	—	—	—
Linguistics	—	—	—	—
Retrocomputing	1688	6801	2142	1757
Food	169	338	47	0
Tourism	14	56	8	0

by Semantic Web solutions. Also, a ‘holistic’ description approach is needed, that includes and interconnects all branches and types of Cultural Heritage, and that is not limited to describing just the formal metadata of cultural objects, but can deal with their content, physicality, context and lifecycle, as well. The GraphBRAIN framework and technology for Knowledge Graph management enforces all these ideas and enjoys improved efficiency, expressiveness, and flexibility thanks to the use of the LPG model for knowledge representation. This paper described GraphBRAIN and its application to several Cultural Heritage-related fields, including digital libraries, archives and museums, history of computing, and tourism as a way to boost fruition of these items.

Future work is ongoing in several directions, to extend and refine the GraphBRAIN framework, its API and interfaces, and its Cultural Heritage-related KG. In particular, within the effort for Spoke 3 “Digital Libraries, Archives and Philology” of project CHANGES “Cultural Heritage Active innovation for Next-Gen Sustainable society”, winner of the NRRP program of the Italian Ministry of University and Research, funded by the NextGenerationEU, GraphBRAIN plays a key role in the development and exploitation of a ‘holistic’ ontology that, starting from the aforementioned descriptions for digital libraries and archives, expands them to support the fields of history and archeology of books, intellectual property law enforcement, and economic exploitation of library and archive materials. Also, extensions to make the CH ontology compliant with, or aligned to, the CIDOC-CRM and ArCo initiatives are foreseen. Future work will also include releasing an open source version of the GraphBRAIN API.

**Acknowledgements.** This research was partially supported by projects FAIR – Future AI Research (PE00000013), spoke 6 – Symbiotic AI, and CHANGES – Cultural Heritage Active innovation for Next-Gen Sustainable society (PE00000020), Spoke 3 – Digital Libraries, Archives and Philology, under the NRRP MUR program funded by the NextGenerationEU.

## References

1. Bernasconi, E., Ceriani, M., Di Pierro, D., Ferilli, S., Redavid, D.: Linked data interfaces: a survey. *Information* **14**(9), 483 (2023)
2. Bernasconi, E., Ceriani, M., Mecella, M., Catarci, T.: Design, realization, and user evaluation of the ARCA system for exploring a digital library. *Int. J. Digit. Libr.* **24**(1), 1–22 (2023)
3. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: ArCo: the Italian cultural heritage knowledge graph. In: Ghidini, C., et al. (eds.) *ISWC 2019. LNCS*, vol. 11779, pp. 36–52. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-30796-7\\_3](https://doi.org/10.1007/978-3-030-30796-7_3)
4. Casillo, M., Santo, M.D., Mosca, R., Santaniello, D.: Sharing the knowledge: exploring cultural heritage through an ontology-based platform. *J. Ambient. Intell. Humaniz. Comput.* **14**(9), 12317–12327 (2023). <https://doi.org/10.1007/s12652-023-04652-3>
5. Celano, G.G.: An automatic morphological annotation and lemmatization for the IDP Papyri. In: *Digital Papyrology II*, pp. 139–148. De Gruyter, Berlin and Boston (2018)
6. Depauw, M., Gheldof, T.: Trismegistos: an interdisciplinary platform for ancient world texts and related information. In: Bolikowski, L., Casarosa, V., Goodale, P., Houssos, N., Manghi, P., Schirrwagen, J. (eds.) *TPDL 2013. CCIS*, vol. 416, pp. 40–52. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-08425-1\\_5](https://doi.org/10.1007/978-3-319-08425-1_5)
7. Doerr, M.: Ontologies for cultural heritage. In: Staab, S., Studer, R. (eds.) *Handbook on Ontologies. IHIS*, pp. 463–486. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-540-92673-3\\_21](https://doi.org/10.1007/978-3-540-92673-3_21)
8. Ferilli, S.: Holistic graph-based representation and AI for digital library management. In: Silvello, G., et al. (eds.) *Linking Theory and Practice of Digital Libraries, TPDL 2022. LNCS*, vol. 13541, pp. 485–489. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-16802-4\\_50](https://doi.org/10.1007/978-3-031-16802-4_50)
9. Ferilli, S., Redavid, D.: An ontology and knowledge graph infrastructure for digital library knowledge representation. In: Ceci, M., Ferilli, S., Poggi, A. (eds.) *IRCDL 2020. CCIS*, vol. 1177, pp. 47–61. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-39905-4\\_6](https://doi.org/10.1007/978-3-030-39905-4_6)
10. Ferilli, S., Redavid, D., Pierro, D.D.: Holistic graph-based document representation and management for open science. *Int. J. Digit. Libr. (IJDL)* **11:389**, 23 (2022)
11. Ferilli, S.: Integration strategy and tool between formal ontology and graph database technology. *Electronics* **10**(21), (2021). <https://www.mdpi.com/2079-9292/10/21/2616>
12. Ferilli, S.: GEAR: a general inference engine for automated multistrategy reasoning. *Electronics* **12**(2), 256 (2023)
13. Ferilli, S., Redavid, D.: An ontology and a collaborative knowledge base for history of computing. In: *1st International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH-2019)*, at the *31st International Conference on Advanced Information Systems Engineering (CAiSE 2016)*, Rome, Italy. Central Europe (CEUR) Workshop Proceedings, vol. 2375, pp. 49–60 (2019)
14. Ferilli, S., Redavid, D.: The GraphBRAIN system for knowledge graph management and advanced fruition. In: Helic, D., Leitner, G., Stettinger, M., Felfernig, A., Raś, Z.W. (eds.) *ISMIS 2020. LNCS (LNAI)*, vol. 12117, pp. 308–317. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59491-6\\_29](https://doi.org/10.1007/978-3-030-59491-6_29)

15. Gleyzes, M.A., Perret, L., Kubik, P.: Pleiades system architecture and main performances. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **39**, 537–542 (2012)
16. Grieshaber, F.: Epigraphic Database Heidelberg – data reuse options (2019)
17. Gruber, E., Matsumoto, M., Uleberg, E.: Linked open data for numismatic library, archive and museum integration. In: *Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology (Oxford 2018), CAA 2016*, pp. 55–62 (2018)
18. Haslhofer, B., Isaac, A., Simon, R.: Knowledge graphs in the libraries and digital humanities domain. In: Sakr, S., Zomaya, A.Y. (eds.) *Encyclopedia of Big Data Technologies*. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-63962-8\\_291-1](https://doi.org/10.1007/978-3-319-63962-8_291-1)
19. IFLA Study Group on the FRBR: Functional requirements for bibliographic records - final report. Technical report, International Federation of Library Associations and Institutions (IFLA), February 2009
20. Kansa, E., Kansa, S.W.: Digital data and data literacy in archaeology now and in the new decade. *Adv. Archaeol. Pract.* **9**(1), 81–85 (2021)
21. Lodi, G., et al.: Semantic web for cultural heritage valorisation. In: Hai-Jew, S. (ed.) *Data Analytics in Digital Humanities*. MSA, pp. 3–37. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-54499-1\\_1](https://doi.org/10.1007/978-3-319-54499-1_1)
22. Lohmann, S., Link, V., Marbach, E., Negru, S.: WebVOWL: web-based visualization of ontologies. In: Lambrix, P., et al. (eds.) *EKAW 2014. LNCS (LNAI)*, vol. 8982, pp. 154–158. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-17966-7\\_21](https://doi.org/10.1007/978-3-319-17966-7_21)
23. McGillivray, B., Cassotti, P., Basile, P., Di Pierro, D., Ferilli, S.: Using graph databases for historical language data: challenges and opportunities, vol. 3365, pp. 88–96. *CEUR-WS* (2023)
24. Pavlov, D., Emelyanov, Y., Mouromtsev, D., Morozov, A., Razdyakonov, D., Belyaeva, O.: Ontodia.org - a simple cognitive service to fill the gap in linked open data management tools. In: *ISWC 2016* (2016)
25. Meroño Peñuela, A., et al.: CLARIAH: enabling interoperability between humanities disciplines with ontologies. In: *Applications and Practices in Ontology Design, Extraction, and Reasoning, Studies on the Semantic Web*, vol. 49, pp. 73–90. IOS Press, November 2020
26. Riva, P., Le Boeuf, P., Žumer, M.: IFLA library reference model - a conceptual model for bibliographic information. Technical report, International Federation of Library Associations and Institutions (IFLA), December 2017
27. Robinson, I., Webber, J., Eifrem, E.: *Graph Databases*, 2nd edn. O'Reilly Media, Sebastopol (2015)
28. Žumer, M., Riva, P.: IFLA LRM - finally here. In: *Proceedings of the 2017 International Conference on Dublin Core and Metadata Applications, DCMI 2017*, pp. 13–23. Dublin Core Metadata Initiative (2017)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.














# **Health Care**



# Challenges for AI in Healthcare Systems

Markus Bertl<sup>1</sup>, Yngve Lamo<sup>2</sup>, Martin Leucker<sup>3</sup>, Tiziana Margaria<sup>4</sup>,  
Esfandiar Mohammadi<sup>5</sup>, Suresh Kumar Mukhiya<sup>6</sup>, Ludwig Pechmann<sup>3</sup>,  
Gunnar Piho<sup>1</sup>, and Fazle Rabbi<sup>7</sup>

<sup>1</sup> School of Information Technology, Tallinn University of Technology,  
12616 Tallinn, Estonia

{markus.bertl,gunnar.piho}@taltech.ee

<sup>2</sup> Department of Computer science, Electrical engineering and Mathematical  
sciences, Western Norway University of Applied Sciences, Bergen, Norway  
yngve.lamo@hvl.no

<sup>3</sup> Institute for Software Engineering and Programming Languages,  
University of Lübeck, Lübeck, Germany

{leucker,lpechmann}@isp.uni-luebeck.de

<sup>4</sup> Department of Computer Science and Information Systems,  
University of Limerick, Limerick, Ireland

tiziana.margaria@ul.ie

<sup>5</sup> Institute for IT Security, University of Lübeck, Lübeck, Germany  
esfandiar.mohammadi@uni-luebeck.de

<sup>6</sup> Tryg, Fyllingsdalen, Norway

suresh.mukhiya@tryg.no

<sup>7</sup> Information Science and Media Studies, University of Bergen, Bergen, Norway  
fazle.rabbi@uib.no

**Abstract.** This paper overviews the challenges of using artificial intelligence (AI) methods when building healthcare systems, as discussed at the AISola Conference in 2023. It focuses on the topics (i) medical data, (ii) decision support, (iii) software engineering for AI-based health systems, (iv) regulatory affairs as well as (v) privacy-preserving machine learning and highlights the importance and challenges involved when utilizing AI in healthcare systems.

**Keywords:** healthcare · artificial intelligence · medical systems

## 1 Introduction

Artificial intelligence has gained a lot of attention in previous years in various domains like finance [50], education [30], human resources [110], healthcare [91], etc. To a large part, this is due to the advances in machine learning, especially in deep learning. Moreover, the success of large language models also empowers new applications in various areas. In addition, the advances in huge dedicated

---

This work was partially supported by the BMBF project AnoMed 16KISA050K and 16KISA056.

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 165–186, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_11](https://doi.org/10.1007/978-3-031-73741-1_11)

high-performance computers with enormous GPU power drive this field of computer science. While decades ago it was more the symbolic approaches in artificial intelligence that were the driving force, it is now the sub-symbolic approaches [36]. However, it may be expected that the combination of both approaches yields solutions having the benefits of both techniques [13].

Healthcare is of great importance to society. It is also economically significant and may become even more influential in the future due to an aging society [56]. Therefore, as expected, artificial intelligence techniques are finding their way into the health domain. Healthcare affects humans' lives and is a prime example of a safety-critical domain [7, 115, 127]. Therefore, there are several challenges when applying artificial intelligence to medical applications, ranging from medical care and medicines to medical devices [114].

This paper gives a short overview of the discussions and contributions presented and partially reflected in the subsequent chapters of the healthcare track volume of the AISola Conference 2023. The track consisted of six invited presentations analyzing several challenges when applying artificial intelligence techniques to healthcare applications. We summarize these presentations in the subsequent sections and refer the reader to dedicated papers in this volume or published elsewhere. In the track, we discussed the following challenges:

- medical data,
- decision support,
- software engineering in the healthcare domain,
- regulations when building medical devices and
- privacy in machine learning.

It should be stressed that the papers contained in this volume are interim discussions of ongoing studies and not final results.

This paper is structured as follows: in the next section, we discuss the role of health data in the medical domain, focusing on machine learning applications. Section 3 gives a quick summary of challenges when building medical decision support systems. Software engineering challenges for building medical devices are sketched in Sect. 4. Regulatory requirements for building machine learning applications for the health domain are explained in Sect. 5. Section 6 recalls the privacy challenges in machine learning. Conclusions are drawn in Sect. 7.

## 2 Data Handling

A vast amount of data is generated in the healthcare domain regularly, and it consists of diverse information such as admission records, medical histories, diagnosis reports, laboratory test results, and treatment procedures from various departments and clinics. AI has the potential to analyze this extensive dataset and extract insightful information, such as comorbidity patterns, trends, and correlations, which may play an important role in improving the service quality of healthcare systems. Identifying recurring patterns within healthcare processes is crucial for streamlining healthcare procedures and ultimately improving patient

outcomes. AI provides an opportunity to reduce the cost of healthcare by optimizing processes and maximizing resource utilization, as well as providing better service quality by offering personalized treatments. Effective implementation of AI in healthcare relies on various factors, including data quality and availability; data interpretability and explainability; ethical considerations and bias; AI model complexity and selection; scalability and performance; data privacy and security; integration with existing systems; user acceptance and adoption; and regulatory compliance. In the following subsections, we present various types of healthcare data, their characteristics, and various challenges in healthcare data collection, management, analysis, and reporting.

## 2.1 Types of Healthcare Data

Healthcare data comes in various types and possess distinct characteristics. The following are different types of healthcare data that possess distinct significance for AI.

- Electronic Healthcare Records (EHRs)
- Clinical data
- Administrative data
- Genomics Data, e.g. DNA
- Patient-reported data, e.g., biological markers
- Health Behavior Data, e.g., diet, exercise, substance use
- Public Health Surveillance Data
- Research Data encompasses data
- Imaging and Diagnostic Data, e.g., X-rays, MRIs, CT scans
- Social Determinants of Health (SDOH), data encompasses factors outside the healthcare system that influence health outcomes, such as socioeconomic status, education, and environmental conditions

## 2.2 Characteristics of Healthcare Data

Healthcare data has several characteristics that distinguish it from data in other domains.

- *Complexity*: Healthcare data involves a wide range of information related to patient health, medical treatments, and administrative processes.
- *Variability*: Healthcare data exhibits variability in formats, structures, and types. It includes structured data (e.g., databases), various kinds of images, and unstructured data (e.g., free-text clinical notes).
- *Volume*: Healthcare data are often voluminous, with large datasets generated from various sources, such as diagnostic tests, medical imaging, and continuous monitoring.
- *Velocity*: Healthcare data is generated and updated in real-time. Continuous monitoring, streaming data, and rapid updates contribute to high-velocity data.

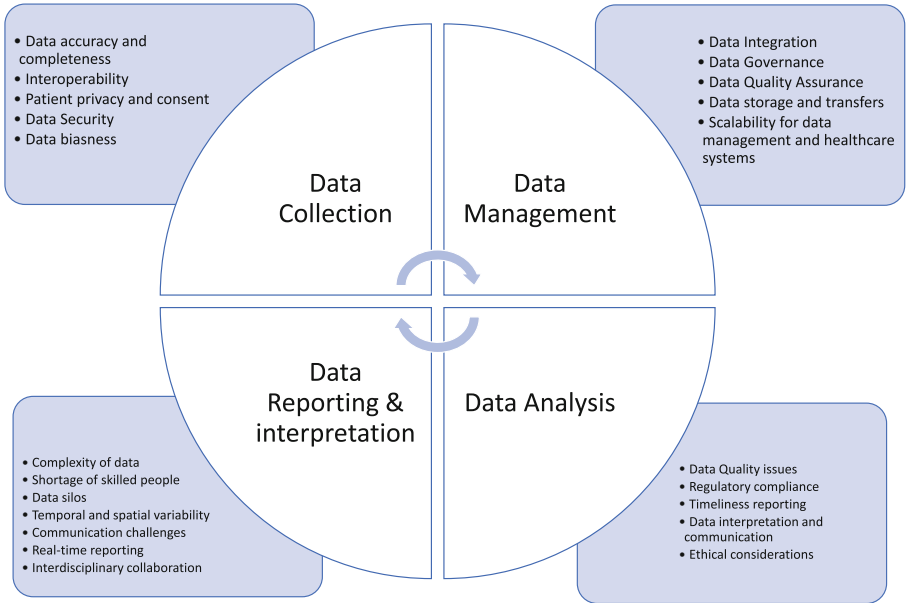
- *Variety*: Healthcare data comes in diverse formats, including text, numerical values, images, and signals. Integrating and analyzing these varied data types is a challenge.
- *Veracity*: The accuracy and trustworthiness of healthcare data can vary.
- *Privacy and Security*: Healthcare data is sensitive and subject to strict privacy regulations. Protecting patient confidentiality and ensuring data security are paramount.
- *Longitudinal*: Healthcare data often span long periods, providing a longitudinal view of a patient’s health history. This historical context is crucial for comprehensive patient care.
- *Inter-connectedness*: Different healthcare data elements are interconnected. Patient records, diagnoses, medications, and treatments are linked to provide a holistic view of care.
- *Context Dependency*: Healthcare data requires contextual understanding. Clinical data interpretation often depends on the medical context and the patient’s history.
- *Regulatory Compliance*: Healthcare data should comply with regulatory frameworks to protect patient rights and privacy, such as compliance with HIPAA, GDPR, and other data protection laws.
- *Multidimensional*: Healthcare data are often multidimensional, involving data from various sources and aspects of patient care, including clinical, financial, and operational dimensions.

### 2.3 Challenges of Healthcare Data

Addressing the challenges of healthcare data is pivotal for advancing the quality and effectiveness of healthcare services. Several vital challenges emerge as we delve into the intricacies of managing healthcare data. Figure 1 outlines different healthcare data collection, management, analysis, and reporting challenges.

*Interoperability*: There is a need for standardized interoperability to ensure the seamless exchange of healthcare data between different systems and providers. Common data standards and interoperability frameworks are necessary to facilitate better care coordination and enhance data sharing for research purposes. Even though there exist several healthcare ontologies, e.g., ICD-10 [82] and SNOMED-CT [29], the healthcare dataset often consists of unstructured data. Image processing and Natural Language Processing (NLP) may play an important role in extracting structured information from unstructured information. However, there are still many challenges as the state-of-the-art technique for image processing, and NLP could provide inaccurate results.

*Data Quality and Accuracy*: Maintaining data quality and accuracy is challenging due to errors in data entry, variations in documentation practices, and evolving standards. Implementing data validation processes and quality assurance measures are essential to ensure the reliability of healthcare data. Adopting cutting-edge techniques such as machine learning algorithms for anomaly detection [55, 118], natural language processing for semantic validation [62, 98], and



**Fig. 1.** Challenges in healthcare data collection, management, analysis and reporting

blockchain technology for immutable data records [102, 104] can strengthen data validation processes and quality assurance measures, thereby enhancing the reliability and integrity of healthcare data.

*Bias and Fairness:* Ethical and responsible AI practices are paramount when implementing AI methods in the healthcare domain. Quality of data is essential for the application of AI. Bias in data poses a significant risk to the development of responsible AI. Systems trained on biased data may produce inaccurate and harmful predictions. This bias is particularly concerning when it affects individuals from specific demographic groups. Advanced methodologies such as adversarial de-biasing [125, 126], fairness-aware learning [66, 124], and counterfactual fairness [122, 129] are pivotal for promoting fairness and mitigating bias in AI healthcare applications.

In this volume, the contribution *Towards a Multi-dimensional Health Data Analysis Framework* by Rabbi et al. [90] studies a framework for analyzing health data.

### 3 Tests vs. Learning from Massive Data Sets

The dichotomy between traditional testing methods and learning from massive data sets through Artificial Intelligence (AI) presents a compelling discourse in healthcare. Traditional testing methods in healthcare often suffer from limitations such as time inefficiency, capacity, human resources, and practicality.

In healthcare, traditional testing refers to the conventional medical testing methods that have been used for years. These methods typically involve a multi-step process that includes collecting samples from the patient at the bedside or the clinic, transporting the samples to a centralized laboratory (often located far away), and then subjecting them to several processing steps. Examples of traditional tests include blood tests, urine tests, and tissue biopsies, among others. These tests are usually performed in a laboratory by trained professionals, and the results are returned to the healthcare provider. This process can take some time, which can delay treatment [49,64].

In contrast, point-of-care testing is a more modern approach where tests are conducted close to the site of patient care, providing a rapid turnaround of test results. This can lead to improved clinical or economic outcomes compared to traditional laboratory testing [49,64]. Examples of point-of-care tests include blood glucose monitoring and home pregnancy tests [80]. Traditional testing and point-of-care testing play crucial roles in healthcare, each with its own strengths and limitations [49]. However, advances in ICT can also add to and enhance current medical testing processes.

With its capacity to learn from massive data, AI offers unprecedented opportunities for predictive analytics, pattern recognition, and decision-making support. Rule-based AI algorithms [27], machine learning [24], and deep learning algorithms [19] have been applied to healthcare with good results in terms of algorithmic accuracy, often even outperforming humans in diagnosing illnesses or predicting outcomes of treatment trajectories. However, the rise of AI in medicine also brings challenges regarding data privacy, algorithmic transparency, and validation of AI models. With increasing use, AI systems in healthcare are also increasingly targeted by cyber attacks [17]. Therefore, an optimal approach may lie in integrating both methodologies, leveraging the robustness of traditional tests and the innovative potential of AI to drive a new era of precision medicine and personalized healthcare. This approach necessitates rigorous regulatory frameworks to ensure the ethical and responsible use of data and continuous evaluation to maintain the accuracy and reliability of AI systems. Thus, the interplay between tests and learning from massive data through AI is not a competition but rather a symbiotic relationship that could redefine the future of healthcare.

Besides its potential, AI in healthcare still suffers from low adoption rates [26]. Even large tech companies like IBM with Watson for Healthcare failed to deliver on the promise of revolutionizing healthcare with AI [108]. The reasons for that were a missing interdisciplinary approach between IT researchers and healthcare professionals [21] and the complexity, quality, and large quantity of data needed for training algorithms [20]. Those factors could have also led to a bias in the good research results of machine and deep learning algorithms, especially because most published deep learning research in healthcare is only based on small datasets [20,22,25].

In conclusion, the healthcare industry stands at a crossroad between traditional testing methods and the innovative potential of AI. The integration of

AI into healthcare represents a transformative shift from traditional methodologies. While traditional testing has been the backbone of medical diagnostics for decades, it is encumbered by efficiency, capacity, and resource allocation constraints. Point-of-care testing has emerged as a viable alternative, offering rapid results and the potential for better clinical outcomes. However, the true paradigm shift lies in the application of AI, which provides a level of predictive analytics, pattern recognition, and decision-making support previously unattainable through its ability to analyze vast data sets. This fusion of AI with traditional and point-of-care testing methods does not diminish the value of either; instead, it augments the healthcare ecosystem, creating a more robust, responsive, and efficient framework for patient care. We see the future of healthcare testing as one that harmonizes the reliability of traditional methods with the agility of point-of-care solutions and the innovative prowess of AI, paving the way for a more proactive and patient-centric approach to medical diagnostics. Despite this potential, the reliance on small data sets in AI research poses a risk of bias, and the lack of robust Randomized Controlled Trials (RCTs) to validate AI's efficacy is a significant gap in the literature. Moving forward, the healthcare sector must embrace a balanced approach that leverages the strengths of traditional and AI-driven testing methods while also addressing the challenges of data representativeness and empirical validation challenges. Only through such a comprehensive strategy can we ensure the delivery of efficient, accurate, and equitable healthcare services.

Note that this section presents a summary of research on Systematic AI Support described in detail in [18]. In this volume, the contribution *Future Opportunities for Systematic AI Support in Healthcare* by Bertl et al. [23] studies further opportunities for using AI in healthcare.

## 4 Software Engineering for Developing AI-Intensive Healthcare Systems: Opportunities and Challenges

Integrating Artificial Intelligence (AI) in healthcare systems is related to the expectation to transform medicine [70], including economic and viability considerations of Disruption, Discontinuity and Differentiation (3-D-Model) [94]. It offers unprecedented opportunities to enhance patient care outcomes [48, 119] by fostering personalized care that leads to continuity of care [37] and active and healthy longevity. These opportunities can be summarized as follows:

- *Personalized Medicine:* AI's capability to analyze vast datasets enables the development of personalized treatment plans [48]. By considering an individual's genetic predisposition, lifestyle, and environmental factors [109], AI may help physicians predict the most effective evidence-based treatments, hopefully reducing medication errors and therapy selections [4].
- *Predictive Analytics:* AI algorithms can identify patterns and predict outbreaks of diseases [111], helping public health officers plan proactive healthcare measures [120]. For example, machine learning models can forecast the



spread of infectious diseases by analyzing data from various sources, including social media, citizens' mobility patterns, and climate changes, probably enabling more timely interventions [54].

- *Enhanced Diagnostic Accuracy:* AI has demonstrated superior performance in diagnosing diseases using medical image processing [67]. Deep learning models, trained on thousands of images, may help medical specialists detect anomalies such as tumors [34] and fractures [121], often with higher accuracy and speed than human experts alone can do, leading to earlier treatment and better outcomes [4, 61].
- *Operational Efficiency:* AI may help physicians streamline healthcare operations, reducing the burden on healthcare professionals and improving patient care [128]. From scheduling appointments to managing patient flow and automating administrative tasks, AI can significantly enhance efficiency in healthcare settings [78].
- *Bridging the Accessibility Gap:* AI-powered telemedicine and mobile health applications can deliver healthcare services to remote and underserved populations [28]. By reducing geographical barriers, AI has the potential to democratize access to healthcare services, making it possible for more individuals to receive timely and appropriate care [83].

#### 4.1 Challenges

However, robust, efficient, and ethical AI-intensive healthcare systems bring with them and amplify complex challenges for software engineers who engineer such systems [42]. We summarize these challenges as follows:

- *Data Privacy and Security:* The backbone of AI in healthcare is data, which often includes sensitive personal information [123]. Ensuring the privacy and security of this data is paramount, requiring robust encryption methods, secure data-sharing protocols, and compliance with regulations [35] such as HIPAA (Health Insurance Portability and Accountability Act) in the United States and GDPR in the EU.
- *Bias and Ethical Concerns:* AI systems are only as unbiased as the data they are trained on [81]. If the training data is skewed, the AI's decisions may be as well, potentially leading to unequal treatment outcomes among different demographic groups [100]. Addressing these biases and ensuring ethical considerations are integrated into AI systems is a complex challenge [69].
- *Interoperability:* Healthcare data is fragmented across various systems and formats, making it difficult to aggregate and analyze comprehensively [65]. Software engineers must tackle the interoperability challenge to enable seamless data exchange and semantic integration [9, 41, 44], ensuring AI systems can leverage diverse data sources for comprehensive analysis [103]. A further complexity dimension is that data is often processed while the representation format remains the same (e.g. TIFF files). So semantics has also the role of a refinement of the concept of type, captured for example through semantic data types in [63] and [33]. Also this needs to be included in the scope of AI tools.

- *Regulatory Compliance:* Navigating the regulatory landscape in the medical domain is challenging [101]. AI-based healthcare systems must comply with many regulations governing medical devices and patient data [14, 45, 86]. Ensuring these systems are effective and legally compliant requires a deep understanding of both technological and regulatory domains.
- *Trust and Adoption:* Building trust among healthcare professionals and patients is crucial for adopting AI-based systems [11, 40]. This involves demonstrating AI interventions' reliability, safety, and efficacy [68, 96]. Software engineers must work closely with healthcare professionals to design systems that complement clinical workflows, enhancing rather than replacing human judgment [53].

Therefore, integrating AI into healthcare systems offers a promising avenue for enhancing healthcare delivery, making personalized medicine a reality, and improving access to care. However, the journey is fraught with challenges ranging from data privacy and bias to regulatory hurdles and the need for interoperability.

## 4.2 Opportunities

As software engineering continues to evolve in response to these challenges, the collaboration between software engineers, healthcare professionals, and policy-makers will be critical [8]. By navigating these complexities, we can harness the full potential of AI in healthcare, ensuring that it serves as a tool for equitable, efficient, and effective patient care. Therefore, we summarize the following possible opportunities in software engineering for developing AI-intensive healthcare systems:

- *Advanced Tool Development:* There's a growing demand for sophisticated tools to manage and analyze health data at scale [16, 113] as well as recommender systems to manage imbalance [105]. Software engineers have the opportunity to develop and refine platforms that facilitate the training of AI models on vast datasets, including electronic health records, imaging data [31], diagnostics [99], health information [32] and genomics [39]. These tools must be powerful in terms of computational capabilities and user-friendly for healthcare professionals.
- *Interoperability Solutions:* One of the significant opportunities lies in creating solutions that ensure seamless semantic interoperability among diverse healthcare systems [88, 92, 106], including considerations of simplicity [93] in times of disruption. By engineering advanced APIs and data exchange protocols, software engineers can enable different systems to communicate effectively, enhancing data sharing and collaboration across the healthcare sector. This impacts also the availability and interoperability of data for long-lived interdisciplinary research [72].
- *Data Privacy and Security Innovations:* With the sensitivity of healthcare data, there's an urgent need for innovative solutions that protect patient

information as data breaches [6,79] harm the trust and privacy. Architectural solutions that enforce privacy [5] and blockchain-based privacy enhancing technologies have been proposed [52]. Software engineers are at the forefront of designing encryption methods, secure data storage solutions [58–60], and privacy-preserving algorithms, ensuring that AI-based systems adhere to strict data protection standards.

- *Scalable Infrastructure*: Developing AI models requires significant computational resources. There is a tremendous opportunity for software engineers to build scalable infrastructures [12,51] that can support the development and deployment of AI models, making advanced healthcare analytics accessible to institutions of all sizes.

### 4.3 AI-Intensive Healthcare Systems

However, those opportunities for software engineers are in harmony with the complex challenges that software engineers face in developing software and systems for AI-intensive healthcare systems. We summarize some of these challenges as follows:

- *Managing Complex Data*: Healthcare data is notoriously complex, heterogeneous, and voluminous [2,3,65,112]. Software engineers face the challenge of creating systems capable of handling this complexity, including different data formats, incomplete datasets, and the integration of real-time data streams, all while maintaining high performance. An example is in [31], where the sheer size of TIFF files for highly-plexed tissue image analysis required an extension to the underlying platform.
- *Ensuring Model Explainability*: AI models, especially deep learning, are often seen as “black boxes” due to their complex nature [10]. Developing methodologies and tools that enhance the transparency and explainability of these models is a significant challenge but essential for gaining trust among healthcare providers and patients [95]. Alternatively, one can use different ML techniques, where explainability can be supported by formal methods [46,47].
- *Addressing Bias and Fairness*: Data bias is a critical issue that can lead to skewed AI predictions [116]. Software engineers must devise strategies for identifying and mitigating bias in training datasets and algorithms, ensuring that AI-intensive healthcare solutions are fair and equitable [11,40,95].
- *Navigating Regulatory Landscapes*: The healthcare industry is heavily regulated, and AI-intensive systems must comply with many regulations and standards [45,101]. Software engineers must stay abreast of these evolving requirements, integrating compliance into the software engineering life-cycle, which can be complex and time-consuming.
- *Integration with Existing Healthcare IT Ecosystems*: Integrating AI solutions into existing IT infrastructures without disrupting clinical workflows represents a significant challenge [38]. Software engineers must design AI systems that are not only interoperable [89,97] but also align with the needs and processes of healthcare professionals, ensuring smooth deployment, adoption, and effective use.

Integrating AI into healthcare systems offers software engineers an opportunity-rich ground for innovation, with substantial opportunities to impact patient care positively. The combination with model driven development of software systems, including the current push towards Low-Code/No-Code can be a help. Here, approaches based on eXtreme Model Driven Development (XMDD) [76,77], based on a Digital Thread approach [71,74] and enriched by formal methods [73,75] have proven useful in many areas including cyberphysical systems, which has direct application to medical devices and Health IoT. However, the path is full of complex technical, ethical, and regulatory challenges that require thoughtful navigation. By addressing these challenges head-on and leveraging the opportunities, software engineers can play a pivotal role in shaping the future of healthcare, making it more accurate, efficient, and accessible for all by supporting transformation in medicine for enhanced patient care outcomes and personalized and continuity of care towards the active and healthy longevity of citizens.

In this volume, the contribution *Model Driven Development for AI-based Healthcare Systems: A Review* by Colm Brandon, Amandeep Singh and Tiziana Margaria [33] reviews four case studies that illustrate different quadrants in the bidimensional space of AI/ML and advanced model driven development, specifically in a low-code/no code fashion.

## 5 Regulatory Affairs

Manufacturers of Medical Devices (MD) and Software as a Medical Device (SaMD) encounter a variety of challenges when aligning with regulatory standards, both in the US under FDA regulations and in the EU under the Medical Device Regulation (MDR) [57]. These challenges are compounded by technology's dynamic and evolving nature, particularly in AI and machine learning. The US Food and Drug Administration (FDA) has a comprehensive set of regulations and guidelines for medical devices, which include detailed requirements for the safety, efficacy, and quality of these products. These regulations ensure that medical devices are safe for patients and effective in their intended use. The FDA's approach to regulation is comprehensive, covering every stage of a medical device's life cycle, from design and development to post-market surveillance. The Medical Device Regulation (MDR) plays a similar role in the European Union. The MDR sets stringent standards for medical devices, focusing on safety and performance. It emphasizes the importance of clinical evaluation and post-market surveillance, ensuring that medical devices meet high standards throughout their life cycle. Both sets of regulations demand rigorous risk management like ISO 14971, testing and validation processes, extensive documentation, and adherence to quality management systems, especially one compliant with standards like ISO 13485 [84]. The complexities in these regulatory landscapes arise from the need to balance technological innovation with patient safety and product efficacy.

The task is further complicated by the evolving nature of Artificial Intelligence (AI) and Machine Learning (ML) in medical devices, which challenges

traditional models of software change management and regulatory compliance [87]. We highlight how the adaptive capabilities of ML tasks challenge traditional software change management. Unlike conventional software, ML systems, especially those engaged in continual learning, where an ML can autonomously evolve their algorithms based on new data, blurring the lines of standard change control practices. This autonomous evolution poses unique regulatory challenges, as it may not fit within established frameworks that expect static, well-tested software versions before market release. Understanding and navigating these nuances is crucial for regulatory compliance in the ML-driven landscape of medical technology.

One approach to support evolving software for medical devices is to anticipate or predetermine the system's evolution and to foresee to which extent these changes do not affect the safety of the medical device. The FDA introduced a so-called Predetermined Change Control Plan (PCCP) [117], in which anticipated changes are described and evaluated according to their impact on the medical device. If the impact is moderate, the FDA may approve such evolving systems. However, the systematic identification of potential changes in the software and their criticality assessment is difficult. To support this complex task of identifying and documenting the development and potential changes, we introduce the CRISP-PCCP as a methodology for developing AI/ML-enabled medical devices in the context of FDA approval. CRISP-PCCP facilitates the identification of potential changes in AI/ML processes and ensures that these changes are compliant and safe. It aims to streamline the development process, focusing on quality assurance and effective project management in the complex area of medical device regulations. This makes it an important tool for manufacturers seeking FDA approval for AI/ML-enabled medical devices.

In this volume, the contribution *CRISP-PCCP - A Development Methodology Supporting FDA Approval for Machine Learning Enabled Medical Devices* by Pechmann et al. [85] explains the CRISP-PCCP approach in detail.

## 6 Privacy-Challenges in Machine Learning

Learning predictive machine learning (ML) models from patient data and similar medical applications requires a particularly careful treatment of the patient data. Recent work in ML security has illustrated that using classical learning methods can lead to models that leak information about the training data, i.e., the patient data [43]. A classical countermeasure against data leakage has been to sanitize the patient data before using it, e.g., via methods that achieve k-anonymity. Yet, data sanitization has been shown to be ineffective against deanonymization attacks [107].

Consequently, modern methods for protecting yet learning from patient data rely on data aggregation, which is what many machine learning methods are based on [1]. By solely using the patient data in aggregated statistics (e.g., the mean of gradients), the impact of single data points is limited. Furthermore, adding random noise to these statistics can prevent partial deidentification

attacks. The state-of-the-art definition used to prove that no deanonymization attack is possible against a given data processing algorithm is differential privacy.

Differentially private ML algorithms aim to protect single data points yet try to preserve an acceptable degree of usefulness, e.g., classification accuracy. For many differentially private ML algorithms, the degree of usefulness increases with increasing data points, which in medical applications translates to patients in a study. The same studies can be conducted at several medical institutions to maximize the number of patients in a study. To ensure that no party has to collect all patient data centrally, so-called secure distributed learning algorithms have been developed.

Secure distributed learning algorithms [15] ensure that no party leaks their locally collected data while each party can contribute their data to a joint learning protocol. Secure distributed learning algorithms aim to achieve performance similar to classical learning algorithms, where the data is centrally collected. Secure differentially private distributed learning algorithms additionally ensure that the result of the learning, the resulting ML model, does not leak information about the training data, i.e., the patient data.

## 7 Conclusion

In AI's integration into healthcare, a domain that directly impacts human well-being and is a complex, data-rich environment, we face many opportunities and challenges that probably impact the medical informatics future. This paper summarizes contributions and discussions at the AIsola Conference 2023. It explains opportunities where software engineering not only profits from the advancement of healthcare through AI but also faces complex challenges inherent in such a critical domain.

The promise of AI in healthcare is enormous, offering to revolutionize patient care through personalized medicine, predictive analytics, and enhanced diagnostic accuracy. The potential to streamline operational efficiencies, bridge accessibility gaps, and ultimately improve patient care outcomes and quality of life underscores a future where healthcare is accessible, effective, and more evidence-based. Software engineers find themselves at the heart of translating this potential into reality, developing the tools, systems, and algorithms that empower AI-intensive healthcare.

Yet, the path to realizing this potential has diverse and significant challenges. Data privacy and security are paramount concerns, reflecting the sensitive nature of healthcare information and the imperative to protect patient confidentiality in an increasingly digital world. AI systems' possible bias and ethical considerations further complicate the landscape, raising questions about equity and fairness in healthcare outcomes. Interoperability, regulatory compliance, and integrating AI systems into existing healthcare infrastructures present complex technical and bureaucratic issues that require understanding and innovative solutions.

In addressing these challenges, software engineering is not just a technical endeavor but a multidisciplinary one, demanding a synthesis of healthcare,

ethics, law, and beyond expertise. It requires a balance between innovation and caution, pushing the boundaries of what is possible with AI while ensuring the developed systems' safety, reliability, and fairness. The collaborative effort between software engineers, healthcare professionals, policymakers, and patients is crucial in facing the medical informatics domain's ethical, legal, and technical complexities.

The opportunities for software engineering in developing AI-intensive healthcare systems are vast, ranging from advanced tool development and interoperability solutions to innovations in data privacy and scalable infrastructure. Each opportunity enhances healthcare systems' capabilities and opens new avenues for research, development, and application in an ever-evolving field. The challenges prompt reevaluating traditional approaches and encourage a critically creative forward-thinking mindset.

As we look to the future, the convergence of AI and healthcare mediated through software engineering will probably lead to a transformative change in medicine. Yet, this convergence also requires a thoughtful approach considering healthcare's ethical, social, and technical facets. By embracing the opportunities and addressing the challenges, software engineering stands to play a pivotal role in shaping a future where AI not only enhances healthcare but does so in a manner that is equitable, secure, and deeply attuned to the democratic values of society. In this endeavor, the lessons learned and the strategies developed will benefit healthcare and offer valuable insights for applying AI across other domains, reflecting the broader implications of this work for society at large.

## References

1. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318. Association for Computing Machinery (2016)
2. Abdalla, H., Artoli, A.M.: Towards an efficient data fragmentation, allocation, and clustering approach in a distributed environment. *Information* **10**(3), 112 (2019)
3. Agrawal, R., Prabakaran, S.: Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity* **124**(4), 525–534 (2020)
4. Ahmed, Z., Mohamed, K., Zeeshan, S., Dong, X.: Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* **2020**, baaa010 (2020)
5. Alamri, B., Javed, I.T., Margaria, T.: A gdpr-compliant framework for iot-based personal health records using blockchain. In: 2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. IEEE (2021)
6. Alder, S.: Largest healthcare data breaches of 2021. *The HIPAA Journal* (2021). Accessed 10 Sep 2022
7. Ali, N., Punnekkat, S., Rauf, A.: Modeling and safety analysis for collaborative safety-critical systems using hierarchical colored petri nets. *J. Syst. Softw.* **210**, 111958 (2024)

8. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., Consortium, P.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **20**, 1–9 (2020)
9. Amar, F., April, A., Abran, A.: Electronic health record and semantic issues using fast healthcare interoperability resources: systematic mapping review. *J. Med. Internet Res.* **26**, e45209 (2024)
10. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**, 82–115 (2020)
11. Asan, O., Bayrak, A.E., Choudhury, A.: Artificial intelligence and human trust in healthcare: focus on clinicians. *J. Med. Internet Res.* **22**(6), 7 (2020)
12. Asch, M., et al.: Big data and extreme-scale computing: pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *Int. J. High Perform. Comput. Appl.* **32**(4), 435–479 (2018)
13. Azevedo, B.F., Rocha, A.M.A., Pereira, A.L.: Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Mach. Learn.* 1–43 (2024)
14. Beckers, R., Kwade, Z., Zanca, F.: The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. *Physica Med.* **83**, 1–8 (2021)
15. Bell, J.H., Bonawitz, K.A., Gascón, A., Lepoint, T., Raykova, M.: Secure single-server aggregation with (poly) logarithmic overhead. In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1253–1269 (2020)
16. Belle, A., Thiagarajan, R., Soroushmehr, S.M., Navidi, F., Beard, D.A., Najarian, K.: Big data analytics in healthcare. *BioMed Res. Int.* (2015)
17. Bertl, M.: News analysis for the detection of cyber security issues in digital healthcare: a text mining approach to uncover actors, attack methods and technologies for cyber defense. *Young Inf. Sci.* **4**, 1–15 (2019)
18. Bertl, M.: Systematic AI support for psychiatry: a framework on how to implement decision support systems. Phd thesis, Tallinn University of Technology, Tallinn, Estonia (November 2023). <https://digikogu.taltech.ee/en/Item/a77022c7-b2af-46f9-ae83-62fe04f9bc72>
19. Bertl, M., Bignoumba, N., Ross, P., Yahia, S.B., Draheim, D.: Evaluation of deep learning-based depression detection using medical claims data. *Artif. Intell. Med.* (2023)
20. Bertl, M., Kankainen, K.J.I., Piho, G., Draheim, D., Ross, P.: Evaluation of data quality in the Estonia national health information system for digital decision support. In: *Proceedings of the 3rd International Health Data Workshop. CEUR-WS* (2023)
21. Bertl, M., Klementi, T., Piho, G., Ross, P., Draheim, D.: How domain engineering can help to raise adoption rates of artificial intelligence in healthcare. In: Delir Haghghi, P., et al. *Information Integration and Web Intelligence. iiWAS 2023. LNCS*, vol. 14416, pp. 3–12. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-48316-5\\_1](https://doi.org/10.1007/978-3-031-48316-5_1)
22. Bertl, M., Metsallik, J., Ross, P.: A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. *Front. Psychiatry* **13** (2022). <https://doi.org/10.3389/fpsy.2022.923613>
23. Bertl, M., et al.: Future opportunities for systematic AI support in healthcare. In: Steffen, B. (ed.) *Bridging the Gap Between AI and Reality - First International*



- Conference, AISoLA 2023, Crete, Greece, 23–28 October 2023, Post-Proceedings (2024). this volume
24. Bertl, M., Ross, P., Draheim, D.: Predicting psychiatric diseases using AutoAI: a performance analysis based on health insurance billing data. In: Strauss, C., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DEXA 2021. LNCS, vol. 12923, pp. 104–111. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86472-9\\_9](https://doi.org/10.1007/978-3-030-86472-9_9)
  25. Bertl, M., Ross, P., Draheim, D.: A survey on AI and decision support systems in psychiatry - uncovering a dilemma. *Expert Syst. Appl.* **202**, 117464 (2022). <https://doi.org/10.1016/j.eswa.2022.117464>. <https://www.sciencedirect.com/science/article/pii/S0957417422007965>
  26. Bertl, M., Ross, P., Draheim, D.: Systematic AI Support for Decision Making in the Healthcare Sector: Obstacles and Success Factors. *Health Policy and Technology* (2023)
  27. Bertl, M., Shahin, M., Ross, P., Draheim, D.: Finding indicator diseases of psychiatric disorders in bigdata using clustered association rule mining. In: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, pp. 826–833. SAC '23, Association for Computing Machinery (2023). <https://doi.org/10.1145/3555776.3577594>
  28. Bhatt, P., Liu, J., Gong, Y., Wang, J., Guo, Y.: Emerging artificial intelligence-empowered mhealth: scoping review. *JMIR Mhealth Uhealth* **10**(6), e35053 (2022)
  29. Bodenreider, O., Cornet, R., Vreeman, D.J.: Recent developments in clinical terminologies - snomed ct, loinc, and rxnorm. *Yearb. Med. Inform.* **27**, 129–139 (2018)
  30. Boscardin, C.K., Gin, B., Golde, P.B., Hauer, K.E.: Chatgpt and generative artificial intelligence for medical education: potential impact and opportunity. *Acad. Med.* **99**(1), 22–27 (2024)
  31. Brandon, C., et al.: Cinco de bio: a low-code platform for domain-specific workflows for biomedical research. *BioMedInformatics* (to appear) (2024)
  32. Brandon, C., Doherty, A.J., Kelly, D., Leddin, D., Margaria, T.: Hipp: health information portal for patients and public. *Appl. Sci.* **13**(16), 9453 (2023)
  33. Brandon, C., Singh, A., Margaria, T.: Model driven development for AI-based healthcare systems: a review. In: Steffen, B. (ed.) Bridging the Gap Between AI and Reality - First International Conference, AISoLA 2023, Crete, Greece, 23–28 October 2023, Post-Proceedings (2024). this volume
  34. Chartrand, G., et al.: Deep learning: a primer for radiologists. *Radiographics* **37**(7), 2113–2131 (2017)
  35. Chenthara, S., Ahmed, K., Wang, H., Whittaker, F.: Security and privacy-preserving challenges of e-health solutions in cloud computing. *IEEE Access* **7**, 74361–74382 (2019)
  36. Ciatto, G., Sabbatini, F., Agiollo, A., Magnini, M., Omicini, A.: Symbolic knowledge extraction and injection with sub-symbolic predictors: a systematic literature review. *ACM Comput. Surv.* (2024)
  37. Cohen, E., Lindman, I.: Importance of continuity of care from a patient perspective—a cross-sectional study in Swedish health care. *Scand. J. Prim. Health Care* **42**(1), 195–200 (2024)
  38. Directorate-General for Health and Food Safety: A european health data space: harnessing the power of health data for people, patients and innovation. European Commission (2022). [https://health.ec.europa.eu/document/download/17c7065c-c432-445f-9b27-8ccf283581bc\\_en?filename=com\\_2022-196\\_en.pdf](https://health.ec.europa.eu/document/download/17c7065c-c432-445f-9b27-8ccf283581bc_en?filename=com_2022-196_en.pdf). Accessed 28 Jan 2024

39. Dlamini, Z., Francies, F.Z., Hull, R., Marima, R.: Artificial intelligence (AI) and big data in cancer and precision oncology. *Comput. Struct. Biotechnol. J.* **18**, 2300–2311 (2020)
40. Esmaeilzadeh, P.: Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC Med. Inform. Decis. Mak.* **20**, 1–19 (2020)
41. Fennelly, O., Moroney, D., Doyle, M., Eustace-Cook, J., Hughes, M.: Key interoperability factors for patient portals and electronic health records: a scoping review. *Int. J. Med. Inform.* 105335 (2024)
42. Floridi, L., et al.: Ai4people—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind. Mach.* **28**, 689–707 (2018)
43. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333 (2015)
44. Gansel, X., Mary, M., van Belkum, A.: Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. *Eur. J. Clin. Microbiol. Infect. Dis.* **38**(6), 1023–1034 (2019). <https://doi.org/10.1007/s10096-019-03501-6>
45. Gerke, S., Babic, B., Evgeniou, T., Cohen, I.G.: The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit. Med.* **3**(1), 53 (2020)
46. Gossen, F., Margaria, T., Steffen, B.: Towards explainability in machine learning: the formal methods way. *it prof.* **22** (4), 8–12 (2020)
47. Gossen, F., Margaria, T., Steffen, B.: Formal methods boost experimental performance for explainable ai. *it prof.* **23** (6), 8–12 (2021)
48. Gouripur, K., et al.: The impact of artificial intelligence on healthcare: a revolution in progress. *North and West Lond. J. Gen. Pract.* **10**(1) (2024)
49. Gubala, V., Harris, L.F., Ricco, A.J., Tan, M.X., Williams, D.E.: Point of care diagnostics: status and future. *Anal. Chem.* **84**(2), 487–515 (2012)
50. Hidayat, M., Defitri, S.Y., Hilman, H.: The impact of artificial intelligence (AI) on financial management. *Manag. Stud. Bus. J. (PRODUCTIVITY)* **1**(1), 123–129 (2024)
51. Hu, H., Wen, Y., Chua, T.S., Li, X.: Toward scalable systems for big data analytics: a technology tutorial. *IEEE Access* **2**, 652–687 (2014)
52. Javed, I.T., Alharbi, F., Margaria, T., Crespi, N., Qureshi, K.N.: Petchain: a blockchain-based privacy enhancing technology. *IEEE Access* **9**, 41129–41143 (2021)
53. Jeyaraman, M., Ramasubramanian, S., Balaji, S., Jeyaraman, N., Nallakumarasamy, A., Sharma, S.: Chatgpt in action: harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. *World J. Methodol.* **13**(4), 170–178 (2023)
54. Jing, F., Li, Z., Qiao, S., Zhang, J., Olatosi, B., Li, X.: Using geospatial social media data for infectious disease studies: a systematic review. *Int. J. Digit. Earth* **16**(1), 130–157 (2023)
55. Kavitha, M., Srinivas, P.V.V.S., Kalyampudi, P., F, C.S., Srinivasulu, S.: Machine learning techniques for anomaly detection in smart healthcare. In: *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1350–1356 (2021). <https://doi.org/10.1109/ICIRCA51532.2021.9544795>

56. Khanna, N.N., et al.: Economics of artificial intelligence in healthcare: Diagnosis vs. treatment. *Healthcare* **10**(12) (2022). <https://doi.org/10.3390/healthcare10122493>. <https://www.mdpi.com/2227-9032/10/12/2493>
57. Klar, E.: Medical Device Regulation als aktuelle Herausforderung für die rechtssichere Einführung neuer Technologien. *Chirurg* **89**(10), 755–759 (2018). <https://doi.org/10.1007/s00104-018-0705-3>
58. Klementi, T., Kankainen, K.J.I., Piho, G., Ross, P.: Prospective research topics towards preserving electronic health records in decentralised content-addressable storage networks. In: HEDA@ Petri Nets, p. 14 (2022)
59. Klementi, T., Piho, G.: Method and system for managing data using decentralized content-addressable storage networks. European patent Office, priority number EP24166173.5 (March 2024). submitted patent
60. Klementi, T., Piho, G., Ross, P.: A reference architecture for resolving health data accessibility, comprehensiveness, and ownership dilemmas through preserving semantically interoperable personal health records in decentralized content-addressable storage networks. (2024), will be published soon
61. Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C., Faisal, A.A.: The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**(11), 1716–1720 (2018)
62. Kraljevic, Z., et al.: Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artif. Intell. Med.* **117**, 102083 (2021). <https://doi.org/10.1016/j.artmed.2021.102083>. <https://www.sciencedirect.com/science/article/pii/S0933365721000762>
63. Lamprecht, A., Naujokat, S., Margaria, T., Steffen, B.: Semantics-based composition of EMBOSS services. *J. Biomed. Semant.* **2**(S-1), S5 (2011). <http://www.jbiomedsem.com/content/2/S1/S5>
64. Larkins, M.C., Thombare, A.: Point-of-care testing. In: StatPearls [Internet]. StatPearls Publishing (2023)
65. Li, H., et al.: Review on security of federated learning and its application in healthcare. *Futur. Gener. Comput. Syst.* **144**, 271–290 (2023)
66. Liu, M., et al.: Fairness-aware interpretable modeling (faim) for trustworthy machine learning in healthcare (2024)
67. Liu, X., et al.: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**(6), e271–e297 (2019)
68. Liu, X., et al.: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *Lancet Digit. Health* **2**(10), e537–e548 (2020)
69. Magrabi, F., et al.: Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications. *Yearb. Med. Inform.* **28**(01), 128–134 (2019)
70. Maleki Varnosfaderani, S., Forouzanfar, M.: The role of AI in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering* **11**(4), 337 (2024)
71. Margaria, T., Chaudhary, H.A.A., Guevara, I., Ryan, S., Schieweck, A.: The interoperability challenge: building a model-driven digital thread platform for CPS. In: Margaria, T., Steffen, B. (eds.) *ISoLA 2021*. LNCS, vol. 13036, pp. 393–413. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-89159-6\\_25](https://doi.org/10.1007/978-3-030-89159-6_25)
72. Margaria, T., Floyd, B.D., Gonzalez Camargo, R., Lamprecht, A.-L., Neubauer, J., Seelaender, M.: Simple management of high assurance data in long-lived interdisciplinary healthcare research: a proposal. In: Margaria, T., Steffen, B. (eds.)

- ISoLA 2014. LNCS, vol. 8803, pp. 526–544. Springer, Heidelberg (2014). [https://doi.org/10.1007/978-3-662-45231-8\\_44](https://doi.org/10.1007/978-3-662-45231-8_44)
73. Margaria, T., Kubczak, C., Steffen, B.: Bio-jeti: a service integration, design, and provisioning platform for orchestrated bioinformatics processes. *BMC Bioinformatics* **9**(4), S12 (2008). <https://doi.org/10.1186/1471-2105-9-S4-S12>
  74. Margaria, T., Schieweck, A.: The digital thread in industry 4.0. In: Ahrendt, W., Tapia Tarifa, S.L. (eds.) IFM 2019. LNCS, vol. 11918, pp. 3–24. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-34968-4\\_1](https://doi.org/10.1007/978-3-030-34968-4_1)
  75. Margaria, T., Steffen, B.: Backtracking-free design planning by automatic synthesis in metaframe. In: Astesiano, E. (ed.) FASE 1998. LNCS, vol. 1382, pp. 188–204. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0053591>
  76. Margaria, T., Steffen, B.: Business process modeling in the jabc: the one-thing approach. In: *Handbook of Research on Business Process Modeling*, pp. 1–26. IGI Global (2009)
  77. Margaria, T., Steffen, B.: Extreme model-driven development (xmdd) technologies as a hands-on approach to software development without coding. *Encyclopedia Educ. Inf. Technol.* 732–750 (2020)
  78. Munavalli, J.R., Boersma, H.J., Rao, S.V., Van Merode, G.: Real-time capacity management and patient flow optimization in hospitals using AI methods. *Artif. Intell. Data Min. Healthc.* 55–69 (2021)
  79. Murray-Watson, R.: Healthcare data breach statistics. *The HIPAA Journal* (2021). Accessed 10 Sep 2022
  80. Nichols, J.H.: Utilizing point-of-care testing to optimize patient care. *EJIFCC* **32**(2), 140 (2021)
  81. Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al.: Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(3), e1356 (2020)
  82. Organization., W.H.: ICD-10 : international statistical classification of diseases and related health problems/World Health Organization. World Health Organization Geneva, 10th revision, 2nd edn. (2004)
  83. Patibandla, R.L., Rao, B.T., Murty, M.R.: Revolutionizing diabetic retinopathy diagnostics and therapy through artificial intelligence: a smart vision initiative. In: *Transformative Approaches to Patient Literacy and Healthcare Innovation*, pp. 136–155. IGI Global (2024)
  84. Pechmann, L., Mildner, M., Suthau, T., Leucker, M.: Regulatorische anforderungen an lösungen der künstlichen intelligenz im gesundheitswesen. In: Pfannstiel, M.A. (ed.) *Künstliche Intelligenz im Gesundheitswesen: Entwicklungen, Beispiele und Perspektiven*, pp. 175–198. Springer Fachmedien, Wiesbaden (2022). [https://doi.org/10.1007/978-3-658-33597-7\\_8](https://doi.org/10.1007/978-3-658-33597-7_8)
  85. Pechmann, L., Potdevin, Y., Brehmer, K., Nowotka, D., Leucker, M.: Crisp-pccp – a development methodology supporting fda approval for machine learning enabled medical devices. In: Steffen, B. (ed.) *Bridging the Gap Between AI and Reality - First International Conference, AISoLA 2023, Crete, Greece, 23–28 October 2023, Post-Proceedings* (2024). this volume
  86. Pesapane, F., Volonté, C., Codari, M., Sardanelli, F.: Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the united states. *Insights Imaging* **9**, 745–753 (2018)
  87. Petersen, E., et al.: Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE Access* **10**, 58375–58418 (2022). <https://doi.org/10.1109/ACCESS.2022.3178382>

88. Piho, G., Tepandi, J., Thompson, D., Woerner, A., Parman, M.: Business archetypes and archetype patterns from the hl7 rim and openehr rim perspectives: towards interoperability and evolution of healthcare models and software systems. *Procedia Comput. Sci.* **63**, 553–560 (2015)
89. Raab, R., et al.: Federated electronic health records for the European health data space. *Lancet Digit. Health* **5**(11), e840–e847 (2023)
90. RABBI, F., Mukhiya, S.K., Norge, T., Lamo, Y.: Towards a multi-dimensional health data analysis framework. In: Steffen, B. (ed.) *Bridging the Gap Between AI and Reality - First International Conference, AISoLA 2023, Crete, Greece, 23–28 October 2023, Post-Proceedings* (2024). this volume
91. Rahman, M.A., Victoros, E., Ernest, J., Davis, R., Shanjana, Y., Islam, M.R.: Impact of artificial intelligence (AI) technology in healthcare sector: a critical evaluation of both sides of the coin. *Clinical Pathology* **17**, 2632010X241226887 (2024)
92. Randmaa, R., Bossenko, I., Klementi, T., Piho, G., Ross, P.: Evaluating business meta-models for semantic interoperability with fhir resources. In: *HEDA@Petri Nets* (2022)
93. Rasche, C., Margaria, T., Floyd, B.D.: Service model innovation in hospitals: beyond expert organizations. *Service Business Model Innovation in Healthcare and Hospital Management: Models, Strategies, Tools*, pp. 1–20 (2017)
94. Rasche, C., Reinecke, A.A., Margaria, T.: Künstliche Intelligenz im Gesundheitswesen als Kernkompetenz? Status quo, Entwicklungslinien und disruptives Potenzial, pp. 49–79. *Springer Fachmedien Wiesbaden, Wiesbaden* (2022). [https://doi.org/10.1007/978-3-658-33597-7\\_2](https://doi.org/10.1007/978-3-658-33597-7_2)
95. Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M., St Amant, R.: Recent advances in trustworthy explainable artificial intelligence: status, challenges, and perspectives. *IEEE Trans. Artif. Intell.* **3**(6), 852–866 (2021)
96. Rivera, S.C., et al.: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *Lancet Digit. Health* **2**(10), e549–e560 (2020)
97. Saelaert, M., Mathieu, L., Van Hoof, W., Devleeschauwer, B.: Expanding citizen engagement in the secondary use of health data: an opportunity for national health data access bodies to realise the intentions of the European health data space. *Arch. Public Health* **81**(1), 168 (2023)
98. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J.: Natural language processing and the representation of clinical data. *J. Am. Med. Inform. Assoc.* **1**(2), 142–160 (1994). <https://doi.org/10.1136/jamia.1994.95236145>
99. Seelaender, M., et al.: Inflammation in cachexia (2015)
100. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**(12), 2176–2182 (2021)
101. Shah, V., Konda, S.R.: Cloud computing in healthcare: Opportunities, risks, and compliance. *Revista Espanola de Documentacion Cientifica* **16**(3), 50–71 (2022)
102. Shen, B., Guo, J., Yang, Y.: Medchain: efficient healthcare data sharing via blockchain. *Appl. Sci.* **9**(6) (2019). <https://doi.org/10.3390/app9061207>. <https://www.mdpi.com/2076-3417/9/6/1207>
103. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **22**(5), 1589–1604 (2018). <https://doi.org/10.1109/JBHI.2017.2767063>

104. Simpson, G., Nana, L., Kester, Q.A.: A centralized data validation system model for healthcare systems based on blockchain. In: 2021 International Conference on Cyber Security and Internet of Things (ICSIoT), pp. 55–58 (2021). <https://doi.org/10.1109/ICSIoT55070.2021.00019>
105. Singh, A., Minguett, O., Margaria, T.: Binary decision diagrams and composite classifiers for analysis of imbalanced medical datasets. *Electron. Commun. EASST* **82** (2023)
106. Söerd, T., Kankainen, K., Piho, G., Klementi, T., Ross, P.: Towards specification of medical processes according to international standards and semantic interoperability needs. In: MODELSWARD, pp. 160–167 (2023)
107. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data - anonymisation groundhog day. In: 31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, 10–12 August 2022, pp. 1451–1468. USENIX Association (2022)
108. Strickland, E.: Ibm watson, heal thyself: how ibm overpromised and underdelivered on ai health care. *IEEE Spectr.* **56**(4), 24–31 (2019)
109. Subramanian, M., et al.: Precision medicine in the era of artificial intelligence: implications in chronic disease management. *J. Transl. Med.* **18**, 1–12 (2020)
110. Sucipto, H.: The impact of artificial intelligence (AI) on human resource management practices. *Manag. Stud. Bus. J. (PRODUCTIVITY)* **1**(1), 138–145 (2024)
111. Surya, L.: How government can use ai and ml to identify spreading infectious diseases. *Int. J. Creat. Res. Thoughts (IJCRT)* 2320–2882 (2018). ISSN
112. Syed, R., et al.: Digital health data quality issues: systematic review. *J. Med. Internet Res.* **25**, e42615 (2023)
113. Tariq, M.U.: Advanced wearable medical devices and their role in transformative remote health monitoring. In: Transformative Approaches to Patient Literacy and Healthcare Innovation, pp. 308–326. IGI Global (2024)
114. Tettey, F., Parupelli, S.K., Desai, S.: A review of biomedical devices: classification, regulatory guidelines, human factors, software as a medical device, and cybersecurity. *Biomed. Mater. Dev.* **2**(1), 316–341 (2024)
115. Turner, N., Baxter, D.: How to enable ambidexterity in safety-critical software development: this article reports on a safety-critical development project that achieved ambidexterity and highlights four leadership behaviors that can enhance project outcomes. *Res. Technol. Manag.* **67**(2), 35–43 (2024)
116. Ueda, D., et al.: Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn. J. Radiol.* **42**(1), 3–15 (2024)
117. U.S. Food and Drug Administration: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions, April 2023. <https://www.fda.gov/media/166704/download>
118. Šabić, E., Keeley, D., Henderson, B., Nannemann, S.: Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data. *AI Soc.* **36**(1), 149–158 (2021). <https://doi.org/10.1007/s00146-020-00985-1>
119. Wachter, R.M., Brynjolfsson, E.: Will generative artificial intelligence deliver on its promise in health care? *JAMA* **331**(1), 65–69 (2024)
120. Wang, Q., Su, M., Zhang, M., Li, R.: Integrating digital technologies and public health to fight COVID-19 pandemic: key technologies, applications, challenges and outlook of digital healthcare. *Int. J. Environ. Res. Public Health* **18**(11), 6053 (2021)
121. Wang, S., Summers, R.M.: Machine learning and radiology. *Med. Image Anal.* **16**(5), 933–951 (2012)

122. Wastvedt, S., Huling, J.D., Wolfson, J.: An intersectional framework for counterfactual fairness in risk prediction. *Biostatistics*, p. kxad021 (August 2023). <https://doi.org/10.1093/biostatistics/kxad021>
123. Wu, Q., Chen, X., Zhou, Z., Zhang, J.: Fedhome: cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Trans. Mob. Comput.* **21**(8), 2818–2832 (2020)
124. Wu, Z., He, J.: Fairness-aware model-agnostic positive and unlabeled learning. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1698–1708. FAccT '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3531146.3533225>
125. Yang, J., Soltan, A., Eyre, D., Yang, Y., Clifton, D.: An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Med.* **6**(1) (2023)
126. Yang, J., Soltan, A.A.S., Yang, Y., Clifton, D.A.: Algorithmic fairness and bias mitigation for clinical machine learning: Insights from rapid COVID-19 diagnosis by adversarial learning. *medRxiv* (2022). <https://doi.org/10.1101/2022.01.13.22268948>, <https://www.medrxiv.org/content/early/2022/01/14/2022.01.13.22268948>
127. Yao, Y., Han, T., Yu, J., Xie, M.: Uncertainty-aware deep learning for reliable health monitoring in safety-critical energy systems. *Energy* **291**, 130419 (2024)
128. Zheng, Y., Wang, L., Feng, B., Zhao, A., Wu, Y.: Innovating healthcare: the role of chatgpt in streamlining hospital workflow in the future. *Ann. Biomed. Eng.* **52**(4), 750–753 (2024)
129. Zuo, A., Wei, S., Liu, T., Han, B., Zhang, K., Gong, M.: Counterfactual fairness with partially known causal graph. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) *Advances in Neural Information Processing Systems*, vol. 35, pp. 1238–1252. Curran Associates, Inc. (2022). [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/08887999616116910fccc17a63584b5-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/08887999616116910fccc17a63584b5-Paper-Conference.pdf)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.







# Towards a Multi-dimensional Health Data Analysis Framework

Fazle Rabbi<sup>1</sup>(✉), Bahareh Fatemi<sup>1</sup>, Suresh Kumar Mukhiya<sup>2</sup>,  
and Yngve Lamo<sup>3</sup>

<sup>1</sup> Information Science and Media Studies, University of Bergen, Bergen, Norway  
{fazle.rabbi,bahareh.fatemi}@uib.no

<sup>2</sup> Tryg Norge, Bergen, Norway  
suresh.mukhiya@tryg.no

<sup>3</sup> Department of Computer science, Electrical engineering and Mathematical sciences, Western Norway University of Applied Sciences, Bergen, Norway  
yngve.lamo@hvl.no

**Abstract.** Healthcare processes need to be streamlined to offer better healthcare services. Data analysis can be crucial in reducing costs, optimizing processes, and analyzing treatment effectiveness. However, data analysis in healthcare is complex due to the variety and complexity of patient data. This paper proposes a multi-dimensional comparative analysis method that offers healthcare professionals a lens to delve into healthcare datasets from various perspectives. The paper discusses the importance of comparative analysis in healthcare illustrated by two examples on how we can understand the pattern of comorbidity and how we can analyze the effectiveness of internet delivered psychological interventions. The paper presents a multi-dimensional comparative analysis framework covering various use cases in analysing healthcare data. The framework allows healthcare professionals to compare and contrast healthcare data across multiple dimensions, including clinical dimensions such as diagnosis, outcome measures, time dimension, patient dimensions (engagement, involvement), cost dimension, and other relevant factors. This approach offers a more insightful understanding of healthcare data and facilitates informed decision-making in healthcare practices.

**Keywords:** Healthcare data · Comorbidity · Abstraction · Category theory · Knowledge graph · Context

## 1 Introduction

Healthcare expenses are increasing due to various factors such as rising costs of medicines and healthcare equipment, complexity of managing chronic diseases and comorbidity conditions, and growing number of mental illness. Identifying recurring patterns within healthcare processes is crucial for streamlining healthcare procedures and ultimately improving patient outcomes. Data analysis provides an opportunity to reduce the cost in healthcare by optimizing processes and



maximizing resource utilization. Data analysis can also be used to analyze the effectiveness of various treatment options. However, data analysis in the healthcare domain is complex due to the variety of patients' complex healthcare-related issues. Healthcare professionals need to analyse data in the healthcare domain, which comes from multiple sources, for example, admission records, diagnosis reports, laboratory test results, and treatment procedures from various departments and clinics. Although there exist several healthcare ontologies, for example, ICD-10 [13] and SNOMED-CT [1], there needs to be a uniform framework for data analysis in the healthcare domain.

Researchers in the healthcare domain also carry out research activities that may require the use of new data formats. Researchers often use randomized controlled trials (RCTs) to assess the effectiveness of a particular intervention, treatment, or medical approach. RCTs serve as a gold standard for assessing treatment interventions. The primary purposes of conducting randomized controlled trials include determining causation, evaluating effectiveness. Analyzing the data from RCTs requires thorough investigation from various perspectives to identify potential variables that influence the intervention and treatment method. To provide a cost-effective solution to treat mental illness, several RCTs have been conducted. Internet-Delivered Psychological Treatment (IDPT) systems have the potential to provide evidence-based mental health treatments for a far reaching population at a lower cost [10, 11]. In [9], the author presented a framework to develop an adaptive IDPT system that can adapt psychological interventions according to the users need, context, and preferences.

In this paper, we focus on the analysis of identifying recurring patterns within healthcare processes. This is a crucial need for streamlining healthcare procedures and ultimately improving patient outcomes. We propose to use a multi-dimensional comparative analysis method which offers a lens through which healthcare professionals can delve into the healthcare dataset from various perspective and allows them to explore diverse use-cases that span from evaluating treatment interventions to understanding the patterns of patients with comorbidities. The paper discusses the importance of comparative analysis in understanding the effectiveness of interventions such as dropout rates in RCTs. A multi-dimensional comparative analysis framework is proposed as a means to explore patient patterns and tailor interventions to enhance engagement.

The paper is organized as follows. In Sect. 2, we present several needs for comparative analysis in the healthcare domain. In Sect. 3 we present a multi-dimensional comparative analysis framework and show the applicability of the framework in analyzing healthcare dataset. In Sect. 4 we present closely related work and in Sect. 5 we conclude the paper.

## 2 Needs for Comparative Analysis

In this section, we present the need for comparative analysis in the healthcare domain. Comparative analysis can be a valuable tool for making informed decisions about healthcare. Below is a list of use-cases of the application of comparative analysis in the healthcare domain:

- Comparing the cost-effectiveness of two different treatments for a chronic disease.
- Comparing the long-term outcomes of different surgical procedures.
- Comparing the effectiveness of different public health interventions.
- Discovering the pattern of patients with comorbidity.

In the following subsections we present two specific cases that can benefit from comparative analysis of healthcare data.

## 2.1 Discovering the Pattern of Comorbidity Patients

Comorbidity introduces a higher risk of complications in many cases. For instance, diabetes and cardiovascular disease are often comorbid [5], and their coexistence can increase the risk of heart-related complications in patients. The treatment of comorbidity is not straightforward due to many reasons, such as delayed diagnosis, medication interactions, and side effects. The combined effects of different conditions can result in greater physical and mental health challenges that can have a cumulative impact on the well-being of the patients and can lead to a reduced quality of life. Therefore, understanding the patterns of comorbidities is essential as it will allow healthcare professionals to identify potential risks and take steps to prevent the worsening of conditions. However there are many challenges to discover the pattern of comorbidity patients and their progression of diseases. We need an efficient technique to extract the necessary and relevant information from healthcare data. Eliminating noise and irrelevant information from healthcare data is essential for data analysis. However, this task is not easy due to the huge amount of data captured in the healthcare system from multiple sources. A comparative analysis method which can filter healthcare data across various dimensions and abstraction levels would be useful for analyzing the pattern of comorbidities, as it would allow healthcare professionals to effectively test out their hypothesis about comorbidity patterns.

## 2.2 Comparing the Effectiveness of Psychological Interventions

Randomized control trials are often used to measure the effectiveness of different treatment interventions, such as IDPT, for mental illness. In randomized control trials, populations are separated into two groups: (i) the experimental group that receives the intervention that is being tested and (ii) the control group that receives an alternative (conventional) treatment.

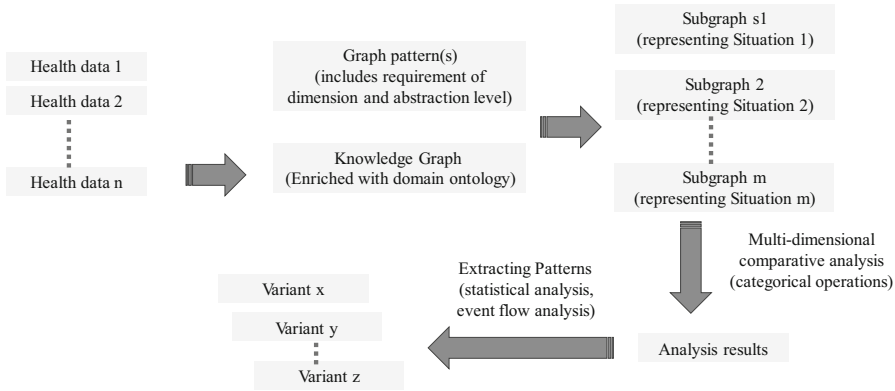
The outcome of treatment in these groups is then followed up to determine the effectiveness of the interventions. The multi dimensional analysis could be means for understanding causal effect in IDPT. In current approaches, the focus is on measuring the effect change but too little focus was provided on the causal relationships. While measuring the effectiveness of treatment interventions, it is essential to understand the subjects, determine the treatments' parameters, and address potential challenges such as dropout rates, particularly in IDPT. Various factors can influence dropouts in IDPT systems. Significant causes include a)

lack of participation, b) technical issues, c) privacy concerns, d) perceived ineffectiveness of online therapies, e) time constraints, f) lack of support or proper guidance during online therapy, g) unexpected life events, h) limited motivation, i) complexity of the program, and j) poor user experience [11].

Understanding these causes can guide developing and implementing strategies to reduce dropout rates, such as improving user engagement, promptly addressing technical issues, enhancing privacy measures, and providing adequate support throughout treatment. To effectively address and mitigate the identified causes of dropout in IDPT programs, a multi-dimensional analysis framework is essential to adapt the IDPT program based on the user's needs and preferences. A multi-dimensional comparative analysis framework proposed in this paper allows us to discover patient patterns. Using such patterns, IDPT systems can proactively tailor interventions. Equipped with insight into patient behaviors, the IDPT system can dynamically adapt its content, pacing, and interactive elements to enable real-time adjustments to address issues such as lack of engagement, perceived ineffectiveness, or technical challenges.

### 3 Proposed Framework

We present a conceptual framework for the analysis of health data in Fig. 1 that supports comparative analysis from a variety of perspectives. Data from various sources are enriched with healthcare ontologies e.g., ICD-10, SNOMED-CT, and stored in a knowledge graph. The framework allows us to search for healthcare data across various dimensions and abstraction levels using graph patterns.

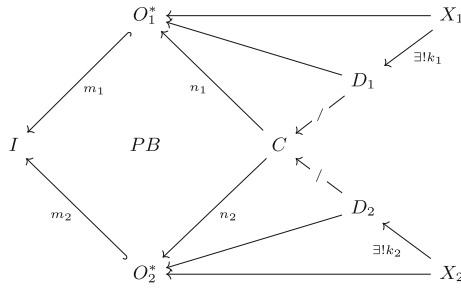


**Fig. 1.** Conceptual framework for healthcare data analysis

The proposed framework incorporates a multi-dimensional comparative analysis method which is based on categorical operations. The results of this comparative analysis help identify patterns among patients who share similar issues or

disease trajectories. In this framework, we emphasize on augmenting healthcare data with domain ontology which is allowing us to do comparison and extract patterns and variants from various perspectives and abstraction levels. While we demonstrate a proof of concept using a graph database in this paper, the framework is also adaptable to SQL databases, employing set and relational operations as an alternative implementation approach.

Figure 2 illustrates an overview of the categorical operations for the comparative analysis. A knowledge graph is represented in the figure as  $I$ ; Subgraphs  $O_1^*, O_2^*$  indicate two objects that are subject to comparison. These subgraphs may include the diagnosis or the symptoms of some patients. We identify the commonality of  $O_1^*, O_2^*$  by a pullback operation which is represented as object  $C$  in the figure. From  $O_1^*$  and  $C$  we can compute the object  $D_1$  that represents the nodes and relationships present in  $O_1^*$  but not in  $C$ . Similarly we can compute  $D_2$ . With these objects, we can perform statistical analysis to measure the similarity and dissimilarity of different aspects of the healthcare data.

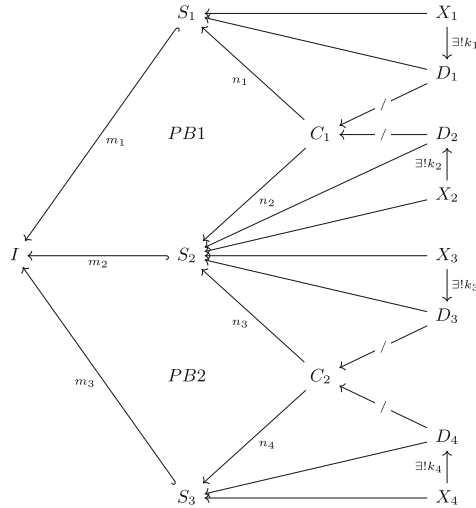


**Fig. 2.** Pullback object ( $C$ ) computes the commonality between  $O_1^*$  and  $O_2^*$ ;  $D_1$  and  $D_2$  objects are used to compute the dissimilarities between  $O_1^*$  and  $O_2^*$

The idea of using comparative analysis can be further extended to study the progression of diseases among certain population of patient individuals. Figure 3 shows some categorical operations that can be performed to perform computational analysis to find the progression of diseases of some groups of patients. Subgraphs  $S_1, S_2, S_3..$  shows the weekly progression of diseases of a patient group. This approach can also be used to analyze the progression of symptoms or the movement of patients in the hospital in various departments or clinics.

### 3.1 Comparative Analysis for Understanding Comorbidity

In our previous work [6], we presented a validated learning approach for healthcare process analysis which incorporates two sequential processes. The first step is to identify groups of patients with various comorbidity issues and the second step is to analyze the progression of diseases in those groups of patients. The first step of the process relies on a community detection technique to identify patient



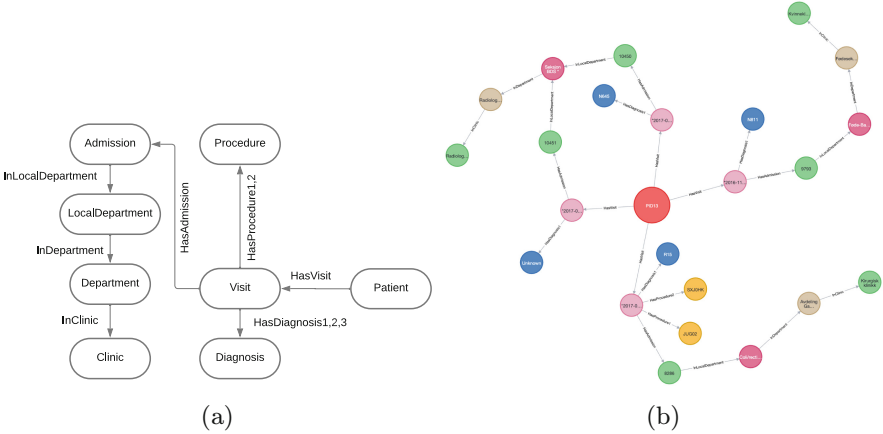
**Fig. 3.** Capturing the progression of diseases with pullback operation

groups with comorbidity diseases; and the second step involves human input regarding potential pattern of disease progression. In this step, the user makes different hypothesis about disease progression and the hypothesis is validated by extracting evidences from a healthcare dataset. We presented a variant of Linear Temporal Logic (*LTL*) [16] language to specify the following pattern of comorbidity issues:

- Diseases that appear one after another.
- Diseases that develop one after the other over short or extended periods of time.
- Diseases that occur after a continuous period of a condition such as high blood pressure.
- Diseases that appear continuously and at determinable time intervals.

While community detection proves to be beneficial in identifying certain groups of patients sharing common diseases, its application has limitations when analyzing comorbidity. This method offers limited insights about the factors that contribute to the comorbidity cases. If there is a group of patients who suffer from disease  $d_1$  and  $d_2$ , it would be important to study the dissimilarity of these groups of patients with patients who are diagnosed with one of the diseases, but not both. For example, we may be interested in studying patients who have diabetes or kidney problems or both. A community detection algorithm may find a patient group who have both diseases but we also need to study the patients who have one of the diseases but not both. Hence, a framework for comparative analysis is required. A multi-dimensional comparative analysis framework would be appropriate to deal with the detailed analysis of the progression of diseases for patients with comorbidities.

Figure 4a presents a schema of a knowledge graph that allows us to instantiate a knowledge graph with patients' clinical information such as admission, diagnosis, procedure, etc. The knowledge graph is enriched with the ICD-10 ontology. Figure 4b shows an instance of this graph schema that represents the information about the administration of a patient in our dataset.



**Fig. 4.** Model for patient profile and its relationship with ICD-10 and department ontologies [6] (4a), graph instance for patient with id=PID13 (4b).

In this section, we elaborate three comparative analysis tasks that can be performed using our categorical approach:

1. **Comorbidity analysis:** Extracting information about patients who have been diagnosed with multiple diseases.
2. **Progression of diseases:** Analysing the progression of diseases of some patient groups.
3. **Comparison of progression of diseases:** Analysing the progression of diseases of two different patient groups.

### Comorbidity Analysis

Suppose we wish to study the patients who have been diagnosed with ‘*diseases of the digestive system (K00-K95)*’ and ‘*diseases of the circulatory system (I00-I99)*’. In Fig. 2, subgraph  $O_1^*$  and subgraph  $O_2^*$  represents patients with ‘*diseases of the digestive system*’ and ‘*diseases of the circulatory system*’ respectively. A pullback operation would give us object  $C$  representing patients who have been diagnosed with both diseases. We can extract the patients individuals from the knowledge graph using the Cypher query. Table 1 shows the Cypher query to extract this pullback object. From subgraph  $O_1^*$  and the pullback object  $C$  we can

compute the object  $D_1$  which includes patient individuals who have been diagnosed with ‘*diseases of the digestive system*’ but have not been diagnosed with ‘*diseases of the circulatory system*’. While this comparative study allows us to identify the number of patients who have such comorbidity issues, we can use similar comparative analysis to check the progression of diseases or symptoms or the movement of patients flow in different departments or clinics.

**Table 1.** Cypher query for computation of pullback: patients diagnosed with ‘*diseases of the digestive*’ and ‘*diseases of the circulatory systems*’

---

Cypher Query

---

```
MATCH (p1:Patient)-[]→(v1:Visit)-[]→(d1:Diagnosis {Level1:'K00-K95'})
MATCH (p2:Patient)-[]→(v2:Visit)-[]→(d2:Diagnosis {Level1:'I00-I99'})
WHERE p1 = p2 RETURN DISTINCT p1;
```

---

## Progression of Diseases

Suppose we are interested to study the progression of disease in patients who have been diagnosed with ‘*diseases of the digestive system*’. To study this progression, we can perform the categorical operations in Fig. 3 to carry out the computational analysis. Here in this situation, the subgraphs  $S_1, S_2, S_3..$  indicate the monthly progression of diseases of a group of patients who have been diagnosed with ‘*diseases of the digestive system*’. The Cypher query used to extract this progression is shown in Table 2.

**Table 2.** Cypher query for extracting the progression of diseases in patients with ‘*diseases of the digestive system*.’

---

Cypher Query

---

```
MATCH (p1:Patient)-[]→(v1:Visit)-[]→(d1:Diagnosis {Level1:'K00-K95'})
WHERE date(v1.visitDate) ≥ date("2015-01-01") AND
date(v1.visitDate) ≤ date("2015-01-31")
WITH collect(d1) AS d1_collection
MATCH (p2:Patient)-[]→(v2:Visit)-[]→(d2:Diagnosis)
WHERE date(v2.visitDate) ≥ date("2015-02-01") AND
date(v2.visitDate) ≤ date("2015-02-28") AND p1 = p2 AND
NOT d2 IN d1_collection RETURN DISTINCT d2;
```

---

### Comparison of progression of diseases

In Fig. 5 we present another comparative analysis using categorical operations where we compare the progression of diseases of two different patient groups e.g., the weekly progression of patients diseases who have been diagnosed with ‘*diseases of the digestive system*’. The subgraph  $\alpha_1, \alpha_2, \alpha_3$  represents the diseases of patients with ‘*diseases of the circulatory system*’ who have also been diagnosed with ‘*diseases of the digestive system*’ and  $\beta_1, \beta_2, \beta_3$  represent the diseases of patients who have been diagnosed with ‘*diseases of the digestive system*’, but NOT ‘*diseases of the circulatory system*’. This comparative analysis will give us an idea of the variety of diseases that patients with ‘*diseases of the circulatory system*’ are prone to suffer while they are diagnosed with ‘*diseases of the digestive system*’. The Cypher query employed for comparing the progression of diseases of two patient groups is shown in Table 3. This comparative analysis can be performed at a variety of abstraction levels. Figure 6 illustrates a comparative analysis performed over two different abstraction level. The subgraphs  $\alpha_1, \alpha'_1$  represent a group of patients diagnosis with lower and higher abstraction level (respectively). The progression of diseases in this group of patients is shown in subgraphs  $\alpha_2, \alpha'_2$ . These patient groups progression of diseases can be compared with another patient groups (i.e., subgraphs  $\beta_1, \beta'_1$  and  $\beta_2, \beta'_2$ ) by means of categorical operations as shown in Fig. 6.

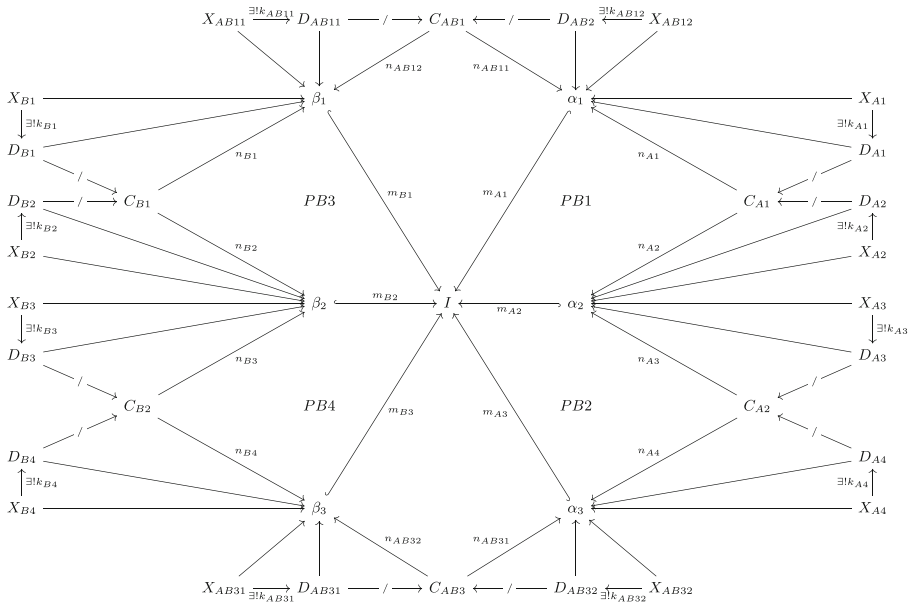


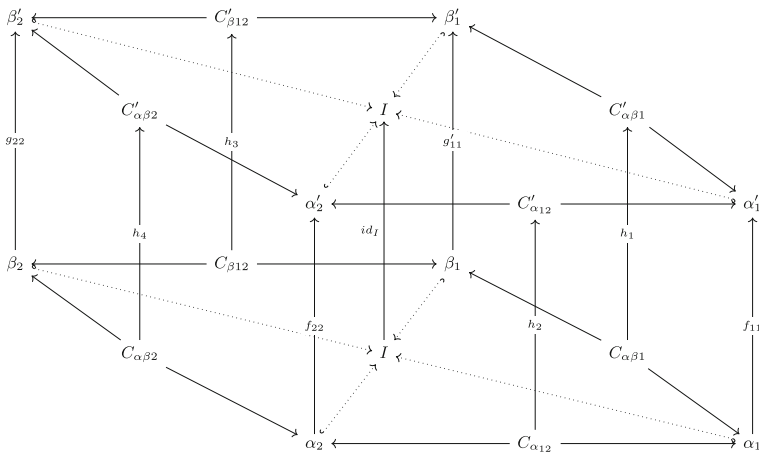
Fig. 5. Comparison of progression of diseases



**Table 3.** Cypher query for comparative analysis of diseases

```

Cypher Query
MATCH (p1:Patient)-[]→(v1:Visit)-[]→(d1:Diagnosis {Level1:'K00-K95'})
WHERE date(v1.visitDate) ≥ date("2015-01-01") AND
date(v1.visitDate) ≤ date("2015-12-31")
WITH collect(d1) AS d1_collection
MATCH (p2:Patient)-[]→(v2:Visit)-[]→(d2:Diagnosis {Level1:'I00-I99'})
WHERE date(v2.visitDate) ≥ date("2015-01-01") AND
date(v2.visitDate) ≤ date("2015-12-31") AND NOT d2 IN d1_collection
RETURN DISTINCT d2;
    
```



**Fig. 6.** Comparison of progression of diseases at a higher level of abstraction

Figure 7 illustrates a computational model for analyzing two sets of progressions.  $\alpha_1, \alpha_2, \alpha_3..$  (resp.  $\beta_1, \beta_2, \beta_3..$ ) represents the progression of situations (e.g., diseases, symptoms) at a certain abstraction level  $j$ ;  $\alpha'_1, \alpha'_2, \alpha'_3..$  (resp.  $\beta'_1, \beta'_2, \beta'_3..$ ) represents the progression of situations specified at a higher level of abstraction  $k$ .  $C_{\alpha12}$  is the pullback object of  $\alpha_1 \rightarrow I$  and  $\alpha_2 \rightarrow I$ . Similarly,  $C_{\alpha23}$ ,  $C_{\beta12}$  and  $C_{\beta23}$  are the other pullback objects computed from the situations represented at level  $j$ . The evolution of progressions are represented at a higher level of abstraction at the top of the figure in  $\alpha_1, \alpha_2, \alpha_3..$  (resp.  $\beta_1, \beta_2, \beta_3..$ ). Figure 7 shows the co-limit objects  $CL_\alpha, CL_\beta$  at level  $j$  which represents combined commonality of the progressions. The pullback of  $r, t$  (resp.  $r', t'$ ) is shown in the figure as  $Z$  ( $Z'$ ). The pullback object  $Z$  represents the commonality in the progressions in  $\alpha_1, \alpha_2, \alpha_3..$  and  $\beta_1, \beta_2, \beta_3..$ . An empty pullback object  $Z$  would indicate that the progressions in  $\alpha_1, \alpha_2, \alpha_3..$  and  $\beta_1, \beta_2, \beta_3..$  are considered to be different at abstraction level  $j$ .

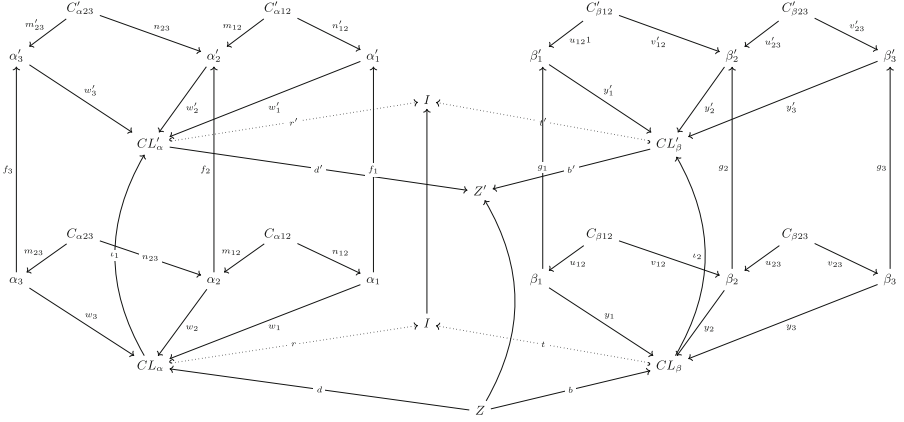


Fig. 7. Computational model for analyzing two sets of progressions.

### Variant Identification

The result of the comparative analyses can be used to identify variants. In [6], we proposed a linear temporal logic language  $LTL_{EOT}$  to represent a variety of comorbidity patterns for determining patient variants.  $LTL_{EOT}$  allows us to formulate queries to find patients who exhibit similar patterns in their healthcare event log consisting of patients admission, diagnosis, procedure related information with timestamp. The syntax of  $LTL_{EOT}$  includes the incorporation of the following:

- ontologies so that the base case of the satisfaction relation refers to an instance of ontology concept; and,
- specific time intervals.

With this formulation, we specify comorbidity patterns, specifying the diagnoses and time intervals between events. A few examples are shown below:

- Find all events of patients who are diagnosed with iron deficiency anemia (D50), right after being diagnosed with chronic kidney disease (N18):

$$\diamond_{(\geq 0s)}(N18) \wedge \square_{(\geq 0s)}\left((N18) \rightarrow \bigcirc_{(\geq 0s)}(D50)\right) \quad (1)$$

- Find all the events of patients who have been diagnosed with diabetes mellitus (E08-E13) and within 2 years have been diagnosed with chronic kidney disease (N18):

$$\diamond_{(\geq 0s)}(E08 - E13) \wedge \square_{(\geq 0s)}\left((E08 - E13) \rightarrow \diamond_{(\leq 2\text{ years})}(N18)\right) \quad (2)$$

The formulas include temporal operators,  $\square$  (always),  $\diamond$  (eventually),  $\bigcirc$  (next-time), which enables us to specify properties of events as they evolve over time.

### 3.2 Comparative Analysis for IDPT

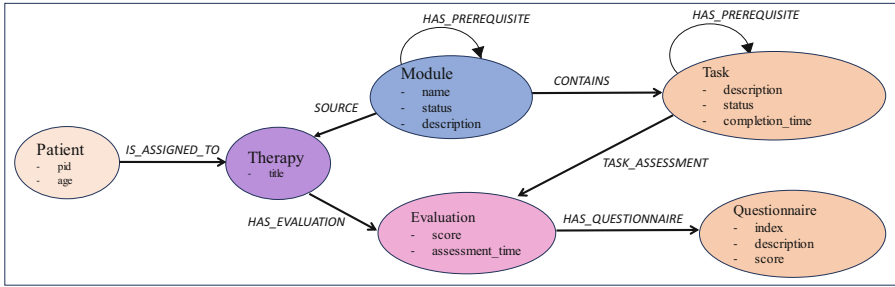
In this section, we present an application of comparative analysis for analyzing the effectiveness of treatment interventions. Patients suffering from mental health morbidities can be assigned one or more therapies (IDPT). IDPT involves one or more modules. Each module has some prerequisites that each patient must fulfil to complete. Each module consists of different tasks. These tasks can be interactive or informative. In Fig. 8, we present a knowledge graph schema which allows us to store information about patients involvement in IDPT (adapted from [9]). *Informative tasks* provide learning materials about mental health issues, therapy, symptoms, use cases, and several ways to manage them. The main objective of such educational materials is to provide psycho-education so that:

- Patients and their associated families can learn about symptoms, causes, remedies and treatment concepts;
- Patients can understand self-help programs and steps required to manage their illness;
- Patients can correlate their situations with similar others, which helps to vent their frustrations.
- Such educational materials are in the form of reading tasks (text), listening (audio), and watching (video).

*Interactive tasks* differ from informative tasks in that they involve user interactions, often in the form of exercises. Such exercises can be physical activities or computerized tasks. Examples of physical activities include workouts and mindfulness exercises like breathing exercises, walking certain distances, stretching, or performing other activities. Examples of computerized exercises include filling in the blanks, answering questions (Q/A), multiple-choice questions (MCQ), and providing feedback. Feedback tasks involve using free text, rating systems, or multiple-choice questions.

Evaluation is essential in all the parts of IDPT. Each task and module has an evaluation method. Overall evaluation of module tasks gives the gross evaluation of the therapy.

*Analyzing Drop-out in IDPT:* Dropout is a major problem in IDPT programs. There might be a variety of reasons for the dropout and it is essential to understand the cause of dropouts. Our proposed comparative analysis framework would allow researchers to study the dropout cases from a variety of perspectives. For example, in Attention Deficit/Hyperactivity Disorder (ADHD) interventions, we might be interested to study the role of concentration issues against completing modules and tasks. The comparative analysis method of two patients group as shown in Fig. 2 can be adopted for analyzing dropout of patients. We may represent subgraphs  $O_1^*$ ,  $O_2^*$  consisting of patients who have concentration issues and dropped out from the intervention program accordingly. The pullback object  $C$  would indicate the patients who have concentration issue and dropped out from the intervention program. Similarly we can analyze the dominance of other parameters that influences the drop-out from IDPT programs.



**Fig. 8.** Knowledge graph schema for Internet Delivered Psychological Treatments

*Analyzing the Dominance of Symptoms:* The comparative analysis method of capturing the progression of issues as presented in Fig. 3 may be used to analyze the significance of certain symptoms among patients. Suppose we wish to study the effectiveness of modules in a therapeutic program among patients who have high concentration issue. The subgraphs  $S_1, S_2, S_3, \dots$  in Fig. 3 may represent patients who have started module  $M_1, M_2, M_3, \dots$  (respectively) and the patients have high concentration issue found at the time of starting each module. A categorical pullback of  $S_1 \rightarrow I$  and  $S_2 \rightarrow I$  would give us patients who have concentration issue after completing module  $M_1$ ; The object  $D_1$  represent the patients whose concentration problem have been resolved after the completion of module  $M_1$ .

*Personalized Treatment:* In order to develop personalized treatment interventions it is required to identify variants of patient groups. In [10], Mukhiya et al., presented a rule based approach for adaptive IDPT. Comparative analysis would provide us a mechanism to identify the variants of various patients group who have participated in a treatment program.

The comparative analysis as presented in Fig. 5 may be used to analyze the significance of certain mental conditions of patient groups and the effectiveness of a therapeutic program. We may represent subgraphs  $\alpha_1, \alpha_2, \alpha_3$  consisting of patients completion of modules over different weeks who have scored more than  $X$  in Adult ADHD Self-Report Scale (ASRS); and  $\beta_1, \beta_2, \beta_3$  may represent subgraphs consisting of patients completion of modules over different weeks who have scored more than  $Y$  in Patient Health Questionnaire (PHQ-9). This comparative analysis would indicate the effectiveness of individual modules in reducing certain mental conditions in a therapeutic program. This knowledge can be useful identify variants of patient population and offer personalized treatments using adaptive treatment method.

## 4 Related Works

In [14], Partington et al., presented a comparative analysis of clinical processes of four Australian hospitals. They presented the use of process-mining tech-

nique to discover the control-flow and the performance of the processes at each hospital and the discovered process models were used for comparison. Through an exploratory approach they have identified four comparative points based on known drivers of costs and/or patient health outcomes: the proportion of patients admitted to an inpatient care setting; the throughput timing between ED presentation and movement to an inpatient setting (Admission); the frequency of procedures (diagnostic/treatment) provided; the total length of stay for patients. Partington et al., did not present any comparative analysis technique for identifying patterns of comorbidity issues and effectiveness of treatment interventions. In this paper, we have presented the need for comparative analysis across various healthcare scenarios and presented a generic framework for comparative analysis.

Scientists from various fields of research have investigated comorbidity and have used different methodologies to deal with its complexities. To identify the prevalence of comorbidities of mental and behavioral disorders, Cha et al. [4] propose an analysis based on association rules. Boytcheva et al. [3] propose a cascade data mining approach specifically tailored for frequent pattern mining in comorbidity studies. Several studies fall into the category of network analysis. Jones et al. [8] define four network statistics to identify symptoms that connect two mental disorders. Social network analysis and graph theory have also been used to understand the comorbidity of two chronic diseases [7]. Bottrighi et al. [2] proposed a knowledge-based approach to run-time comorbidity management to support physicians during the execution of the Clinical Practice Guidelines (CPGs) on a specific patient. Piovesan et al. [15] using Computer-Interpretable Guidelines, the history of the status of the patient, and the log of the clinical actions executed on them, propose an Answer Set programming-based method for the treatment of comorbid patients. However, these works do not generalize the need for comparative analysis for analyzing comorbidity patterns and do not promote the use of a multi-dimensional comparative analysis framework as we have proposed in this paper.

The study [12] examined the impact of weekly SMS reminders on adherence to an IDPT for adults with ADHD. The results indicated that the overall module completion, logins, and coping strategy practice slightly improved. Although SMS reminders can influence engagement, they do not uniformly improve overall adherence. To enhance the effectiveness of interventions such as the self-guided IDPT for ADHD, the proposed multi-dimensional framework for analyzing user patterns can be employed, especially in the context of SMS reminders. The framework considers user engagement patterns, allowing for the identification of personalized strategies to improve adherence. By tailoring reminders based on a comprehensive analysis of user patterns, the self-guided IDPT can be adapted to better suit the diverse needs and preferences of adults with ADHD, ultimately fostering more sustained engagement and positive outcomes.

## 5 Conclusion

In this paper, we explored the significance of comparative analysis in the healthcare domain and proposed a multi-dimensional comparative analysis framework to address two challenges of data analysis in healthcare: (i) Comparing the effectiveness of different public health interventions; and (ii) Discovering the pattern of patients with comorbidity. We presented a formal approach to address the common need for comparative analysis in healthcare. The proposed framework facilitates comparative analysis of healthcare datasets from various perspectives, essential for comprehending the patterns of patients with comorbidities. The paper also highlighted the importance of comparative analysis in understanding the effectiveness of interventions and presented computational methods based on formal method techniques. The presented approach is generic enough to be applicable in various healthcare scenarios.

## References

1. Bodenreider, O., Cornet, R., Vreeman, D.J.: Recent developments in clinical terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb. Med. Inform.* **27**, 129–139 (2018)
2. Bottrighi, A., Piovesan, L., Terenziani, P.: Run-time support to comorbidities in GLARE-SSCPM. In: HEALTHINF, pp. 498–505 (2019)
3. Boytcheva, S., Angelova, G., Angelov, Z., Tcharaktchiev, D.: Mining comorbidity patterns using retrospective analysis of big collection of outpatient records. *Health Inf. Sci. Syst.* **5**(1), 1–9 (2017)
4. Cha, S., Kim, S.S.: Discovery of association rules patterns and prevalence of comorbidities in adult patients hospitalized with mental and behavioral disorders. *Healthcare* **9**, 636 (2021)
5. Davis, J., Chung, R., Juarez, D.: Prevalence of comorbid conditions with aging among patients with diabetes and cardiovascular disease. *Hawaii Med. J.* **70**, 209–13 (2011)
6. Fatemi, B., Rabbi, F., MacCaull, W.: A validated learning approach to healthcare process analysis through contextual and temporal filtering. In: Koutny, M., Bergenthum, R., Ciardo, G. (eds.) *Transactions on Petri Nets and Other Models of Concurrency XVII. Lecture Notes in Computer Science*( ), vol. 14150, pp. 108–137. Springer, Berlin, Heidelberg (2024). [https://doi.org/10.1007/978-3-662-68191-6\\_5](https://doi.org/10.1007/978-3-662-68191-6_5)
7. Hossain, M.E., Khan, A., Uddin, S.: Understanding the comorbidity of multiple chronic diseases using a network approach. In: *Proceedings of the Australasian Computer Science Week Multiconference*, pp. 1–7 (2019)
8. Jones, P.J., Ma, R., McNally, R.J.: Bridge centrality: a network approach to understanding comorbidity. *Multivar. Behav. Res.* **56**(2), 353–367 (2021)
9. Mukhiya, S.K.: A software framework for adaptive and interoperable internet-delivered psychological treatments. Ph. D. thesis, Høgskulen på Vestlandet (2021). <https://hvlopen.brage.unit.no/hvlopen-xmlui/handle/11250/2778982>
10. Mukhiya, S.K., Ahmed, U., Rabbi, F., Pun, K.I., Lamo, Y.: Adaptation of IDPT system based on patient-authored text data using NLP. 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 226–232 (2020). <https://doi.org/10.1109/CBMS49503.2020.00050>, <https://api.semanticscholar.org/CorpusID:221474223>

11. Mukhiya, S.K., Wake, J.D., Inal, Y., Lamo, Y.: Adaptive systems for internet-delivered psychological treatments. *IEEE Access* **8**, 112220–112236 (2020). <https://doi.org/10.1109/ACCESS.2020.3002793>
12. Nordby, E.S., et al.: The effect of SMS reminders on adherence in a self-guided internet-delivered intervention for adults with ADHD. *Front. Digit. Health* **4**, 821031 (2022)
13. Organization., W.H.: ICD-10 : international statistical classification of diseases and related health problems / World Health Organization. World Health Organization Geneva, 10th revision, 2nd ed. edn. (2004)
14. Partington, A., Wynn, M., Suriadi, S., Ouyang, C., Karnon, J.: Process mining for clinical processes: a comparative analysis of four Australian hospitals. *ACM Trans. Manage. Inf. Syst.* **5**(4), 1–8 (2015). <https://doi.org/10.1145/2629446>
15. Piovesan, L., Terenziani, P., Dupré, D.T.: Conformance analysis for comorbid patients in answer set programming. *J. Biomed. Inform.* **103**, 103377 (2020)
16. Pnueli, A.: The temporal logic of programs. In: 18th Annual Symposium on Foundations of Computer Science (sfcs 1977), pp. 46–57 (1977). <https://doi.org/10.1109/SFCS.1977.32>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Future Opportunities for Systematic AI Support in Healthcare

Markus Bertl<sup>1</sup>(✉) , Gunnar Piho<sup>1</sup> , Dirk Draheim<sup>1</sup> , Peeter Ross<sup>1,2</sup> ,  
Ludwig Pechmann<sup>3</sup> , Nicholas Bucciarelli<sup>4,5</sup>, and Rahul Sharma<sup>1,6</sup>

<sup>1</sup> School of Information Technology, Tallinn University of Technology,  
12616 Tallinn, Estonia

{markus.bertl,gunnar.piho,peeter.ross}@taltech.ee,  
sharmarahul@akgec.ac.in

<sup>2</sup> Department of Research, East-Tallinn Central Hospital,  
Ravi 18, 10138 Tallinn, Estonia

<sup>3</sup> Institute for Software Engineering and Programming Languages,  
University of Lübeck, Maria-Goeppert-Str. 3, 23562 Lübeck, Germany  
lpechmann@unitransferklinik.de

<sup>4</sup> Cornell University, 616 Thurston Avenue, Ithaca, NY 14853, USA  
Ncb57@cornell.edu

<sup>5</sup> Rochester Institute of Technology, One Lomb Memorial Drive, Rochester,  
NY 14623, USA

<sup>6</sup> Department of Information Technology, Ajay Kumar Garg Engineering College,  
27th Km Milestone, Delhi-Meerut Expressway, P.O. Adhyatmik Nagar,  
Ghaziabad 201009, India

**Abstract.** Artificial Intelligence (AI) holds transformative potential to revolutionize healthcare delivery and outcomes. However, the literature suggests that focusing solely on AI algorithms leads to low adoption rates. AI needs to be introduced systematically into healthcare. This paper builds on this approach and synthesizes existing literature and authors' insights to critically examine the current landscape and future opportunities for systematic AI support in healthcare. The multifaceted applications of AI, ranging from disease prediction to personalized medicine, are explored with a focus on AI's potential to optimize employee performance, alleviate healthcare staff burdens, and enhance patient care. However, challenges such as limited access to unbiased data sets, connectivity issues, and ethical concerns pose significant barriers to AI adoption in healthcare.

**Keywords:** Healthcare · Electronic health record · Digital Decision Support · Artificial Intelligence · Machine learning · Medical data reuse · Primary use · Secondary use · Genome data · personalized medicine · preventive medicine

## 1 Introduction

Artificial Intelligence (AI) can be a game-changer in the healthcare industry, potentially revolutionizing how healthcare is delivered and received. According

© The Author(s) 2025

B. Steffen (Ed.): AISO LA 2023, LNCS 14129, pp. 203–224, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_13](https://doi.org/10.1007/978-3-031-73741-1_13)



to Sharma [66], AI has been used in the integrated management of cancer, to support the diagnosis and prediction of function changes in urinary bladder control, in stroke prediction as well as in risk prediction for cardiovascular diseases, and to support the decision-making process of diagnosis. In addition, AI can assist in disease diagnosis, drug discovery, and personalized medicine [20]. Secinaro et al. [65] state that AI can help predict and prevent diseases, improve clinical trials, and enhance patient outcomes.

According to [20], AI applications in healthcare can be broadly categorized into three groups: patient-oriented AI, clinician-oriented AI, and administrative and operational-oriented AI. Patient-oriented AI includes tasks such as medical record review, population health trending and analytics, therapeutic drug and device design, reading radiology images, making clinical diagnoses and treatment plans, and even talking with patients. Clinician-oriented AI includes answering the phone, reviewing medical records, and making clinical diagnoses and treatment plans. Administrative and operational-oriented AI includes tasks such as medical record review, population health trending and analytics, and answering the phone.

The benefits of AI in healthcare can be summed up as follows: AI can improve and optimize the performance and productivity of employees, reduce the burden on healthcare staff, and improve patient care and treatment [20,66]. Nevertheless, the integration of AI into healthcare faces several hurdles. These include the scarcity of high-quality, unbiased data sets, challenges with internet connectivity, the absence of robust systems and processes that facilitate adoption, gaps in mindset and knowledge among stakeholders, and significant ethical considerations [5,65,66].

One of the issues of implementing high-quality and unbiased data sets is the semantic heterogeneity of health data. In an ideal world, AI models are trained based on routine clinical and health data. The routine clinical data are considered precious [81], and their secondary use [60] is considered beneficial for policymakers, public health officers, scientists, clinicians, citizens, and industry [37]. Several initiatives, including the European Health Data Space [22], are searching for solutions to use routine clinical and health data in global content. However, as stated in a survey [71], we currently do not have a unified approach for the semantic heterogeneity of health data and use divide-and-conquer approaches instead. Such semantic heterogeneity-related issues are highlighted also in 2024 [2,28].

AI is often called a game changer for healthcare. However, recent studies show that it is often over-promised and under-delivered [6,8,10,72]. In previous research, we showed obstacles and success factors [11] and introduced the idea of “systematic AI support” to bring AI algorithms for digital decision support into production [6].

Based on our previous work, the current paper analyzes and critically discusses future opportunities for systematic AI support based on a literature review and informed arguments from the authors. Firstly, AI algorithms require large amounts of high-quality data for training, but healthcare data is often frag-

mented, incomplete, or inconsistent [49]. We explore this challenge in detail in Sect. 3.1 and provide a potential solution in Sect. 4.1. Secondly, there are significant ethical and legal considerations, such as patient privacy and data security, that must be addressed [82]. Therefore, we look into the EU Medical Device Regulation, the EU AI Act, and FDA requirements in Sect. 3.2. Thirdly, the generalizability of AI algorithms is often limited, as models trained on data from one population or healthcare system may not perform well when applied to another [16]. This is where personalized medicine comes into play, tailoring healthcare to individual patients based on their predicted response or risk of disease. Lastly, the complexity of AI algorithms often leads to a lack of transparency, or a “black box” problem, making it difficult for clinicians to understand and trust the output of AI (Sect. 4.3). Further challenges of AI in healthcare, like domain complexity and ethical challenges, have already been comprehensively researched, e.g., in [3,33]. These challenges necessitate careful consideration and ongoing research to ensure AI’s safe and effective application in healthcare. We discuss the results of the current paper in Sect. 5, and we conclude in Sect. 6.

## 2 Related Work

In [76], Pasi Tyrväinen et al. report on an idea-generating design thinking [18] project in service of understanding the potential of AI applications in healthcare. Stakeholders from various groups, including AI users, technology providers, and resource providers, have been involved in a series of design thinking workshops to innovate on the utilization of AI in healthcare in four specific focus areas, namely digital hospital, aging, preventive healthcare, and wellbeing related to sports. A total of 34 best-use case descriptions have been condensed from the overall idea pool of the conducted workshops.

In [64], Deepti Saraswat et al. conducted a systematic literature review (SLR) to investigate the utilization of explainable AI (EXAI) in healthcare applications. For this purpose, they distinguish between Healthcare 4.0 [35,36], and Healthcare 5.0, where Healthcare 4.0 stands for a patient-centric healthcare paradigm that involves sensor-driven analytics, and Healthcare 5.0 [51] transcends Healthcare 4.0 into a healthcare paradigm that involves “smart control, interpretable healthcare analytics, three-dimensional view models, and augmented and virtual reality” [64]. Starting from the general investigation of “evolution and technology trends of EXAI in healthcare 5.0 applications” [64], the SLR in [64] contributes more specifically by analyzing the operational challenges and the data collection process as requirements of EXAI in Healthcare 5.0.

In [74], Vaibhav Thakare et al. discuss the challenges and opportunities in utilizing AI and IoT in healthcare, focusing on applicability, trends, and potential future developments. According to [74], AI and IoT show potential to “improve the accessibility of preventative public health services and transition our current secondary and tertiary healthcare to be a more proactive, continuous, and coordinated system” [74].

In [32], Akshit Garg et al. show a series of medical AI solutions for diagnosis, prognosis, and treatment, particularly in fighting the COVID-19 pandemic.

They also give an account of challenges to the wide-scale adoption of AI in the healthcare sector.

In [61], Bhupesh Rawat et al. analyze a series of AI applications in regard to their potential for personalized medicine and disease diagnosis, i.e., in regard to “their ability to provide faster and more accurate predictions, risk stratification, and improved outcomes through augmented intelligence” [61]. According to [61], the challenges of the utilization of AI algorithms and applications in healthcare are data privacy, ethics, and regulation.

In [55], Balakrishnan Mullachery outlines a vision for how emerging technologies from the fields of 5G, Internet of Things (IoT), geographical information systems (GIS), and AI technologies can be orchestrated in service of smart healthcare, eventually enabling “a future smart-connected society emphasizing healthcare and well-being” [55].

### 3 Current Challenges of AI in Healthcare

#### 3.1 Health Data-Related Challenges and Issues

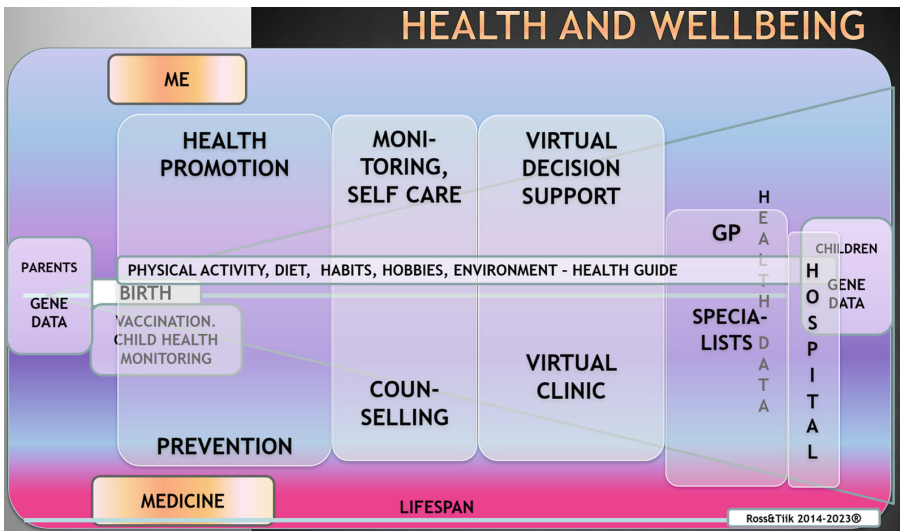


Fig. 1. Variety of a person’s health data

Health data plays a pivotal role in the healthcare sector, serving as a cornerstone for informed decision-making, research, and the delivery of quality patient care [48,83]. It encompasses a wide range of information related to individuals’ health and well-being, including medical history, treatment plans, laboratory results, and demographic details [23,50]. Ross and Tiik have illustrated (Fig. 1) the

veracity of a person's health data [62]. Some of these data types are electronic health records (EHRs), data generated during clinical encounters, public health data, surveys, and questionnaires, data on lifestyle factors and patient-reported outcomes, data from smartphones, gene sequencing data, environmental health data, social and economic data like income and education, and different biometric data.

A person's health data is usually located in different data repositories and formats, and much of this data might be in the form of unstructured free text [7]. In addition, health data are sensitive personal data [19], and their handling is subject to specific rules of HIPAA or GDPR. All these add complexity and challenges like semantic heterogeneity, data fragmentation, and data preprocessing (anonymization/pseudonymization, cleaning, formalization, etc.) that need to be addressed before using health data for the needs of ML and AI.

The *semantic heterogeneity* in health data refers to challenges arising from differences in the meaning or interpretation of terms and concepts used across various healthcare systems or organizations [2, 28]. Semantic heterogeneity occurs when a lack of common understanding or agreement exists on the semantics (meaning) of data elements. The sources for semantic heterogeneity are diverse data standards and terminologies, such as HL7 CDA, openEHR, IHE, FHIR, SNOMED CT, LOINC, and ICD, each with its own set of codes and meanings for clinical concepts. Integrating data from systems that use different standards can lead to semantic mismatches, making it challenging both for humans and machines to accurately interpret and exchange information [75]. In addition, healthcare providers or institutions may use local or customized terminologies that deviate from widely accepted standards. Another source for semantic heterogeneity is ambiguous or inconsistent clinical documentation, such as free-text notes that introduce uncertainty in the interpretation of medical terms and concepts. Automated systems may struggle to extract meaningful information from unstructured data, and human interpreters may derive different meanings from the same text [1]. Further, healthcare standards and terminologies evolve over time to accommodate new knowledge and medical practices, and older systems or data sources may use outdated standards, leading to disparities in the representation and interpretation of clinical concepts [17]. Healthcare involves various disciplines, each with its own vocabulary and understanding of terms.

Interdisciplinary collaboration in healthcare can also result in semantic differences, especially when terms are used across medical, nursing, and administrative domains. Also, inconsistent data entry practices, such as variations in abbreviations or coding conventions, can contribute to semantic heterogeneity challenges. Inaccuracies and discrepancies in data representation hinder the seamless exchange and integration of health information. The meaning of a term can vary based on the context in which it is used. Semantic variations due to contextual differences can lead to misunderstandings, particularly when data are shared across different healthcare settings. In addition, cultural and regional differences may influence the interpretation of certain health terms and concepts.

Health data that are shared globally may need to account for these variations to ensure accurate cross-cultural communication and understanding.

The other challenge with health data is *data fragmentation*. This is an even more complex and challenging issue than semantic heterogeneity [34,44]. Data fragmentation in healthcare poses a significant challenge as crucial health information is dispersed across institutions and data sources, hindering the creation of a comprehensive and cohesive patient health profile. The decentralization of healthcare data among different organizations, such as hospitals, clinics, specialized care facilities, smartphones, and wellbeing apps, leads to a fragmented view of an individual's health history. However, based on big data analysis, Bertl et al. have shown [9,12] that there can be a reliably strong correlation between illnesses that occur in the same person. Therefore, data fragmentation not only obstructs the seamless flow of information critical for providing timely and effective patient care but also jeopardizes the accuracy of diagnoses and treatment plans. In cases of the primary use of health data (diagnosis, treatment, prevention of diseases, longevity, etc.), the person is usually the one who knows and remembers and can, if necessary, restore missing data; in secondary use cases of health data (public health, medical science, policy-making, etc.), the existence of missing data may not be known at all, and the decisions in such a case may be made based on totally incomplete and even totally biased data. When analyzing data on the basis of two or more sets of data, if one and the same person is present in two or more sets of data, there is also no guarantee that said person's data will be correctly linked.

The third serious challenge is the current one-sided nature of health data, which is mainly inclined towards disease and treatment-related data in the second half of a person's life expectancy. While quite a lot of data on human health behavior and physical activity exist, linking these data to medical data is difficult, and so it is also difficult to combine these data in ML and AI applications.

Interoperability issues, incompatible data systems, and privacy concerns exacerbate the problem, impeding healthcare professionals from accessing a holistic understanding of a patient's medical journey. As the demand for integrated, patient-centric care grows, addressing data fragmentation emerges as a pressing imperative in the ongoing evolution of healthcare systems worldwide. Establishing standardized data exchange protocols and promoting interoperability is essential to overcoming these challenges and fostering a more connected, data-driven approach to healthcare delivery. In addition, health data fragmentation also affects primary use, as treating physicians may not know the data entered by a doctor from another hospital and are therefore forced, in many cases, to focus on episodic symptom treatment instead of focusing on continuity of care.

### 3.2 Regulatory and Legal Basis

The development of AI-supported medical devices is a complex process that meets a variety of regulatory and legal requirements. In the European Union, the Medical Device Regulation (MDR) forms the legal basis for the development of

these products. The MDR sets the requirements for the safety and performance of medical devices [57]. The MDR is supplemented by the General Data Protection Regulation (GDPR), which regulates the protection of personal data [24].

In order to obtain approval for a medical device in the EU, a manufacturer must be able to demonstrate that it is capable of producing safe and efficient products. This is preferably achieved by basing its development on harmonized standards that reflect the state of the art, which creates a presumption of conformity that facilitates and accelerates the certification process.

The manufacturer provides proof of this by implementing a quality management system (QMS), for example, in accordance with ISO 13485, and having it certified in conjunction with a notified body. This certification confirms that the manufacturer's QMS meets the requirements of the relevant standards and that the company is able to consistently deliver safe and effective medical devices. It is important to note that compliance with these standards and regulations is not only a legal requirement but also a matter of ethical responsibility. It ensures that patient safety is paramount and that the benefits of AI-enabled medical devices are maximized while potential risks are minimized [40].

As there are currently no harmonized standards that support the development of AI-supported medical devices, conformity cannot be assumed for the development of medical devices. Manufacturers are therefore faced with the challenge of comprehensively documenting the manufacturing process in order to demonstrate and prove the performance of their products [58]. This can be done through extended clinical studies, for example, which is time-consuming and costly. Nevertheless, 219 medical devices for the EU market can be found in the free database "AI for Radiology" [21]. For the US market, over 500 products can be found in the database of the Food and Drug Administration (FDA) [13], the authority responsible for the approval of medical devices.

**EU vs FDA.** The approval of medical devices under the FDA and the MDR has many similarities. Both require manufacturers to demonstrate the performance and safety of their products using a quality management system (QMS).

The differences lie mainly in the details. These include the requirements for the risk classification of the product, proof through clinical studies, and the requirements for technical documentation. These differences can increase the workload for manufacturers when preparing documents for different markets.

The use of internationally recognized processes and documents, such as those from the International Medical Device Regulators Forum, can help to reduce the effort involved. These international standards and guidelines can help to reduce the complexity of regulatory requirements while ensuring the safety and efficacy of products [41].

**FDA Activities for the Use of Continuous Learning Systems.** One thing all markets have in common at present is the use of frozen ML/AI models or the abandonment of continuously learning AI. The background to this is that the approval of a product or, in particular, Software as a Medical Device (SaMD) is

granted with a defined version. If the SaMD changes in the field, e.g. through active learning, the SaMD no longer complies with the subject of the approval and is therefore illegal on the market. An evaluation of the change is necessary to decide what the next steps are. In the case of minor changes, such as a bug fix, this would only have to be reported to the authorities with a modified version. If the change is more extensive or has an impact on the performance of the product, it requires prior validation and re-release of the product. And if the change has an impact on the intended use of the product, this must be re-approved [53, 79].

The FDA has recognized the challenges in regulating AI-based medical devices and published the discussion paper “Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)” in April 2019 [78]. In this paper, a number of activities were presented to promote the development and use of continuously learning AI in medicine. These include the creation of guidelines for the development and validation of AI algorithms.

In April of this year, a proposal for a Predetermined Change Control Plan (PCCP) was presented [80]. This plan is intended to enable manufacturers to identify and evaluate changes to the model. The PCCP defines three areas in which continuous learning should be tolerated differently. These areas include changes to the performance of the model that has a high chance of acceptance, changes to the input that has a medium chance of acceptance and which, in case of doubt, should be evaluated with the FDA in good time, and finally changes to the intended use that requires re-approval of the application and is unlikely to be approved.

**EU AI Act and Sync Between MDR and AI Act.** The recently introduced EU AI Act aims to create a legal framework for AI systems that minimizes risks while promoting innovation [27]. The Act is based to a certain extent on the regulations of the MDR [57], for example by providing for a risk-based approach. The Regulatory Framework defines four levels of risk in AI:

- Unacceptable risk
- High risk
- Limited risk
- Minimal or no risk

However, in the case of an acceptable risk, all AI-supported medical devices fall under the category of high-risk products. The AI Act also requires extensive logging of the algorithm’s activities. As part of a medical device manufacturer’s post-market surveillance activities, extensive monitoring measures are already required for the products on the market. However, the AI Act goes further and seeks to require the manufacturer to design the AI-powered application to keep detailed logs of what it evaluates and why it made a decision in order to monitor the performance of the model in the field [25]. This requires additional consent from the patient and the operator to monitor the entire processing of such a program.



Another area that has not yet been further defined is the requirement for sandbox operation, which can best be understood as an independent test laboratory [26]. This should have an insight into the development and verification of the AI model and also have test data available in order to be able to verify the results by themselves. The aim is for the independent authority to check the AI and certify it, similar to the way it is already necessary to carry out an EMC test in accordance with ISO 60601-1-2 in an independent test laboratory for the acceptance of electronic medical devices [39].

The AI Act is currently in trilogue negotiations and is expected to come into force in 2024. The question of how existing regulations, such as the MDR and the AI Act, can be brought together is still open. In order to not inhibit or even suppress innovation, a sensible harmonization of regulations is necessary. The transition to the MDR is already a complex task for all stakeholders. The notified bodies need to be certified for the MDR, which leads to less accredited notified bodies. The Manufacturers need to re-certify their products under MDR or may fall, in some cases, into a higher-risk class, which also needs to be certified [42]. So, the few named bodies are occupied by re-certification and there are manufacturers who need a new body to be notified for their medical devices. The transition of the MDR led to some withdrawal of medical devices in the area of implants, for example, [30]. The synchronization between the EU AI Act and the MDR is crucial to ensure that the regulations for AI-supported medical devices take into account both technological progress and patient protection. The upcoming challenge for legislators and regulators, as well as for organizations and scientists, would be to take appropriate action to ensure a smooth transition while protecting patients and supporting the innovation and development of safe and effective products.

## 4 Future Opportunities for AI in Healthcare

### 4.1 Medical Data Storage

One of the promising solutions to tackle all the issues related to health data mentioned in Sect. 3.1 is a total paradigm shift from the current institution-based health data management [54] towards the individuals' privately owned and fully controlled health data. In [45–47], Klementi et al., have explained the idea, as well as proposed prospective research topics and a reference architecture towards such a system. They describe an architectural framework as a solution for the pervasive issue of health data fragmentation, which has posed a considerable impediment to the secondary use of health data. This approach facilitates the consolidation of an individual's comprehensive health data within a unified logical repository exclusively controlled by the data owner. Remarkably, this consolidation occurs without incurring an augmented risk to the privacy of such data. Notably, the proposed solution extends beyond primary health data, encompassing other pertinent data categories commonly referred to as “lifestyle data,” including information from wearable devices, home measurement results, and similar sources. This inclusive approach fosters the development of a holistic



perspective on an individual's health status, empowering each person to actively monitor their own well-being. Consequently, it has the potential to engender a paradigm shift in the healthcare domain, transitioning the overarching focus from reactive disease management to proactive prevention, with individuals assuming an increasingly prominent role and the associated responsibility. Given the innovative nature of this solution, it necessitates paradigmatic changes in societal conceptions surrounding personal data storage. The authors posit that technological readiness for such a transformation may outpace societal preparedness. Accordingly, prompt initiation of broad-based discussions concerning this approach assumes paramount importance.

## 4.2 AI Technology and Algorithms

In terms of AI technology and algorithms, we see *humanoid robotics* and *quantum computing* as enablers for *personalized and preventive medicine* as big emerging clusters.

**Convergence of Synthetic Intelligence and Cognitive Agents in Humanoid Robotics.** The fusion of Synthetic Intelligence (SI) and Cognitive Agents (CAs) represents a significant advancement in the field of robotics. Humanoid robots equipped with SI and CA capabilities are not only capable of mimicking human-like behaviors but also adapting and learning from their environment [43]. This convergence holds the promise of addressing critical issues in healthcare. We delve into the role of SI and CAs in humanoid robotics and their positive impacts on healthcare efforts by analyzing the intricate interplay between SI, CAs, and humanoid robots and elucidating the dynamic ecosystem where these intelligent agents collaborate with systems, remote devices such as IoT, equipment, and humans. This highlights the imperative of fostering responsible AI development and promoting the ethical deployment of Synthetic Intelligence, Cognitive Agents, and humanoid robotics to ensure the harmonious and productive coexistence between humans and intelligent machines as the technological amalgamation is moving forward across many domains.

Synthetic Intelligence (SI) refers to the artificial creation of intelligent entities through the emulation of human cognitive processes [38]. SI systems are designed to learn, adapt, and make decisions, mimicking human intelligence to varying degrees. This technology forms the foundation of intelligent robots. Cognitive Agents are software or hardware entities that possess the ability to perceive, reason, and make decisions in complex environments [63]. They are equipped with advanced reasoning, problem-solving, and learning capabilities, making them crucial components in the development of intelligent robots.

Simulate Human Critical Thinking and Reasoning and Humanoid Cognitive Robotics is the implementation of a Decision by these CAs. Cognitive Agents reference entities that have the ability to perceive their environment, process information, and make decisions or take actions based on their understanding. These agents often leverage cognitive abilities like human thinking processes. Critical

Thinking and Reasoning are cognitive skills associated with the ability to analyze information, evaluate arguments, and make decisions based on evidence and sound logic. In the context of cognitive agents, incorporating critical thinking and reasoning involves endowing these artificial entities with the capability to assess information, weigh evidence, and make decisions in a manner analogous to human cognitive processes. Critical Thinking Framework (CTF-HA) decision-making, problem-solving, and information evaluation by promoting systematic, rational, and evidence-based thinking. It serves as a guide for structuring cognitive processes such as Behavioral Characteristics of Cognitive Agents that Access information specific to the subject matter being evaluated, analyze the logical consistency and validity of arguments, check facts, data sources, and evidence for accuracy and relevance, identify gaps, limitations, or biases in information/data, generate alternate hypotheses and conclusions based on evidence, update assessments as new evidence/data becomes available, and highlight uncertainties and areas needing further analysis.

**Quantum Machine Learning.** Quantum computing and quantum machine learning represent transformative frontiers in the advancement of AI in medicine [31]. The integration of quantum computing's immense processing power with the sophisticated algorithms of quantum machine learning can potentially revolutionize how we approach complex biological systems and healthcare challenges. Quantum computers, with their ability to perform specific calculations at unprecedented speeds, offer a promising avenue for analyzing and detecting anomalies in vast datasets that are characteristic of the medical field. This capability could lead to the identification of novel therapeutic targets and the development of personalized medicine strategies at a pace far beyond current computational methods. Furthermore, quantum machine learning algorithms are particularly adept at identifying patterns and correlations within these large datasets, which could enhance diagnostic accuracy and predict patient outcomes with greater precision. The synergy of these technologies could also facilitate the design of new drugs by simulating molecular interactions at a quantum level, thus reducing the reliance on costly and time-consuming laboratory experiments [14]. Moreover, the application of quantum machine learning in genomics could accelerate the understanding of genetic factors in disease, leading to breakthroughs in preventive medicine [77]. We see the future of AI in medicine not just as iterative improvements, but quantum leaps in our capabilities to heal and prevent illness.

**Personalized and Preventive Medicine.** Personalized and preventive medicine represents a transformative paradigm in healthcare, leveraging medical and health data to tailor interventions to individual characteristics, genetic makeup, and lifestyle choices [15, 56]. Central to this approach is the integration of data from various sources, including smartphones and wellbeing applications. These technologies provide a continuous stream of real-time data on an individual's daily activities, exercise routines, sleep patterns, and other health-related metrics. By incorporating this information into medical assessments,

healthcare practitioners can gain a more nuanced understanding of a patient’s overall well-being, enabling the development of personalized preventive strategies. Analyzing data from smart devices allows for early detection of potential health issues, facilitating proactive interventions to mitigate risks and promote long-term health [73]. As personalized and preventive medicine advances, the seamless integration of diverse health data sources, including those derived from mobile technologies, plays a pivotal role in optimizing healthcare delivery and empowering individuals to actively participate in their health management.

### 4.3 Explainable AI and Trustworthiness of AI Systems

For several reasons, a human in the loop is crucial for successfully implementing AI-based decision support systems and automation in healthcare. Firstly, it ensures that the complexity and nuances of medical decision-making, which often require human judgment and experience, are not overlooked. Secondly, it allows for real-time oversight and the ability to intervene if the AI system makes a decision that could potentially harm a patient. This is also of vital importance from a legal perspective. Lastly, it helps build trust among healthcare professionals and patients, as they can be assured that there is human oversight, thus fostering acceptance of AI technology in healthcare.

Currently, we see novel AI technologies gaining momentum, particularly in the field of deep learning, which includes transformers. With such novel AI technologies emerging, we recently witnessed an increasing debate about the explainability and trustworthiness of AI systems. Basically, the utilization of AI systems is considered risky; the trustworthiness of AI systems is the goal, and their explainability is seen as a pre-condition for their trustworthiness; compare with [59]: “The AI community is pursuing explainability as one of many desirable characteristics for trustworthy AI systems.” [59] Unfortunately, both explainability and trustworthiness are social constructs rather than naturally given facts and measurable features of an AI system, and, as such, they are hard to grasp and characterize.

With [59], the NIST (National Institute of Standards and Technology) attempts to provide a characterization of explainable AI through the identification of four principles as follows:

- “*Explanation*: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.” [59]
- “*Meaningful*: A system provides explanations that are understandable to the intended consumer(s).” [59]
- “*Explanation Accuracy*: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system’s process.” [59]
- “*Knowledge Limits*: A system only operates under the conditions for which it was designed and when it reaches sufficient confidence in its output.” [59]

Different sectors and business domains vary greatly in terms of their terminology, perception of phenomena, and social constructs, as is the case with

regard to the explainability and trustworthiness of AI systems. Therefore, we argue that it needs domain-specific discussions of explainable AI and its trustworthiness, and consequently also a domain-specific discussion of explainable AI and its trustworthiness in regard to the healthcare sector.

The current discussion of explainable AI puts particular emphasis on the understandability of AI mechanisms, in particular on the explanation of AI systems that are understandable by the customer, i.e., the end-consumer, who is usually a layman in regard to AI technology. We suggest that such focus should be questioned and discussed. For the healthcare domain, consider any other diagnostic tool, such as a laboratory test for a disease. It is unlikely that the layman would understand the biochemistry of the laboratory test, independent of how good the explanation of the test would be. Even the expert, including the medical practitioner, might struggle to completely understand the provided explanation. The problem is that explainability and understandability are not measurable. The level of understanding is measurable, i.e., by an exam, however, at a closer look, such a measure of the level of understanding is also purely a social construct. The very concept of understandability remains un-measurable; all understanding is, eventually, an illusion.<sup>1</sup> As explainability and trustworthiness are social constructs, they can be made subject to the *social science* research technique [4, 52, 84] with its quantitative and qualitative analyses; however, it is impossible to investigate them *scientifically*, i.e., as subject to the *exact sciences* technique.<sup>2</sup>

Whereas the explainability and trustworthiness of a diagnostic tool are not measurable (in the sense of *exact sciences*), the predictive power of a diagnostic tool is indeed measurable. The confusion matrix of a diagnostic tool is fully specified by its positive predictive value (PPV), its negative predictive value (NPV), and the prevalence of the disease in the investigated population. Assuming that both the group of affected and the group of non-affected individuals in the investigated population are sufficiently statistically significant, the pair of PPV and NPV provides the medical practitioner with a relatively stable measure for their decision-making. The same applies to any diagnostic tool, including those based on AI algorithms.

Furthermore, the current discussion of explainability and trustworthiness in AI seems to be focused rather on AI algorithms than on the quality of the underlying data. In extreme cases, the underlying data can be subject to all kinds of statistical paradoxes [67–70], and even if not, the existence of confounding effects (no matter in terms of latent confounders as well as known confounders) can have severe impact on over-estimating or under-estimating the predictive

---

<sup>1</sup> For example, even if we have perfectly *learned* to apply the field equations of general relativity as the current established theory of gravity, we need to admit that we still do not *understand* what gravity is.

<sup>2</sup> Whereas social sciences, in particular, whenever they follow the stance of *positivist research* as opposed to *interpretivism* [52], might often mimic the *exact science* technique, as an attitude, they would still remain not being *exact sciences*; compare with [29].

power of a diagnostic tool. A particular challenge is identifying confounding effects in sparse data such as that utilized by novel AI techniques. Therefore, we see the investigation of confounding effects in sparse data as an important field of research for the applicability of novel AI systems in healthcare.

#### 4.4 Potential Clinical and Healthcare Management Approaches

Similar to other service areas, the success of information technology implementation in healthcare depends on how it is possible to involve content people and those who know the business side, i.e., clinicians and healthcare managers. The relatively few implementations of AI-based Digital Decision Support Systems (DDSS) in clinical practice over the past decade indicate that collaboration between AI specialists and clinicians has not yet gained momentum.

When implementing AI in healthcare, it is reasonable to look at how the introduction of new technologies has taken place in the past. What has proven successful is not technological discoveries per se but how they have been used to assess or improve human health. For example, X-rays or fiber optics do not have a separate value in medicine but are seamlessly integrated into diagnostic activity or surgery. In the case of AI applications in healthcare, the focus is still on the development of individual algorithms. However, integration of multiple algorithms into a clinical workflow or health promotion is required. Similarly to digital radiography, where X-ray photons act directly on a photoconductor layer collectively, producing positive and negative charges and thus forming images of respective anatomic areas, the application of AI algorithms should be more comprehensive and bring new findings to human health status and support new and more effective approaches in the clinical workflow. It is not only one data source and one deep learning algorithm but an interdisciplinary approach combined with medical, IT, and healthcare process knowledge integrated into current systems.

The material for AI-powered medical applications and tools is data. Since a large part of data in medicine is entered by clinicians and carries potential bias, especially when written in free text, the development of AI algorithms should focus on non-obtrusive data collected by various sensors, genetic data, and well-structured and terminologically consistent data. These sources provide an opportunity primarily for preventive health care, where the biosignals collected by the person from everyday life with portable sensors, enriched with genetic data, make it possible to start preventive and health promotion activities well in advance of the development of disease symptoms. It is also possible to find indicator diagnoses that point to possible hidden new health problems based on electronic health records and personal health records data.

AI solutions can certainly support healthcare financing and health management more broadly. As is known, the determinants of human health are, in addition to medical care and genetic factors, social and economic conditions, education, behavior, and environment. The mutual integration and analysis of these data with AI-supported tools, together with person-reported data and biosignals, has the potential to change healthcare from reactive to proactive.

Also, new funds can be brought into healthcare financing if an additional bonus is paid to the funder of prevention activities in the case of the successful completion of a person's personal health plan. This is how, for example, health impact bonds work – a funding model that supports large-scale prevention.

## 5 Discussion

The integration of AI in healthcare is a topic of significant interest and debate. While the potential benefits are substantial, several critical discussions and considerations arise:

- **Ethical Implications:** As AI algorithms become more involved in clinical decision-making, ethical questions emerge. How do we ensure transparency, fairness, and accountability in AI-driven diagnoses and treatment plans? Striking the right balance between automation and human judgment is crucial.
- **Data Quality and Bias:** AI models heavily rely on data. However, biases present in historical medical data can perpetuate inequalities. Ensuring diverse and representative datasets is essential to prevent biased outcomes. Additionally, addressing data privacy concerns is paramount.
- **Harnessing new data sources:** personalized and preventive medicine based on genome or unobtrusive data for digital phenotype still holds unexplored potential for improving health care.
- **Human-AI Collaboration:** Rather than replacing clinicians, AI should augment their capabilities. Discussions about how AI can complement human expertise, enhance diagnostic accuracy, and improve patient outcomes are ongoing.
- **Regulatory Frameworks:** Developing robust regulations for AI in healthcare is challenging. Balancing innovation with safety requires collaboration among policy-makers, researchers, and industry stakeholders.
- **Interoperability and Integration:** Integrating AI seamlessly into existing healthcare systems remains a hurdle. Ensuring interoperability, data sharing, and compatibility with electronic health records are critical aspects.
- **Patient Acceptance and Trust:** Patients need to trust AI-driven decisions. It is essential to educate patients about AI's role, its limitations, and the shared decision-making process.
- **Resource Allocation:** While AI promises efficiency gains, initial implementation costs and ongoing maintenance must be considered. Allocating resources effectively is vital.
- **Legal and Liability Issues:** Who bears responsibility when an AI algorithm makes an incorrect diagnosis? Legal frameworks need to adapt to this evolving landscape.
- **Global Collaboration:** AI transcends borders. Collaborative efforts across countries can accelerate research, data sharing, and best practices.
- **Long-Term Impact:** Finally, discussions should extend beyond immediate benefits. How will AI shape healthcare in the long term? What societal changes will it bring?

With increasing computing power, the potential of quantum computing, and the corresponding impact of areas like personalized and preventive medicine and robotics, the above-mentioned points have become increasingly important. While AI holds immense potential, thoughtful analysis and interdisciplinary collaboration are essential to harness its benefits while mitigating risks.

## 6 Conclusion

In conclusion, the integration of AI into the healthcare industry holds immense promise. As evidenced by recent studies, AI has already made significant contributions to cancer management, urinary bladder control, stroke prediction, cardiovascular disease diagnosis, and personalized medicine. Its potential extends to predicting and preventing diseases, enhancing clinical trials, and improving patient outcomes. AI applications in healthcare can be broadly categorized into three domains:

- Patient-Oriented AI: This category encompasses a wide range of tasks, including medical record review, population health analytics, therapeutic drug and device design, radiology image interpretation, clinical diagnosis, and even patient communication.
- Clinician-oriented AI: Clinicians benefit from AI in tasks such as medical record management, answering phone calls, and aiding in diagnosis and treatment planning.
- Administrative and Operational-oriented AI: Streamlining processes like medical record review and population health analysis fall under this domain.

The advantages of AI adoption in healthcare are manifold: increased productivity, reduced staff burden, and enhanced patient care. However, challenges persist, such as limited access to high-quality, unbiased data sets, internet connectivity issues, and the need for robust systems and protocols.

As technology advances, addressing these challenges and responsibly leveraging AI's potential will be crucial for realizing its transformative impact on healthcare delivery and patient well-being.

**Acknowledgement.** This work in the project ‘ICT programme’ was supported by the European Union through the European Social Fund and the Norway Grants Program “Green ICT” (Nmb. F21009).

## References

1. Adnan, K., Akbar, R.: An analytical study of information extraction from unstructured and multidimensional big data. *J. Big Data* **6**(1), 1–38 (2019)
2. Amar, F., April, A., Abran, A.: Electronic health record and semantic issues using fast healthcare interoperability resources: systematic mapping review. *J. Med. Internet Res.* **26**, e45209 (2024)

3. Aung, Y.Y., Wong, D.C., Ting, D.S.: The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br. Med. Bull.* **139**(1), 4–15 (2021)
4. Babbie, E.R.: *The Practice of Social Research*, 13th Edition. Wadsworth Publishing (2012)
5. Belciug, S.: An introduction to artificial intelligence in healthcare. In: Lim, C.P., Vaidya, A., Chen, Y.W., Jain, T., Jain, L.C. (eds.) *Artificial Intelligence and Machine Learning for Healthcare: Vol. 1: Image and Data Analytics*, pp. 1–16. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-11154-9\\_1](https://doi.org/10.1007/978-3-031-11154-9_1)
6. Bertl, M.: Systematic AI support for psychiatry: a framework on how to implement decision support systems. Ph. D. thesis, Tallinn University of Technology, Tallinn, Estonia (2023). <https://digikogu.taltech.ee/en/Item/a77022c7-b2af-46f9-ae83-62fe04f9bc72>
7. Bertl, M., Kankainen, K.J.I., Piho, G., Draheim, D., Ross, P.: Evaluation of data quality in the Estonia national health information system for digital decision support. In: *Proceedings of the 3rd International Health Data Workshop*. CEUR-WS (2023)
8. Bertl, M., Metsallik, J., Ross, P.: A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. *Front. Psychiatry* **13**, 923613 (2022). <https://doi.org/10.3389/fpsyt.2022.923613>
9. Bertl, M., Ross, P., Draheim, D.: Predicting psychiatric diseases using AutoAI: a performance analysis based on health insurance billing data. In: Strauss, C., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) *DEXA 2021*. LNCS, vol. 12923, pp. 104–111. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-86472-9\\_9](https://doi.org/10.1007/978-3-030-86472-9_9)
10. Bertl, M., Ross, P., Draheim, D.: A survey on AI and decision support systems in psychiatry - uncovering a dilemma. *Expert Syst. Appl.* **202**, 117464 (2022). <https://doi.org/10.1016/j.eswa.2022.117464>
11. Bertl, M., Ross, P., Draheim, D.: Systematic AI support for decision-making in the healthcare sector: obstacles and success factors. *Health Policy Technol.* **12**(3), 8 (2023). <https://doi.org/10.13140/RG.2.2.17159.52646/1>
12. Bertl, M., Shahin, M., Ross, P., Draheim, D.: Finding indicator diseases of psychiatric disorders in bigdata using clustered association rule mining. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pp. 826–833 (2023)
13. Center for Devices and Radiological Health: Artificial intelligence and machine learning (AI/ML)-enabled medical devices. FDA (2023). <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>. Accessed 20 Oct 2023
14. Chen, M., et al.: The emergence of machine learning force fields in drug design. *Med. Res. Rev.* **44**, 1147–1182 (2024)
15. Chevinsky, J., Chirumamilla, S., Caswell, S., Nyoni, L.M., Studer, K.: Clinical preventive medicine, integrative medicine, and lifestyle medicine: current state and future opportunities in the development of emerging clinical areas. *AJPM Focus* **3**(1), 100166 (2024)
16. Chung, C.T., et al.: Clinical significance, challenges and limitations in using artificial intelligence for electrocardiography-based diagnosis. *Int. J. Arrhythmia* **23**(1), 24 (2022)
17. Chute, C.G., Cohn, S.P., Campbell, J.R.: A framework for comprehensive health terminology systems in the united states: development guidelines, criteria for selection, and public policy implications. *J. Am. Med. Inform. Assoc.* **5**(6), 503–510 (1998)



18. Clark, K., Smith, R.: Unleashing the power of design thinking. *Des. Manag. Rev.* **19**, 8–15 (2008)
19. Clayton, E.W., Embí, P.J., Malin, B.A.: Dobbs and the future of health data privacy for patients and healthcare organizations. *J. Am. Med. Inform. Assoc.* **30**(1), 155–160 (2023)
20. Deloitte: The future of artificial intelligence in health care. emerging applications of AI in health care (2019). <https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/future-of-artificial-intelligence-in-health-care.html>. Accessed 03 Dec 2023
21. Diagnostic Image Analysis Group: AI for Radiology. <https://www.AIforRadiology.com>. Accessed 23 Nov 2023
22. Directorate-General for Health and Food Safety: A European health data space: harnessing the power of health data for people, patients and innovation (2022). <https://health.ec.europa.eu/document/download/17c7065c-c432-445f-9b27-8ccf283581bc.en?filename=com.2022-196.en.pdf>
23. Esnault, C., Rollot, M., Guilmin, P., Zucker, J.D.: Qluster: an easy-to-implement generic workflow for robust clustering of health data. *Front. Artif. Intell.* **5**, 1055294 (2023)
24. EU: Regulation (eu) 2016/679 of the European parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance). *Official J. Eur. Union* **119**(1), 1–88 (2016). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
25. European Commission: Article 12. <https://artificialintelligenceact.com/title-iii/chapter-2/article-12/>. Accessed 07 Dec 2023
26. European Commission: Article 53. <https://artificialintelligenceact.com/title-v/article-53/>. Accessed 07 Dec 2023
27. European Commission: EU AI act: institutions debate definition of AI. <https://www.morganlewis.com/pubs/2023/09/eu-ai-act-institutions-debate-definition-of-ai>
28. Fennelly, O., Moroney, D., Doyle, M., Eustace-Cook, J., Hughes, M.: Key interoperability factors for patient portals and electronic health records: a scoping review. *Int. J. Med. Informatics* **183**, 105335 (2024)
29. Feynman, R.P.: Cargo cult science – some remarks on science, pseudoscience, and learning how to not fool yourself. Caltech’s 1974 commencement address. <https://calteches.library.caltech.edu/51/2/CargoCult.htm>
30. Fick, M.: Insight: medical device makers drop products as EU law sows chaos. *Reuters* (2022). <https://www.reuters.com/business/healthcare-pharmaceuticals/medical-device-makers-drop-products-eu-law-sows-chaos-2022-12-19/>. Accessed 07 Dec 2023
31. Fujii, Y.R.: The MicroRNA 2000 Transformer: Quantum Computing and Artificial Intelligence for Health. *Springer Nature* (2023). <https://doi.org/10.1007/978-981-99-3165-1>
32. Garg, A., Venkataramani, V.V., Karthikeyan, A., Priyakumar, U.D.: Modern AI/ML methods for healthcare: opportunities and challenges. In: Bapi, R., Kulkarni, S., Mohalik, S., Peri, S. (eds.) *ICDCIT 2022*. LNCS, vol. 13145, pp. 3–25. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-94876-4\\_1](https://doi.org/10.1007/978-3-030-94876-4_1)
33. Gerke, S., Minssen, T., Cohen, G.: Ethical and legal challenges of artificial intelligence-driven healthcare. In: *Artificial intelligence in healthcare*, pp. 295–336. Elsevier (2020)

34. Getzen, E., Ungar, L., Mowery, D., Jiang, X., Long, Q.: Mining for equitable health: assessing the impact of missing data in electronic health records. *J. Biomed. Inform.* **139**, 104269 (2023)
35. Gupta, R., et al.: VAHAK: a blockchain-based outdoor delivery scheme using UAV for Healthcare 4.0 services. In: *IEEE INFOCOM 2020 – IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 255–260. IEEE (2020). <https://doi.org/10.1109/INFOCOMWKSHPS50562.2020.9162738>
36. Gupta, R., Shukla, A., Tanwar, S.: Aayush: a smart contract-based telesurgery system for healthcare 4.0. In: *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6. IEEE (2020). <https://doi.org/10.1109/ICCWorkshops49005.2020.9145044>
37. Hackl, W.O., Ammenwerth, E.: Spirit: systematic planning of intelligent reuse of integrated clinical routine data: a conceptual best-practice framework and procedure model. *Methods Inf. Med.* **55**(2), 114–124 (2016). <https://doi.org/10.3414/ME15-01-0045>
38. Holl, C.: The content intelligence: an argument against the lethality of artificial intelligence. *Discov. Artif. Intell.* **4**(1), 13 (2024)
39. International Electrotechnical Commission (IEC): IEC 60601-1-2:2014 - IEC-Normen - VDE VERLAG. <https://www.vde-verlag.de/iec-normen/220613/iec-60601-1-2-2014.html>. Accessed 07 Dec 2023
40. International Electrotechnical Commission (IEC): IEC62304:2006/AMD1:2015 amendment 1-medical device software-software life cycle processes (2015). <https://webstore.iec.ch/publication/22790>
41. International Medical Device Regulators Forum: International medical device regulators forum (IMDRF) (2023). <https://www.imdrf.org/international-medical-device-regulators-forum-imdrf>. Accessed 07 Dec 2023
42. Juerschik, S.: Survey on the EU medical devices regulation (2022). <https://www.dihk.de/resource/blob/82008/e5950cc020f29581c01d6b5b899855da/eu-medizinprodukteverordnung-englisch-data.pdf>. Accessed 07 Dec 2023
43. Kadalagere Sampath, S., Wang, N., Wu, H., Yang, C.: Review on human-like robot manipulation using dexterous hands. *Cogn. Comput. Syst.* **5**(1), 14–29 (2023)
44. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 1–9 (2019)
45. Klementi, T., Kankainen, K.J.I., Piho, G., Ross, P.: Prospective research topics towards preserving electronic health records in decentralised content-addressable storage networks. In: *HEDA@ Petri Nets* (2022)
46. Klementi, T., Piho, G.: Method and system for managing data using decentralized content-addressable storage networks. European patent Office, priority number EP24166173.5 (2024). Submitted patent
47. Klementi, T., Piho, G., Ross, P.: A reference architecture for resolving health data accessibility, comprehensiveness, and ownership dilemmas through preserving semantically interoperable personal health records in decentralized content-addressable storage networks (2024). Will be published soon
48. Kokshagina, O.: Managing shifts to value-based healthcare and value digitalization as a multi-level dynamic capability development process. *Technol. Forecast. Soc. Chang.* **172**, 121072 (2021)
49. Li, H., et al.: Review on security of federated learning and its application in healthcare. *Futur. Gener. Comput. Syst.* **144**, 271–290 (2023)

50. Mashoufi, M., Ayatollahi, H., Khorasani-Zavareh, D., Boni, T.T.A.: Data quality in health care: main concepts and assessment methodologies. *Methods Inf. Med.* **62**, 005–018 (2022). <https://doi.org/10.1055/s-0043-1761500>
51. Mbunge, E., Muchemwa, B., Jiyane, S., Batani, J.: Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. *Global Health J.* **5**(4), 169–177 (2021). <https://doi.org/10.1016/j.glohj.2021.11.008>
52. McBride, K., Misnikov, Y., Draheim, D.: Discussing the foundations for interpretivist digital government research. In: Charalabidis, Y., Flak, L.S., Viale Pereira, G. (eds.) *Scientific Foundations of Digital Governance and Transformation*. PAIT, vol. 38, pp. 121–147. Springer, Cham (2022). [https://doi.org/10.1007/978-3-030-92945-9\\_6](https://doi.org/10.1007/978-3-030-92945-9_6)
53. Medical Device Coordination Group (MDCG): Guidance on significant changes regarding the transitional provision under Article 120 of the MDR with regard to devices covered by certificates according to MDD or AIMDD (2020). [https://health.ec.europa.eu/system/files/2020-09/md\\_mdmg-guidance\\_significant\\_changes\\_annexes\\_en\\_0.pdf](https://health.ec.europa.eu/system/files/2020-09/md_mdmg-guidance_significant_changes_annexes_en_0.pdf). Accessed 07 Dec 2023
54. Metsallik, J., Ross, P., Draheim, D., Piho, G.: Ten years of the e-health system in Estonia. In: *Proceedings of MMHS 2018 – Workshop on (Meta) Modelling for Healthcare Systems*. Lecture Notes in Computer Science, vol. 2336. CEUR Workshops Proceedings (2019)
55. Mullachery, B., Alismail, S.: A smart healthcare framework: opportunities for integrating emerging technologies (5G, IoT, AI, and GIS). In: Arai, K. (ed.) *Proceedings of the FTC’2022 - the 1st Future Technologies Conference, Volume 3*. Lecture Notes in Networks and Systems, vol. 561, pp. 325–340. Springer International Publishing, Cham (2023). [https://doi.org/10.1007/978-3-031-18344-7\\_22](https://doi.org/10.1007/978-3-031-18344-7_22)
56. Okolo, C.A., Olorunsogo, T., Babawarun, O., et al.: A comprehensive review of AI applications in personalized medicine. *Int. J. Sci. Res. Arch.* **11**(1), 2544–2549 (2024)
57. Parliament, E., of the Council: Regulation (EU) 2017/745 of the European parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/ec, regulation (ec) no 178/2002 and regulation (ec) no 1223/2009 and repealing council directives 90/385/eec and 93/42/eec (text with eea relevance) (2017). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02017R0745-20230320>
58. Petersen, E., et al.: Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE Access* **10**, 58375–58418 (2022). <https://doi.org/10.1109/ACCESS.2022.3178382>. conference Name: IEEE Access
59. Phillips, P.J., et al.: Four principles of explainable artificial intelligence. Tech. Rep. NISTIR 8312, National Institute of Standards and Technology (NIST), US Department of Commerce (2021). <https://doi.org/10.6028/NIST.IR.8312>
60. PWC: Transforming healthcare through secondary use of health data. PriceWaterhouseCoopers (2009)
61. Rawat, B., Joshi, Y., Kumar, A.: AI in healthcare: opportunities and challenges for personalized medicine and disease diagnosis. In: *Proceedings of ICIRCA’2023 – the 5th International Conference on Inventive Research in Computing Applications*, pp. 374–379. IEEE (2023). <https://doi.org/10.1109/ICIRCA57980.2023.10220746>
62. Ross, P.: Healthcare information systems (2023). Unpublished lecture notes at École Nationale Supérieure des Mines de Saint-Étienne

63. Sandini, G., Sciutti, A., Morasso, P.: Artificial cognition vs. artificial intelligence for next-generation autonomous robotic agents. *Front. Comput. Neurosci.* **18**, 1349408 (2024)
64. Saraswat, D., et al.: Explainable AI for healthcare 50: opportunities and challenges. *IEEE Access* **10**, 84486–84517 (2022). <https://doi.org/10.1109/ACCESS.2022.3197671>
65. Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., Biancone, P.: The role of artificial intelligence in healthcare: a structured literature review. *BMC Med. Inform. Decis. Mak.* **21**, 1–23 (2021)
66. Sharma, A.: Artificial intelligence in health care. *Int. J. Humanit. Arts, Med. Sci.* **5**(1), 106–109 (2021)
67. Sharma, R.: Uni?cation of decision support techniques: mitigating statistical paradoxes for enabling trustworthy decision making. Ph. D. thesis, Tallinn University of Technology (2023). <https://digikogu.taltech.ee/et/Item/ae7a764a-f548-4750-8ba8-5efa79038799>
68. Sharma, R., Garayev, H., Kaushik, M., Tiwari, P., Peious, S.A., Draheim, D.: Detecting simpson’s paradox: a machine learning perspective. In: Strauss, C., Cuzocrea, A., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) *Proceedings of DEXA’2022 - the 33rd International Conference on Database and Expert Systems Applications*. Lecture Notes in Computer Science, vol. 13426. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-12423-5\\_25](https://doi.org/10.1007/978-3-031-12423-5_25)
69. Sharma, R., Kaushik, M., Peious, S.A., Shahin, M., Vidyarthi, A., Draheim, D.: Existence of the yule-simpson effect: an experiment with continuous data. In: *Proceedings of Confluence’2022 - the 12th International Conference on Cloud Computing, Data Science and Engineering*, pp. 351–355. IEEE (2022)
70. Sharma, R., Kaushik, M., Peious, S.A., Shahin, M., Vidyarthi, A., Draheim, D.: Why not to trust big data: discovering instances of Simpson’s paradox. In: Rage, U.K., Goyal, V., Reddy, P.K. (eds.) *Proceedings of PMDB’2022@DASFAA’2022 - the 1st Workshop on Pattern mining and Machine learning in Big Complex Databases*, conducted at the 27th International Conference on Database Systems for Advanced Applications (DASFAA-2022). Lecture Notes in Computer Science, vol. 13245. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-11217-1\\_4](https://doi.org/10.1007/978-3-031-11217-1_4)
71. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **22**(5), 1589–1604 (2018). <https://doi.org/10.1109/JBHI.2017.2767063>
72. Strickland, E.: IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* **56**(4), 24–31 (2019)
73. Tariq, M.U.: Advanced wearable medical devices and their role in transformative remote health monitoring. In: *Transformative Approaches to Patient Literacy and Healthcare Innovation*, pp. 308–326. IGI Global (2024)
74. Thakare, V., Khire, G., Kumbhar, M.: Artificial intelligence (AI) and internet of things (IoT) in healthcare: opportunities and challenges. *ECS Trans.* **107**(1), 7941 (2022). <https://doi.org/10.1149/10701.7941ecst>
75. Torab-Miandoab, A., Samad-Soltani, T., Jodati, A., Rezaei-Hachesu, P.: Interoperability of heterogeneous health information systems: a systematic literature review. *BMC Med. Inform. Decis. Mak.* **23**(1), 18 (2023)
76. Tyrväinen, P., Silvennoinen, M., Talvitie-Lamberg, K., Ala-Kitula, A., Kuoremäki, R.: Identifying opportunities for AI applications in healthcare – renewing the national healthcare and social services. In: *Proceedings of SeGAH’2018 - the 6th*

- IEEE International Conference on Serious Games and Applications for Health, pp. 1–7. IEEE (2018). <https://doi.org/10.1109/SeGAH.2018.8401381>
77. Ullah, U., Garcia-Zapirain, B.: Quantum machine learning revolution in healthcare: a systematic review of emerging perspectives and applications. *IEEE Access* **12**, 11423–11450 (2024)
  78. U.S. Food and Drug Administration (FDA): Proposed regulatory framework for modifications to AI/ML software as a medical device (2019). <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
  79. U.S. Food and Drug Administration (FDA): Deciding when to submit a 510(k) for a change to an existing device (2020). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-change-existing-device>. Accessed 07 Dec 2023
  80. U.S. Food and Drug Administration (FDA): Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions (2023). <https://www.fda.gov/media/166704/download>
  81. Wade, T.D.: Refining gold from existing data. *Curr. Opin. Allergy Clin. Immunol.* **14**(3), 181 (2014)
  82. Williamson, S.M., Prybutok, V.: Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Appl. Sci.* **14**(2), 675 (2024)
  83. Xiang, D., Cai, W.: Privacy protection and secondary use of health data: strategies and methods. *Biomed. Res. Int.* **2021**, 6967166 (2021)
  84. Yin, R.K.: *Case Study Research and Applications: Design and Methods*, 6th Edition. Sage (2017)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# CRISP-PCCP – A Development Methodology Supporting FDA Approval for Machine Learning Enabled Medical Devices

Ludwig Pechmann<sup>1</sup>, Yannik Potdevin<sup>2</sup>, Kai Brehmer<sup>3</sup>, Dirk Nowotka<sup>2</sup>,  
and Martin Leucker<sup>1,4</sup>

<sup>1</sup> UniTransferKlinik Lübeck GmbH, Lübeck, Germany

leucker@isp.uni-luebeck.de

<sup>2</sup> Dependable Systems Group, Kiel University, Kiel, Germany

<sup>3</sup> Institute for Electrical Engineering in Medicine, University of Lübeck,  
Lübeck, Germany

<sup>4</sup> Institute for Software Engineering and Programming Languages,  
University of Lübeck, Lübeck, Germany

**Abstract.** The U.S. Food and Drug Administration (FDA) is the regulatory body that ensures the safety, efficacy, and security of medical devices and software in the healthcare sector in the U.S. However, its guidelines and regulations often set a global benchmark, influencing medical device standards in Europe and other regions. The FDA recently published a draft guidance, the Predetermined Change Control Plan (PCCP), aiming to support medical device manufacturers with the release of continual learning Machine Learning-Enabled Device Software Functions (ML-DSF). Such ML-DSFs are intended to change after initial market approval. We present a systematic process to support the implementation of the PCCP. Building upon the Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology (CRISP-ML(Q)), we present an approach that a manufacturer may use to identify and evaluate the impact of anticipated changes to ML-DSF. Our process also indicates a forecast, whether the anticipated change would be accepted by the FDA as a part of the PCCP.

**Keywords:** FDA · Predetermined Change Control Plan (PCCP) · CRISP-ML(Q) · Machine Learning-Enabled Device Software Function (ML-DSF)

## 1 Introduction

Medical devices are safety critical systems that are typically subject to strict regulatory restrictions. In the U.S., these are imposed by the Food and Drug Administration (FDA), which approves the medical product prior to market release. As part of this, software functions of such systems need sophisticated

testing. Depending on the change of software after market approval, a re-approval becomes necessary. In the EU, the Medical Device Regulation (MDR) is mandatory and defines a rigorous certification process. However, for this paper, we concentrate on the development of the FDA.

Today, more and more such software functions are developed using machine learning techniques. Often, such learning Machine Learning-Enabled Device Software Functions (ML-DSF) are developed iteratively, meaning software changes are continually applied, either by manually triggered re-learning of functions or improving the system automatically whenever new data is available.

ML applications that do not change after market approval can be approved under current FDA regulations. However, any manual update would require re-approval. Systems that update automatically could not even get approval. This is because continually learning systems may adapt to changes in the environment or input on their own. Such a system would change dynamically without manufacturer supervision after initial FDA approval. As a result, the device in its updated state no longer conforms to what was originally tested and approved. This is contrary to current regulations, which require that any medical device is used in its approved form, with no subsequent changes that could affect its functionality. However, this hinders the usage of the benefits of continual learning systems which may improve their quality in a continual fashion.

To address this problem, the FDA initiated discussions with corresponding stakeholders and proposed a draft guidance to support the use of continual learning machine learning software as medical devices, hereby assisting medical device manufacturers with the development and the approval process of such systems. The most recent guidance is the Predetermined Change Control Plan (PCCP) [27]. Briefly explained, the FDA expects the medical device manufacturers to state at application for initial approval, what changes to the Machine Learning-Enabled Device Software Function (ML-DSF) are expected to occur during the lifetime of the medical device and how this change would affect the overall device. The anticipated changes are compiled into the PCCP. It is emphasized that the FDA expects to define the PCCP at initial approval, potentially long before the expected changes may occur. The intention is that if an acceptable anticipated change occurs and the manufacturer acts according to what he or she stated in the PCCP, a modification to the ML part of the ML-DSF may be distributed *without* requiring re-approval by the FDA [27]. Otherwise, that is if a change has not been anticipated, or if the steps to address the changes turn out to be not suitable or sufficient, a completely new approval by the FDA may be needed. Note that in this case, distributing the modified medical device would constitute adulteration and misbranding.

To support the development of PCCP, this paper proposes CRISP-PCCP as a new methodology to systematically identify effects and implications of changes during the release process of an ML-DSF. It is inspired by the Cross-Industry Standard Process for the development of Machine Learning applications with Quality assurance methodology (CRISP-ML(Q)) [19]. In simple words, CRISP-PCCP consists of several steps that are iterated. The first step of CRISP-PCCP



is to provide a sufficiently precise description of the anticipated device change. Next, one must walk through the phases and sub-phases (generic tasks) of an ML processing pipeline and determine for each generic task, whether it is affected by the anticipated device change. If so, additional generic questions for further details must be answered. In this way, a deeper understanding of the potential changes is obtained, and a subsequent risk re-evaluation must be performed to determine their potential impact on the overall system safety, effectiveness, and compliance with regulatory standards. This last step is called the change impact analysis. Following the proposed methodology not only sharpens the conception of the anticipated device change and its consequences, but also provides an estimation of whether an anticipated device change will be accepted by the FDA as part of the PCCP.

The rest of the paper is organized as follows: In Sect. 2, the increasing importance of continuously learning AI/ML products is highlighted and the need for a forward-looking change control plan for AI/ML models in the medical device industry is underlined. Then we continue in Sect. 3 with a brief overview of the regulatory context. In Sect. 4 we describe the process, which uses the CRISP-ML(Q) as its foundation: The risk-based approach for manufacturer to identify changes along the ML life cycle and to predetermine their impact during field usage. Further we estimate FDA acceptance within each phase of the process. Finally, we discuss our experience and findings with the development and usage of the approach on real world projects in Sect. 5 and conclude with steps to improve the approach in Sect. 6.

## 2 Machine Learning in Medical Devices

Machine Learning (ML) is an area of computer science dedicated to developing systems that can execute tasks usually associated with cognitive processes. ML systems can analyze data, recognize patterns, make decisions, and adapt to evolving situations without explicit programming. Within the medical domain, ML has transformed medicine by enhancing diagnostics, personalizing treatments, streamlining drug discovery, and enabling predictive analytic [11]. However, it also presents challenges related to data privacy, regulation, and ethical use [2, 14]. The ethical issues referred to include concerns about bias and fairness in ML algorithms. Bias can occur when an ML system generates skewed or prejudiced results due to flawed assumptions in the algorithm or biased data inputs. This in turn can lead to unfair treatment of individuals or groups [14].

In comparison to classical V-Model driven software development, where requirements engineering and testing are typically performed at well-defined stages in sequence [9], the development of ML models is performed slightly different. The typical ML development process involves developing a model that is taken through a series of iterative steps that require constant adaptation and learning as the model interacts with data. At the center of this process are the ML algorithms that serve as a blueprint and dictate how the system should achieve its goal. The model acts as an instance of the algorithm and dynamically adapts



through training iterations to approximate a target function that is initially unknown<sup>1</sup>. This iterative nature of development and focus on adapting models through learning makes the traditional software development life cycle less suitable to the nuanced and evolving landscape of ML development [5, 13].

ML approaches may be distinguished whether they are static or continual: A static model is trained offline. The model will be trained till it reaches a defined goal in predicting certain features. After the training phase, the model will be used without further changes. However, static models often perform well on similar data but could perform poorly in scenarios that are rare in the training process. Also, they prevent the ability to learn from post-approval, real-world data, and thus cannot improve over time in the same way as adaptive systems.

Dynamic models, also known as continual learning ML models, are trained online. The model will also be trained till it reaches a defined goal in predicting certain features. But after the training phase, the data that is continually processed by the system, is also used to update the model [28]. Continual learning ML algorithms are designed to update and improve themselves as their input data, environments, and/or targets change. This ability to adapt to changing data has the potential to create more advanced Machine Learning-Enabled Device Software Function (ML-DSF) that would allow to improve the performance [15, 28].

However, it poses risks that need to be addressed, such as the introduction of new errors, system performance deterioration, if the newly integrated data are biased, and the risk that new information could interfere with what the model has already learned. Therefore, it is important to carefully manage and monitor these systems [15, 28].

In addition, ML-DSF with continuously learning abilities would result in an unknown and undocumented version of the medical device software. This would lead to an illegal product under the current regulations. The FDA has recognized this issue and is actively addressing it within their “Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (ML)-Based Software as a Medical Device [23]”. This framework aims to create a clear pathway for ML-DSF that are subject to continuously learning and adaptation, allowing them to improve over time while ensuring patient safety and device effectiveness.

### 3 Regulatory Context - Approval of Medical Devices

In the United States, the Food and Drug Administration (FDA) serves as the regulatory authority responsible for overseeing Food, Pharmaceuticals, and Medical Devices (MD). Its primary mission is to guarantee that these products adhere to the highest standards of safety, effectiveness, and quality before they enter the market. The FDA achieves this through rigorous and comprehensive regulations designed to safeguard consumers’ health and well-being.

---

<sup>1</sup> This is the essence of machine learning (ML).

The FDA defines a medical device as any product, including accessories, intended for diagnosing, curing, treating, or preventing diseases in people or animals [25]. It covers a vast range of items, from simple tools like bedpans to complex technologies like pacemakers. These devices work primarily through physical or mechanical means, rather than chemical action, and are classified based on their *intended use* and the *level of risk* they pose.

The FDA employs three key processes for approving Medical Devices (MDs) to meet quality and safety standards. The *510(k) Process* is the most common one and used for devices that are substantially equivalent to an approved one. For manufacturers, it suffices to give details showing that their medical device is in substantial equivalence with a previously cleared (approved) one. The *De Novo Classification Process* is for innovative MDs without an existing comparable device. Here, manufacturers request a new classification by providing unique device evidence for FDA review. The *Pre-market Approval (PMA) Process*, which is for high-risk MDs, requires extensive data and a clinical study demonstrating clinical benefits. The FDA reviews this to ensure safety and efficacy. These processes are risk-based, with regulatory scrutiny varying according to each MD's risk level.

### 3.1 Change-Management for Software in Medical Devices

The FDA regulatory frameworks lay out guidelines for managing software changes in MDs to ensure the continued safety and effectiveness of these devices. The FDA emphasizes the importance of effective change control procedures e.g. for software changes [26]. The types of changes differ based on their impact on the device's safety and effectiveness.

*Minor changes* typically have a low impact on the device's safety and effectiveness. Manufacturers can implement these changes without prior FDA approval but must notify the FDA within 30 days of making the change. This type of change encompasses modifications to labeling or bugfixes in the medical device software.

*Moderate changes* have a more significant impact on the device's safety or effectiveness. Manufacturers are generally required to submit a new 510(k) submission or a PMA supplement for these changes, seeking FDA clearance before implementing them. The FDA will review the submission to ensure that the modifications do not compromise the device's safety and efficacy. This type of change encompasses modifications to design or functionality which could affect the medical device software's performance.

*Major changes* have the potential to significantly affect the safety and effectiveness of the device. Manufacturers typically need to submit a new 510(k) submission or a PMA supplement to the FDA for approval before implementing these changes. The FDA's review process for PMAs is more rigorous and involves a comprehensive assessment of the new information to ensure that the modified device continues to meet regulatory standards. In cases where a device that was originally cleared through a 510(k) process undergoes significant changes that might push it into a higher risk category or significantly alter its intended use,

the FDA may require a new PMA application instead of just a new 510(k) submission. This would typically be the case if the changes affect the fundamental technological characteristics or the safety and effectiveness of the device, thereby necessitating a more comprehensive review than what is covered under the 510(k) process. However, simply submitting a PMA supplement for a system that was approved via a 510(k) is not a typical pathway. The decision to require a PMA, instead of another 510(k) submission, is based on the nature of the changes and the potential risks associated with them.

Alterations to the fundamental design or intended use of an MD may be considered significant changes. These changes often require thorough evaluation and may necessitate updated clinical evidence. Changes to software, including updates, bug fixes, or enhancements, are relevant for MDs with software components. The regulations require careful consideration of the potential impact on safety and performance. It is crucial for manufacturers to thoroughly assess and document these changes in accordance with the requirements. Depending on the nature and impact of the change, manufacturers may need to update their technical documentation, conduct additional testing, or even perform a re-approval.

### **3.2 Predetermined Change Control Plan for Machine Learning-Enabled Device Software Function**

The change management described above must also be applied to ML-DSF. This leads to the fact that only static ML models would be accepted by the FDA for approval as continual learning would lead to a modification of the approved MD, resulting in a loss of approval. Manufacturers would be forced to re-validate their devices each time the continual learning ML-DSF would adapt the ML model.

The FDA recognized this as a problem and stated the “Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions” early 2023 [27]. In that recommendation, the FDA renders the PCCP as “the documentation describing what modifications will be made to the ML-DSF and how the modifications will be assessed”. Thus, the FDA requests manufacturers to identify and assess the anticipated changes to their ML-DSF in a PCCP, which will be submitted during the approval process. If the PCCP states that the assessment of the anticipated change has no impact on the general performance or safety of the MD, then the FDA is likely to approve the continual learning-enabled MD. If the ML-DSF changes as defined in the PCCP, the device’s approval persists without the need for re-approval.

According to the FDA, the following modifications fall under the scope of the PCCP: Modifications to an ML model which is “implemented automatically (i.e., for which the modifications are implemented automatically by software)” [27], which does not explicitly involve continual learning but also would not exclude this possibility, and, modifications to an ML model which is “implemented manually (i.e., involving steps that require human input, action, review, and/or decision-making, and therefore are not implemented automatically)” [27].

It is the manufacturer’s responsibility to ensure that the changes are indeed following the PCCP. When conformance with the PCCP is erroneously assumed, the approval of the system vanishes. As such, it is in the benefit of the manufacturer to install monitoring means to ensure compliance with all previously defined requirements, especially the PCCP. To this end, manufacturers prepare Standard Operating Procedures (SOPs) that detail the ongoing monitoring and evaluation processes for PCCP-approved medical devices. These SOPs should outline how data on device performance, safety, and efficacy will be collected and managed post-market, how risk management will be conducted continuously, and how changes to the ML-DSF will be assessed and documented. The SOPs must also specify the roles and responsibilities of personnel involved in monitoring, the methods for reporting and communicating findings, and the procedures for maintaining compliance with FDA regulations. By implementing these SOPs, manufacturers can ensure that any modifications to the ML-DSF remain within the approved scope and that the device continues to meet safety and performance standards without requiring re-approval.

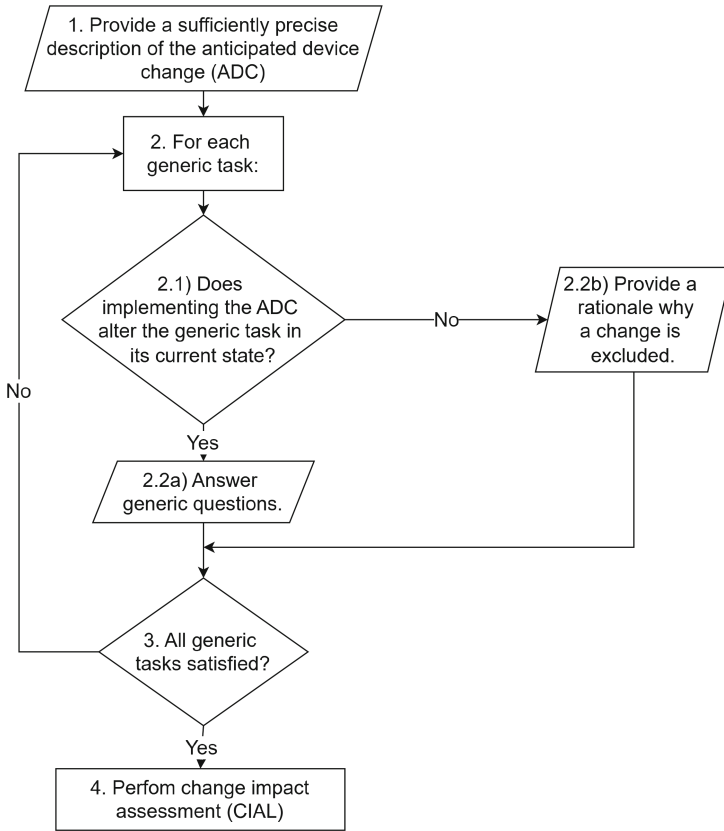
## 4 A CRISP-PCCP

We come now to the main contribution of this paper by introducing CRISP-PCCP, the tailored process to systematically anticipate the consequences of a change of a Machine Learning-Enabled Device Software Function (ML-DSF). It aims to assist in formulating the PCCP described in FDA’s recent draft [27]. If a manufacturer follows the process, he or she will receive an estimation on how high the chance of acceptance by the FDA would be for a particular anticipated device change (ADC). Moreover, the manufacturer may use the CRISP-PCCP as a documentation input for the PCCP.

Note that CRISP-PCCP neither addresses non-ML device changes nor provides general advice on developing ML-DSFs. CRISP-PCCP focuses purely on changes to the ML component. CRISP-PCCP builds heavily on the *CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance methodology* (CRISP-ML(Q)) framework proposed by Studer et al. [19].

CRISP-PCCP is implemented by performing the following steps (see also Fig. 1):

1. provide a sufficiently precise description of the anticipated device change (ADC)
2. for each generic task:
  - 1) argue whether implementing the ADC alters the generic task in its current state
  - 2a) if so, answer generic questions, interpreting them appropriately (see Sect. 4.1)
  - 2b) if not, provide a rationale on why a change is excluded
3. revisit generic tasks as needed (e. g. if side effects and dependencies are discovered), until satisfied

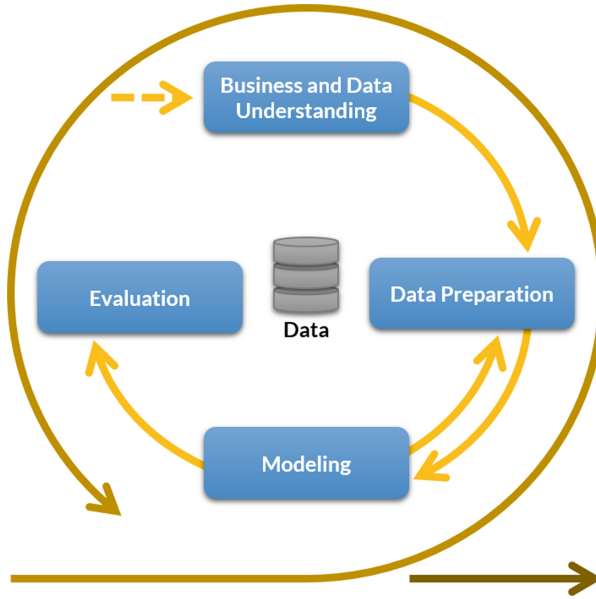


**Fig. 1.** Flowchart for Iterative Assessment and Implementation of Anticipated Device Changes (ADC) in CRISP-PCCP.

4. performing a change impact assessment (see Sect. 4.5)

These steps must be repeated for each phase in the CRISP-PCCP. The phases of CRISP-PCCP can be seen in Fig. 2. It is assumed that if a manufacturer follows this process and document the result as a PCCP request he or she will be able to demonstrate a thorough understanding and control over the anticipated changes to the device, ensuring that each modification is evaluated for its impact on safety and effectiveness.

Let us use an example to guide us through the description of the various generic tasks to come. The example is closely related to one of the prototypes we used to develop our approach. However, depending on the generic task to explain, we vary the purpose and the capabilities of the example, without restricting ourselves to the actual prototype. For more information regarding the prototype, see <https://ki-sigs.de/projekt/AP%20310> (in German).



**Fig. 2.** Overview of the phases of CRISP-PCCP and their interaction.

*Example 1.* Imagine a portable Optical Coherence Tomography (OCT) scanning device [12], which captures an image of the patient’s retina, performs a semantic segmentation to localize macular degeneration (if present) and diagnoses whether the macular degeneration is dry or wet. The computation is performed locally on the device. The image capturing and preprocessing is realized by rule-based (non-ML) software. The semantic segmentation is done by ML software based on a U-Net<sup>2</sup> model. The classification, whether macular degeneration is present and whether it is dry or wet, is performed by rule-based software. Further downstream tasks, like presenting the classification result to a user, are realized by rule-based software.

Starting at Sect. 4.1, we present the relevant generic tasks within our approach. We adjusted some tasks to better fit the development of ML-DSF in the medical context and assigned an impact level to each task, indicating the acceptability of the anticipated device change within a PCCP.

In the overview given above, we mentioned to answer generic questions, if a generic task is altered by an ADC. By that, we mean to answer thoroughly

- (i) *what* is going to be changed,
- (ii) *why* is the generic task going to be changed
- (iii) *where* is the change going to take place (potentially affected region, emphasis on physical location), and

<sup>2</sup> A U-Net is a special Convolutional Neural Network that was primarily developed for the segmentation of image data in medical image processing [17].

- (iv) *who* is going to be responsible for the change (required qualifications still met)?

If there is no clear answer to one of those questions or the change does not affect the ML part of the ML-DSF, a rationale for that generic task has to be given to explain why the change may be excluded. Further we use a set of generic tasks as the basis, derived from CRISP-ML(Q). A manufacturer is free to add or remove generic tasks to the phases of CRISP-PCCP to fully meet the needs of their ML-DSF-enhanced product. Since CRISP-PCCP is modular, adapting the process to suit the manufacturer’s specific requirements is often beneficial. We will now walk through the relevant generic tasks of CRISP-PCCP, phase by phase.

#### 4.1 Data and Business Understanding

Quoting Studer et al. [19], this “initial phase is concerned with tasks to define the business objectives and translate it to ML objectives, to collect and verify the data quality, and to finally assess the project feasibility.” We build on this definition, but leave out the part regarding the project feasibility, as we assume in our context that the feasibility analysis has already been performed at an earlier stage of development, initiated by other requirements. Besides that, we identified several key topics to be considered when assessing the impact of an ADC, like the intended use, quality goals, and capturing processes, just to name a few.

*Intended Use.* The FDA refers to *intended use* as the general purpose of a product, which is the objective intent of the legally responsible representative (e. g. the manufacturer or a reseller) who labels the product [25]. This intent may be claimed over the package or in the instruction for use, the design, or the composite of the product. The FDA offers a tool that shall help to determine whether a product’s software functions may fall potentially under the scope of the FDA’s oversight [24].

*Scope.* By scope, we mean the specific functionality of the ML-DSF that is achieved through an ML model. It is important to distinguish between the intended use and the scope, as the scope may change without affecting the intended use. The following argument explains this distinction: The scope of Example 1 covers performing semantic segmentation to localize macular degeneration, given an RGB image. Stating whether macular degeneration is present (e. g. if the segmentation exceeds some threshold), or even deciding whether the degeneration is wet or dry, is not determined by ML software and therefore not part of our scope. An ADC is to replace the non-ML classification software part with an ML model, leaving everything else as is. The scope is affected but not the intended use.

Therefore, in our Example 1, the generic questions must be answered, which we do in Table 1. For brevity reasons, we omit the questions in the remaining examples.

**Table 1.** Answers to generic questions for an ADC affecting the scope of Example 1.

What?	The rule-based component which classifies the segmentation is replaced by an ML model
Why?	Experiments indicate a higher accuracy of the ML model in comparison to the rule-based approach
Where?	All devices that have the ML-DSF or for which it is subsequently made available by means of an update
Who?	ML engineer in the development unit, test engineer during integration test

*Intended Patient Population.* Citing the patient-focused drug development glossary<sup>3</sup>, the intended patient population is defined as “the group of individuals (patients) about whom one wishes to make an inference.”

To give an example how the intended patient population is affected by an ADC, assume that in the case of Example 1, the initial intended patient population is defined as persons of age 50 to 70 years. However, anticipating that one year after market submission, enough training data will be available of persons of age 40 to 80, the intended patient population is extended to that range.

*Quality Goals.* We highlight several quality goals, relevant to ML in the medical field (derived from the success criteria mentioned in CRISP-ML(Q) [19]). Namely, the *diagnosing performance goal*, *runtime performance goal*, *robustness goal* and the *human understandability goal*. We emphasize the term *goal* in these cases, as we are interested in the intended consequences of an ADC. When applying our method to the example, we obtained several useful insights: Most ADCs can impact certain qualities in some way, even if that was not the original intention. To avoid meaningless answers to our questions, we adopt a specific approach. Often, the side effect of an ADC on a specific quality (not the goal) cannot be ruled out. Therefore, we focus only on intended consequences, which we refer to as quality goals. This means we aim to maintain the effects on diagnosis performance, runtime performance, robustness, and human understandability within reasonable limits. It is worth noting that existing quality assurance processes already cover these aspects.

Ideally, every goal is defined by a measurable metric or verifiable [1, 16, 27]

By diagnosing performance we mean properties like accuracy, specificity, sensitivity, and so on Runtime performance includes computation time, memory and storage consumption, and energy requirements. Robustness denotes “the degree to which a component can function correctly in the presence of invalid inputs or stressful environmental conditions,” see [7]. By human understandability, we refer to “the ability to explain or to present in understandable terms to a human”, see [3].

<sup>3</sup> Refer to <https://www.fda.gov/drugs/development-approval-process-drugs/patient-focused-drug-development-glossary>.



*Capturing Process and Digital Representation.* By capturing process we mean the transition of turning a physical object or an action into a digital representation. The digital representation is the result of the capturing process, usually the output of the capturing device. For example, recording an image is a capturing process and the image file is the digital representation.

Considering again the above mentioned portable OCT device. Assume that the training set of the original device (OD) consists of images captured by a stationary OCT scanner in a lab environment, creating images of high quality. Expanding the training set by retina images collected during production use of the portable OCT device, which by construction generates images of lesser quality and operates in diverse environments, is an example of a change in the capturing process and its digital representation.

## 4.2 Data Preparation

Data preparation covers all tasks involved in the transformation of data in its digital representation (see above) to a form which is accessible by ML models. This includes tasks operating directly on the digital representation, like selection, cleaning, and imputation, but also conversion steps like construction, integration and formatting.

In contrast to CRISP-ML(Q), we do not divide this phase into smaller parts (generic tasks). One reason is that in practice, we experienced that data preparation (preprocessing) closely intertwines the mentioned generic tasks. For example, many deep learning frameworks provide functions which convert JPEG images to a multidimensional array of floating point values in the range of  $[0, 1]$ . These functions combine construction and formatting into one step. Often, they allow to integrate cropping, linearly transforming, and normalizing functionality, which then also covers selection and cleaning. Another reason is that for our approach, it is not necessary to distinguish between the generic tasks, as they are treated equally regarding their Change Impact Assessment Level (refer to Sect. 4.5 for further details).

## 4.3 Modeling

By modeling we mean the declaration (implicitly and explicitly) of a space of learnable functions which is later systematically searched for a good (in terms of some metric) solution. To find, at least in principle, a good solution within said space, it is necessary that the space of learnable functions contains a good solution in the first place (the space should not be too small). To find the solution in a reasonable amount of time, it also should not be too large. Thus, it is important to carefully select the space of learnable functions.

The following paragraphs describe the principles that affect its size and its content.

*Domain Knowledge and Data Assumptions.* Domain knowledge and data assumptions incorporated into the OD may not be valid with regards to an ADC.

For example, assume that the image capturing device of Example 1 can record more image modalities than just those by cameras using visible light. The other modalities were recorded during production use and integrated into the training set, which did not contain such images during the development of the OD. The segmentation component of Example 1 will no longer be a two-dimensional U-Net, but multidimensional. Domain knowledge and data assumptions that expected two-dimensionality may no longer be valid for higher dimensions.

*Modeling Technique.* By modeling technique, we mean the selection of the class of models to choose from. For example, modeling the scope via decision trees is a modeling technique. Modeling it instead via fully connected deep neural networks is another. Pre-training, transfer learning, and assembling also fits into this generic task.

The modeling technique is affected by an ADC, if for example a novel feed forward network module is introduced in the literature and the manufacturer decides to incorporate it into the ResNet<sup>4</sup> of the OD.

*Tuning Procedure.* Identifying the space of potential learnable functions constitutes an initial step within the model development framework, whereas the precise selection of a singular function from this space encompasses a distinct and complex challenge. By tuning we mean the guided selection of a learnable function, typically facilitated by a designated dataset known as tuning data<sup>5</sup>. This selection process commonly adopts the formulation of an optimization problem aimed at evaluating and ranking the candidate functions within the specified space. Usually, the candidates of a given space are ranked along a specifically formulated optimization problem. By defining loss functions and regularization terms, candidates are valued, preferring the ones with higher value (or equivalently, lower loss). To systematically (and hopefully efficiently) search only for promising candidates, optimizer are applied. Depending on the way the optimizer operates (initial solution, local optimization, global optimization, . . .), some candidates are effectively excluded. Considering again the example we gave for the scope, replacing a non-ML unit with an ML unit will affect the tuning procedure of the OD, or introduce a second one.

*Reproducibility.* The modeling is reproducible if, based only on the modeling documentation, the previously learned function can be identically recreated. Often this fails due to implicit data assumptions, non-written domain knowledge, or (hidden) randomness.

Looking again at our running example, an ADC might add more stochasticity to the ML model (e. g. by introducing variational parts). If the manufacturer misses to keep track of the randomness passed through the training, a later reproduction fails.

<sup>4</sup> A ResNet (Residual Network) is a type of Convolutional Neural Network specifically designed to train deeper networks by addressing the vanishing gradient problem [6].

<sup>5</sup> This is often called validation data, when splitting the available data into the training, validation, and test set. We prefer, and so does the FDA, the term tuning, as it avoids confusion with the meaning of validation in the medical context.

#### 4.4 Evaluation

In the evaluation phase, we check whether the performance, robustness, and human understandability goals, defined in phase 1, are met by the learned function obtained from phase 3. In its draft of the PCCP [27], the FDA states that the modification protocol describes “[...] the methods that will be followed when developing, validating, and implementing modifications [...].” Thus, it is mandatory to examine whether the evaluation procedures for the OD are still suitable for evaluating an ADC.

For example, an ADC of the portable OCT device mentioned above is to diversify the intended patient population, by adding corresponding training data which will be acquired after the first market admission. Additional *diagnosis performance* evaluation explicitly targeting the newly affected patient population is necessary.

#### 4.5 Change Impact Assessment and Summary

In the previous sections, we presented those generic tasks of the CRISP-PCCP phases, which we deemed appropriate for our approach. Treating each of these generic tasks and eventually answering the related questions gives a rather detailed view of the impact an ADC may have. However, a broad picture or a conclusion might be unclear. For this reason, we propose a systematic method to condense the change impacts into a single number: the Change Impact Assessment Level (CIAL).

The FDA identified in [23] three broad categories of changes of an ML-DSF: performance changes, input changes and intended use changes. In our view, changes related to the ML-DSF’s performance generally have the highest chance of being compatible with a PCCP (meaning that the FDA will probably accept anticipated changes of this kind in most cases). For example, the usage of additional training data from the intended patient population, gathered from field usage to increase the accuracy of the ML-DSF, is a change that is likely to be compatible with a PCCP. To such a change we assign the Change Impact Assessment Level (CIAL) **3**.

For changes related to the ML-DSF’s input, e. g. the dimensions or resolution of input images and inclusion of additional features, we assume the chance of being PCCP compatible to be on par with the chance of being PCCP incompatible. The broad range of ways to change the input with (more or less) wide-ranging effects justify our view. Therefore, we assign to each input change the CIAL **2**. For each CIAL 2 change individually, we suggest consulting experts and/or reaching out to the FDA as early as possible, to increase PCCP compatibility chance of that change.

Changes to the ML-DSF’s intended use have, according to the FDA ([27, p. 17]), a low (but non-zero nevertheless) chance of being PCCP compatible. We denote the CIAL of such changes by **1**.

Guided by the principles above, we assign to each generic task (or phase) of the CRISP-PCCP process model the most fitting change category (performance

changing, input changing, intended use changing), yielding the corresponding CIAL (see Table 2).

**Table 2.** The CIAL compatibility classes per generic task (*or phase*). The higher the class, the more likely we deem a corresponding change to be PCCP compatible (3 being the highest, 1 being the lowest).

Generic Task ( <i>phase</i> )	CIAL
<i>Data and Business Understanding</i>	
Intended Use	1
Scope	2
Intended Patient Population	2
Quality Goals	3
Capturing Process & Digital Representation	2
<i>Data Preparation</i>	
<i>Modeling</i>	
Domain Knowledge & Data Assumptions	2
Modeling Technique	1
Tuning Procedure	2
<i>Evaluation</i>	
	1

By definition, the *intended use* will receive a CIAL of 1.

To the *evaluation phase*, i. e. the evaluation of the four quality goals diagnosis performance, runtime performance, robustness, and human understandability, we also assign a CIAL of 1. The reason is that the approval of a medical device strongly correlates with the degree to which the ML-DSF fulfills the quality goals and the validity of the degree itself depends on the comprehensiveness and thoroughness of the evaluation method [18,22].

Since increasing the *scope* of an ML-DSF probably lessens the PCCP compatibility, but decreasing the scope of an ML-DSF probably raises the PCCP compatibility, we assign the CIAL 2, to take the unclear situation into account. Considering the case individually may allow for a change of the CIAL to 1 or to 3.

If changing the *intended patient population* resembles rather an extension of the existing population (e.g. enlarging the age interval to both sides) and supporting arguments to do so exist, we suspect a high probability of PCCP compatibility. On the other hand, if the intended patient population is expanded by a rather “orthogonal” group, that only has a minor overlap with the existing population, many other generic tasks would presumably also be affected by this expansion, which in turn make it difficult to argue for PCCP compatibility.

Another generic task with CIAL 3 is *quality goals*. We assume that changing the quality goals will only result in more sophisticated goals (never less sophisticated ones), for economic reasons, and accordingly expect a high PCCP

compatibility. The *tuning procedure* is a generic task that would fall under CIAL 2 with the potential to tend to a CIAL 1. On the one hand, most of the time the tuning procedure only intends to increase the performance or to find even better solutions to the optimization problem. On the other hand, the solution is based on actions performed by the manufacturer and falls more in the category of a software change.

The generic tasks *capturing process & digital representation*, *data preparation*, *domain knowledge & data assumption* and *tuning procedure* each encompass a diverse range of possible changes, for which we do not see ourselves in a position to proclaim either a high or a low probability of approval. That is why we assign the CIAL to all of them, resorting to the consideration of individual cases.

The last step of the change impact assessment is to check for the lowest CIAL from the generic tasks. To do so, first select only those generic tasks, which are affected by the ADC, i. e. which have answers to the generic questions and no rationales supporting their exclusion. Second, order them ascending by their associated CIAL. If there is a CIAL of 1 the chance is low that the FDA would accept the ADC for approval. If there is a CIAL of 2 the chance of FDA approval increases but it is recommended to get in touch with the FDA to discuss further conditions. And finally, if there is a CIAL of 3 the chance that the ADC will be approved by the FDA is high.

One should carefully consider whether it is worth the effort to continue the approval procedure of an ADC with overall CIAL 1. An ADC of overall CIAL 2 is more promising in that regard, but we strongly suggest involving experts and/or the FDA early in the further development process. An ADC of overall CIAL 3 is likely PCCP compatible.

## 5 Discussion

The CRISP-PCCP process was initially developed in response to the FDA’s Proposed “Regulatory Framework for Modifications to ML Software as a Medical Device” [23], as part of the BMWK-funded KI-SIGS project [10]. It was evaluated and refined in three different projects, including “PASBADIA” [21] and two others from the KI-SIGS initiative, all focusing on ML-supported medical devices.

Comprising ML and regulatory experts from the KI-SIGS project, the working groups applied CRISP-PCCP to address specific ML challenges in their respective projects. They aimed to identify and assess potential changes within the ML processing pipeline. The methodology included introducing the CRISP-PCCP’s background and goals, followed by its application. Using the current development state of the corresponding project as a baseline, upcoming development goals were identified and examined using CRISP-PCCP. This process helped in easily identifying changes outside the ML development scope, encouraging the teams to define a rationale at the end of each investigation, which could be incorporated into technical documentation as proof of proper PCCP application.

The CRISP-PCCP process proved to be a relevant procedure for the documentation of ML development results. The underlying CRISP-ML(Q) process enabled detailed examination across various ML development phases. The CRISP-PCCP template aided in posing important questions and assessing potential changes. The ML developers from different research and development projects, which were involved in the development process of the CRISP-PCCP, independently confirmed the process’s utility in evaluating changes in ML. Regulatory experts viewed CRISP-PCCP as a useful method for capturing and assessing changes and their impact on the ML model. CRISP-PCCP has potential beyond continual learning ML applications, such as being part of the development planning for ML models while identifying associated risks. Changes to the ML model can be assessed, and implementation planning can be based on the CIAI, prioritizing changes with a higher chance of acceptance.

CRISP-PCCP can also play a role in the European jurisdiction. The support for change management processes and ML development planning is also required by the Medical Device Regulation (MDR) [20], which CRISP-PCCP can assist with. Additionally, there is a questionnaire titled “Artificial Intelligence (AI) in medical devices” published by the Interest Group of Notified Bodies [8], intended for auditors to ask questions during audits about the ML product development life cycle. Manufacturers can also use this document to identify gaps in their documentation. CRISP-PCCP, due to its structure, can then be expanded to cover identified gaps and serve as appropriate development documentation.

## 6 Conclusion

CRISP-PCCP is a process adapted from CRISP-ML(Q) [19], selectively incorporating its first four phases reflecting the developmental stage of projects. It assumed that the Deployment and Quality Assurance phases from CRISP-ML(Q) are integrated in the existing development process. Despite that, CRISP-PCCP demonstrated its potential in meeting FDA’s PCCP [27] requirements in selected projects. However, its application to projects involving Continual Learning ML models and validation by the FDA itself remains pending, mainly due to the limited number of such projects in medicine.

The integration of CRISP-PCCP in the European jurisdiction, considering the evolving AI-Act [4], poses future research questions. The AI-Act, recently passed in the European Union, seeks to regulate artificial intelligence. It aims to establish a legal framework ensuring AI systems’ safety, compliance with privacy and data protection laws, and the upholding of fundamental rights. The co-existence of the AI Act and the Medical Device Regulation imposes many questions on how to address similar overlapping concerns and requires future research.

## References

1. Altman, D.G., Bland, J.M.: Diagnostic tests. 1: sensitivity and specificity. *BMJ* (Clinical research ed.) **308**(6943), 1552 (1994). <https://doi.org/10.1136/bmj.308.6943.1552>
2. Bende, P., Vovk1, O., Caraveo, D., Pechmann, L., Leucker, M.: A case study on data protection for a cloud- and AI-based Homecare Medical Device. In: Lamo, Y., Rutle, A. (eds.) *Proceedings of The International Health Data Workshop. CEUR Workshop Proceedings*, vol. 3264. CEUR, Bergen, Norway, June 2022. [https://ceur-ws.org/Vol-3264/#HEDA22\\_paper\\_3](https://ceur-ws.org/Vol-3264/#HEDA22_paper_3), iSSN: 1613-0073
3. Doshi-Velez, F., Kim, B.: *Towards a rigorous science of interpretable machine learning* (2017)
4. European Union: EU AI Act: first regulation on artificial intelligence — News — European Parliament, June 2023. <https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>, Accessed 08 Dec 2023
5. Hatcher, W.G., Yu, W.: A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access* **6**, 24411–24432 (2018). <https://ieeexplore.ieee.org/document/8351898>
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, December 2015. <https://doi.org/10.48550/arXiv.1512.03385>, arXiv:1512.03385 [cs]
7. IEEE: IEEE Standard Glossary of Software Engineering Terminology. *IEEE Std 610.12-1990* pp. 1–84 (1990). <https://doi.org/10.1109/IEEESTD.1990.101064>
8. IG-NB: Questionnaire “Artificial Intelligence (AI) in medical devices”. [https://www.ig-nb.de/?tx\\_epxelo\\_file\[id\]=884878&cHash=53e7128f5a6d5760e2e6fe8e3d4bb02a](https://www.ig-nb.de/?tx_epxelo_file[id]=884878&cHash=53e7128f5a6d5760e2e6fe8e3d4bb02a), Accessed: 12 Dec 2023
9. International Electrotechnical Commission: IEC62304:2006/AMD1:2015 Amendment 1-Medical Device Software-Software Life Cycle Processes. <https://webstore.iec.ch/publication/22790> (2015), Accessed: 30 Mar 2021
10. KI-SIGS: AI Space for Intelligent Healthcare Systems KI-SIGS. <https://ki-sigs.de/>, Accessed 01 Feb 2022
11. Obermeyer, Z., Emanuel, E.J.: Predicting the future - big data, machine learning, and clinical medicine. *New Eng. J. Med.* **375**(13), 1216–1219 (2016). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070532/>
12. Ong, J., Zarnegar, A., Corradetti, G., Singh, S.R., Chhablani, J.: Advances in optical coherence tomography imaging technology and techniques for choroidal and retinal disorders. *J. Clin. Med.* **11**(17), 5139 (2022). <https://doi.org/10.3390/jcm11175139>. <https://www.mdpi.com/2077-0383/11/17/5139>, number: 17 Publisher: Multidisciplinary Digital Publishing Institute
13. Pechmann, L., Mildner, M., Suthau, T., Leucker, M.: Regulatorische Anforderungen an Lösungen der künstlichen Intelligenz im Gesundheitswesen. In: Pfanstiel, M.A. (ed.) *Künstliche Intelligenz im Gesundheitswesen*, pp. 175–198. Springer, Wiesbaden (2022). [https://doi.org/10.1007/978-3-658-33597-7\\_8](https://doi.org/10.1007/978-3-658-33597-7_8)
14. Petersen, E., et al.: Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE Access* **10**, 58375–58418 (2022). <https://doi.org/10.1109/ACCESS.2022.3178382>. conference Name: IEEE Access
15. Pianykh, O.S., et al.: Continuous learning AI in radiology: implementation principles and early applications. *Radiology* **297**(1), 6–14 (2020). <https://doi.org/10.1148/radiol.2020200038>. pMID: 32840473

16. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, October 2020. [arXiv:2010.16061](https://arxiv.org/abs/2010.16061) [cs, stat]
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation (2015). [arXiv:1505.04597](https://arxiv.org/abs/1505.04597) [cs]
18. Stewart, J.P.: Software as a Medical Device. SaMD, Clinical Evaluation. International Medical Device Regulators Forum (2017)
19. Studer, S., et al.: Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. *Mach. Learn. Knowl. Extr.* **3**(2), 392–413 (2021). <https://doi.org/10.3390/make3020020>
20. Union, E.: Regulation (EU) 2017/745 of the european parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/ec, regulation (EC) no 178/2002 and regulation (EC) no 1223/2009 and repealing council directives 90/385/eec and 93/42/eec (text with eea relevance. ), April 2017. <https://lexpency.de/eu/32017R0745/>, Accessed 17 Nov 2022
21. University of Lübeck: PASBADIA: COPICOH. <https://www.copicoh.uni-luebeck.de/forschung/projekte/aktuelle-projekte/pasbadia>, Accessed 08 Dec 2022
22. U.S. Food and Drug Administration: Software as a Medical Device (SaMD), December 2018. <https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd>
23. U.S. Food and Drug Administration: Proposed Regulatory Framework for Modifications to AI/ML Software as a Medical Device, 2019. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
24. U.S. Food and Drug Administration: Digital Health Policy Navigator (2022). <https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-policy-navigator>
25. U.S. Food and Drug Administration: Cfr - Code of Federal Regulations Title 21 Part 801.4 (2023). <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=801.4>
26. U.S. Food and Drug Administration: Deciding when to submit a 510(k) for a change to an existing device. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/deciding-when-submit-510k-change-existing-device> (2023). Accessed 29 Nov 2023
27. U.S. Food and Drug Administration: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions, April 2023. <https://www.fda.gov/media/166704/download>
28. Vokinger, K.N., Feuerriegel, S., Kesselheim, A.S.: Continual learning in medical devices: FDA's action plan and beyond. *The Lancet Digital Health* **3**(6), e337–e338 (2021). <https://linkinghub.elsevier.com/retrieve/pii/S2589750021000765>






**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Model Driven Development for AI-Based Healthcare Systems: A Review

Colm Brandon<sup>1,2</sup>, Amandeep Singh<sup>1,2</sup>, and Tiziana Margaria<sup>1,2,3,4</sup>

<sup>1</sup> University of Limerick, Limerick, Ireland  
tiziana.margaria@ul.ie

<sup>2</sup> Centre for Research Training in Artificial Intelligence (CRT-AI), Limerick, Ireland

<sup>3</sup> Lero, Limerick, Ireland

<sup>4</sup> The Health Research Institute (HRI), University of Limerick, Limerick, Ireland

**Abstract.** We review our experience with integrating Artificial Intelligence (AI) into healthcare systems following the Model-Driven Development (MDD) approach. At a time when AI has the potential to instigate a paradigm shift in the health sector, better integrating healthcare experts in the development of these technologies is of paramount importance. We see MDD as a useful way to better embed non-technical stakeholders in the development process. The main goal of this review is to reflect on our experiences to date with MDD and AI in the context of developing healthcare systems. Four case studies that fall within that scope but have different profiles are introduced and summarised: the MyMM application for Multiple Myeloma diagnosis; CNN-HAR, that studies the ability to do AI on the edge for IoT-supported human activity recognition; the HIPPP web based portal for patient information in public health; and Cinco de Bio, a new model driven platform used for the first time to support a better cell-level understanding of diseases. Based on the aforementioned case studies we discuss the characteristics, the challenges faced and the positive outcomes achieved.

**Keywords:** Model Driven Development · Health Informatics · Artificial Intelligence Applications · Low-code/no-code · Digital Thread

## 1 Introduction

The integration of AI into healthcare systems is potentially a paradigm shift for the healthcare sector, with impacts on a variety of levels, such as computer aided diagnostics, personalised medicine, drug discovery, disease understanding, and the processing of healthcare records to name a few. AI's ability to extract actionable knowledge from large amounts of unstructured data has huge opportunities to advance medical discovery and also free medical professionals from a variety of administrative tasks through automation.

There are also potential pitfalls. In computer aided diagnostics, implicit or explicit biases in the training data could lead to higher rates of misdiagnosis in underrepresented patient groups. There is a trade-off between preserving

patient privacy and maximising treatment outcomes with respect to personalised medicine. It is therefore critical to embed medical professions and other stakeholders in the development process as much as possible.

MDD can be a way to include professionals more and earlier in the development of new healthcare software and systems. MDD works using models of a given domain, be they models of organs (like a digital twin), of procedures (like workflows) or of knowledge (like databases, ontologies, and rules). Some of these models are represented as data, others as graphs (like workflows and processes) which can be later transformed into executable code, human or AI generated. These models capture the essence of the domain and the problem the system solves. Most of them enable the creation of Domain-Specific Languages (DSLs) which express the functionalities in the domain in a more intuitive way for the health professionals, and this way improve the usability and effectiveness of the models and the systems. The key property is that models and DSLs must offer a level of abstraction directly accessible to the domain experts, shielding the technicalities. The goal is that domain users may be able to understand the system's design and functioning without the need to also understand the underlying technologies for the implementation and runtime.

This paper explores our experiences to date with respect to the two dimensions of *advanced model driven development* and *AI* in the development of healthcare systems. We summarise four case studies with different characteristics in this spectrum, covering different applications in healthcare: diagnosis in clinical medicine (MyMM), primary care for neurological impairment (CNNHAR), patient information in public health (HIPPP) and the cell-level understanding of diseases (Cinco de Bio (CdB)).

The paper is organized as follows. Section 2 considers literature and previous work pertaining MDD and AI in Healthcare Systems. Section 3 presents and discusses the four cases studies. Section 4 discusses the characteristics, challenges and successes in the cases studies. Finally Sect. 5 offers some final thoughts on the current work and perspectives for future work.

## 2 Model Driven Development and AI in Healthcare Systems

We consider here the two main perspectives of AI and MDD in healthcare.

### 2.1 AI in Healthcare

AI and Machine Learning (ML) are increasingly embedded in healthcare systems all the way from the laboratory, in biomedical research, to clinical medicine and patient care at home.

In health related research, AI models have proven to be extremely effective for the prediction of protein structures [27] in drug discovery, and to predict physiochemical properties [67] of environmental chemicals. When understanding

diseases, AI models enable single cell segmentation in tissues, and thus single cell analysis [23,37].

In the context of primary care, Internet of Things (IoT) devices and AI at the edge have been used effectively to mitigate the negative side effects (without direct human intervention) for individuals with a variety of illnesses. An excellent example is the use of smart devices to predict and detect gait freezing episode of Parkinsons patients, so that an automated intervention can un-freeze the patient [10, 11].

In public health, AI and ML have been successfully used for information analysis and data mining. During the COVID pandemic, Irish researchers developed a method of estimating the reproduction or 'R-number' in COVID-19 cases, the number that describes the likelihood of a virus increasing or decreasing in a population. Their susceptible-exposed-infected-removed - or SEIR - model, based on statistics and ML, was used regularly to provide up-to-date scenario analysis to the National Public Health Emergency Team (NPHET) [21]. Similarly, AI helped identify diseases and foodborne illness outbreaks through data mining social media data [54,58,59]. AI has also been used to identify individuals suffering with mental illnesses through the analysis of social media data [30,48,62]. AI models have been developed to mitigate the effect of low-quality health information on treatment outcomes of patients, by automatically evaluating the quality of web-based health information with respect to some established evaluation frameworks developed by physicians [31].

In clinical medicine, AI models support clinicians in a variety of ways. Medical image classification models assist in the diagnosis of a variety of diseases [25,34,65,68]. Examples are mammogram classification in the context of breast cancer diagnosis [4], heart disease [46], and many more. The possibility of using Large language Models for computer aided diagnosis is being explored [61], along with the development of clinical decision support systems that use other AI methods [66]. LLMs are also expected to help reduce the ever growing administrative workload.

## 2.2 Model Driven Development in Healthcare

MDD has been used extensively in the healthcare domain in recent times with impacts across a broad range of applications such as clinical medicine, information systems for healthcare and health monitoring systems.

In the context of clinical medicine, MDD has been used to simplify the process of developing software systems for clinical scoring and mobile health through the use of meta-modelling [18]. A meta-model that facilitates the development of dynamic checklist support systems for clinical safety helps reduce the cognitive load for clinical stakeholders when designing such systems [49].

With respect to information systems for healthcare, a MDD based versatile Nursing Information System (NIS) can cater for a variety of clinical requirements across a range of clinical settings and be deployed on different devices [63]. For general Hospital Information Systems (HIS), a multidimensional supervision metamodel was developed to create a general framework that mitigates

the challenges of developing such systems which have to cater for a wide variety of end-users and healthcare delivery processes [35]. Agile and model driven methodologies have been shown to aid in the development of Electronic Health Record (EHS)-based population registries and their associated data collection tools [28]. Model-Driven Engineering (MDE) principles have also enabled software engineers to develop HL7 (a healthcare information and protocol standard) models using a Unified Modeling Language (UML) based interface [50].

In the context of personal health monitoring, a model-driven computational framework for Personalized Health Monitoring (PHM) applications using IoT devices poses specific emphasis on security and energy issues [60]. A framework for general IoT-based healthcare monitoring systems presents a meta-model which defines the fundamental components of such systems and their relationships [3].

In the specific context of MDD, we use a set of Integrated Modelling Frameworks that are described next.

### 2.3 Model Driven Development Frameworks

Our approach simplifies the IT and integration tasks by composing *models of behaviour* to achieve the same results. By introducing the different DSLs as the high-level abstraction for each behaviour and plugging them in, we enable a development cycle where stakeholders, instead of only defining requirements, can take action by participating actively in the successive stages of the project. This approach has been successful in several domains, for example in smart agriculture applications [24].

Our contributions cover the demonstration of various aspects of the use of two Low-Code/No-Code (LCNC) development environments, DIME [12] and Pyrus [69], and the extension of their Application Domain Specific Languages (ADSLs) to cover healthcare settings. To accomplish this, we extend the range of A-DSLs that these two environments provide in order to enable our external data sources and systems to become amenable to our LCNC application development approach, allowing a high level of reusability. This extension involves creating several building blocks to communicate with the different systems, that may as well include IoT devices like the low-cost compact multi-sensor platform *Nordic Thingy:53*<sup>1</sup> that we use for example in the Human Activity Recognition (HAR) context. The A-DSLs in particular enable analytics pipelines for a subsequent decision-making.

We embrace a LCNC software development paradigm [9], that is rapidly gaining foot in industry and is predicted to become the development style of choice for 80% of newly developed software by 2026 [1]. However, we specifically adopt a Model Driven Design and development paradigm [38, 45] where the models are not just graphically suggestive but also have an underlying formal model in terms of Kripke Transition Systems [33]. This choice makes them analyzable through

---

<sup>1</sup> [https://docs.nordicsemi.com/bundle/ug\\_thingy53/page/UG/thingy53/intro/frontpage.html](https://docs.nordicsemi.com/bundle/ug_thingy53/page/UG/thingy53/intro/frontpage.html).

well-established techniques like control flow and data flow analysis [36], model checking [8], property checking [56], reachability analysis [29] and more, like synthesis [40, 41], also in robotics and IoT contexts [26]. Specifically, we use here the two Integrated Modelling Environments DIME and Pyrus.

**DIME** is an Eclipse-based LC/NC graphical modelling environment that enables prototyping web applications in a model-driven way. It follows the One Thing Approach (OTA) [43] and the eXtreme Model-Driven Design (XMDD) [44] paradigms for modelling and development, empowering domain experts to model an end-to-end web application with no programming experience. To cover the different aspects of web applications, DIME provides a collection of ready-to-use modelling languages, as well as collections of native DSLs that support the development of new applications via composing models of different natures. In particular, DIME allows users to define the data model, create the user interface models, and create the workflows. Those models are checked for syntactic compliance largely automatically<sup>2</sup>.

**Pyrus** is a web application that offers a graphical, collaborative development environment for Data Analytics. It bridges the gap between Python-based established programming platforms like Jupyter [32] and graphical workflow composition in a data-flow fashion. Individual Python functions are implemented and stored in Jupyter, special signature annotations are added to these functions and exported to Pyrus, so that the functions can be identified and retrieved by the Pyrus web-based orchestration tool, where the pipelines are composed. From the pipelines, Pyrus generates the Python code for the orchestration and configuration, which is again stored and executed in Jupyter. This separation of concerns decouples the coding and development of the single functionalities (in Python/Jupyter), the low-code part of the approach, from the data analytics orchestration modelling, which happens in accordance with model-driven engineering principles and completely graphically, in a no-code fashion.

### 3 Case Studies

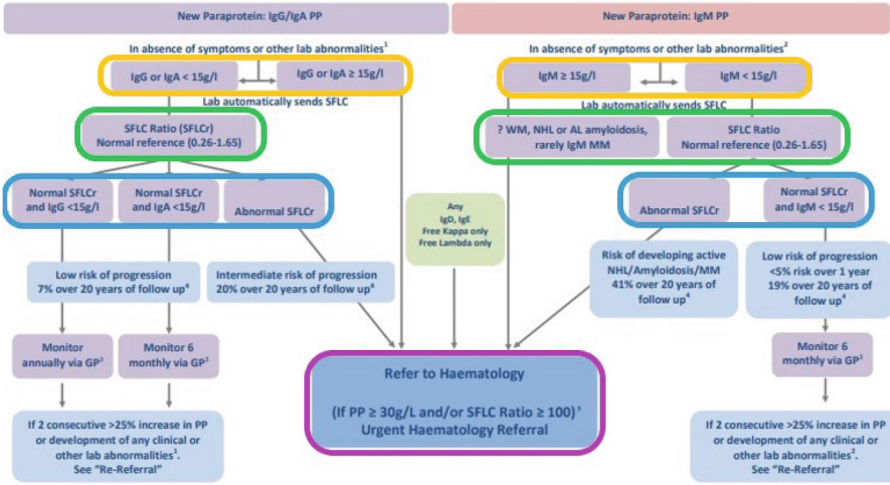
Here we summarise four selected case studies that exemplify our experience of bringing model driven development and/or AI into the healthcare domain.

#### 3.1 Automating the Diagnosis of Multiple Myeloma XMDD and Web

Multiple Myeloma (MM) is a blood cancer that develops in plasma cells in the bone marrow. It is one of the deadliest forms of blood cancer, with a five-year survival rate as low as 40%. In collaboration with the University Hospital Limerick, in the context of the interdisciplinary UL Cancer Network (ULCaN), we automated the referral pathway for MyMM through a Web application developed in DIME, using XMDD.

---

<sup>2</sup> DIME can be downloaded from <https://sccc.gitlab.io/dime/>.



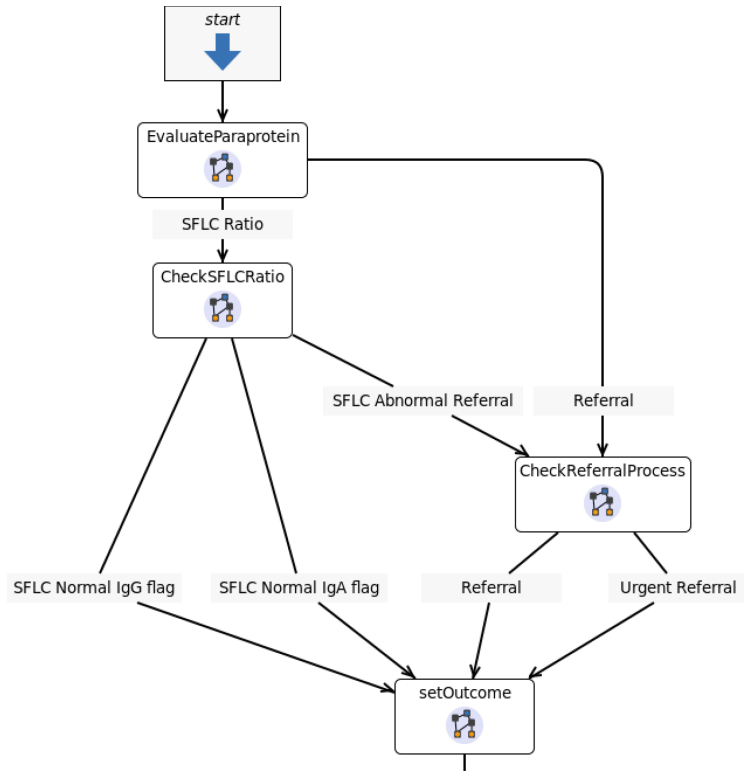
**Fig. 1.** Extract from UHL Guidance on Management of MGUS in Primary Care [51]: risk stratification workflow.

MM is a cancer with circulating monoclonal paraproteins. It is diagnosed by measuring risk factors in the blood: monoclonal proteins called Paraproteins (PP), and Serum-Free Light chain (SFLC) ratios, in blood samples. When lower levels of paraproteins are detected in healthy individuals, this is known as Monoclonal Gammopathy of Uncertain Significance (MGUS). A risk stratification protocol along the levels of such factors developed at the University Hospital Limerick [52] (see Fig. 1) recommends a patient’s further care depending on their outcome. These outcomes can range from a yearly check-up and monitoring via a General Practitioner (GP) to an urgent referral for a patient to be seen by a consultant within the Haematology department. We aimed to improve referral timelines for patients with possible MGUS and automate the risk stratification process by replacing the manual process with a web-based application developed in DIME with XMDD technologies. The MyMM web application implements the manual reference risk stratification algorithm and for patients with possible MGUS it communicates a patient’s further care to the requesting physician or GP. The system includes a process model implemented in the DIME development environment.

Figure 2 shows the corresponding top level decision process modelled in DIME, with a layout resembling the layout of the original algorithm.

It contains several process models that carry out the task of completing each comparison for each layer of the risk stratification seen in Fig. 1. Care was taken to mimic in the DIME process structure and layout of the original algorithm, which is known to the healthcare specialists from their training.

Solving this particular problem is important because the automation makes the evaluation and communication of outcomes immediate, instead of waiting



**Fig. 2.** The top process model of the risk stratification process (in DIME).

for days until a trained specialist finds time to manually score the accumulated batch of samples. In case of the worst outcomes, timeliness is crucial.

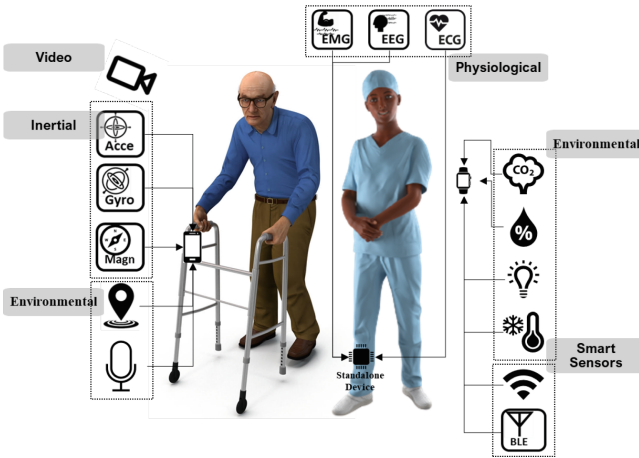
A key challenge was the aim to develop generic processes and components, easily parametrizable, that can be easily reused for other risk stratifications or scoring algorithms. The adopted solution is to abstract parameterizable components and develop generic processes.

The full-stack web-application generated and deployed from the DIME models and Service-Independent Building Blocks (SIBs) also facilitates users to interact with the computational risk stratification algorithm through a web-interface. This reduces the need for skilled healthcare staff to manually apply the algorithm to patient data sets. The DIME development environment lends itself to mapping workflows and process control flows, making development in this manner a utility.



### 3.2 CNN-Based Human Activity Recognition on Edge Computing Devices

Research on HAR involves wearable devices integrating inertial and/or physiological sensors to classify human actions and status across various application domains, such as healthcare, sports, industry, and entertainment [19]. However, executing HAR algorithms on remote devices or the cloud can lead to issues such as latency, bandwidth requirements, and energy consumption. Transitioning towards Edge HAR can be a more effective and versatile solution, overcoming the challenges of traditional HAR techniques. A novel HAR model for computation on edge devices was presented in [55], with a Convolutional Neural Network (CNN) Deep Learning approach, then compared with cloud-computing HAR models.



**Fig. 3.** Full monitoring of a HAR subject with wrist sensors and other monitors (from [19])

Figure 4 describes the adopted HAR processing pipeline, with all the processing steps: Data collection, Noise removal, Segmentation, Feature extraction, Feature selection, Train/Test split, Training HAR model, Deployment on Edge device, Performance exploitation. The CNN model was trained on the Edge Impulse cloud platform by implementing the pre-processing steps of Fig. 4 and deployed on the Thingy 53 device for real-time recognition and evaluation.

The edge computing model achieved promising results ( $\geq 92\%$ ) in terms of Precision, Recall, and F1-score, and with significantly reduced latency and minimal memory. Models with higher computational complexity can be deployed in the cloud but due to network latency and longer inference times they are not suitable for real-time tasks.

As we see, there is here a good outcome for gesture recognition, leveraging CNN-based deep learning, but there is de facto no MDD in the implementation,

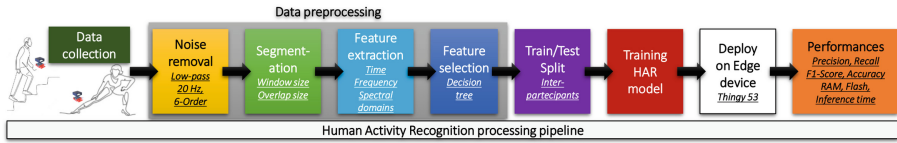


Fig. 4. Detailed HAR system overview. (from [55])

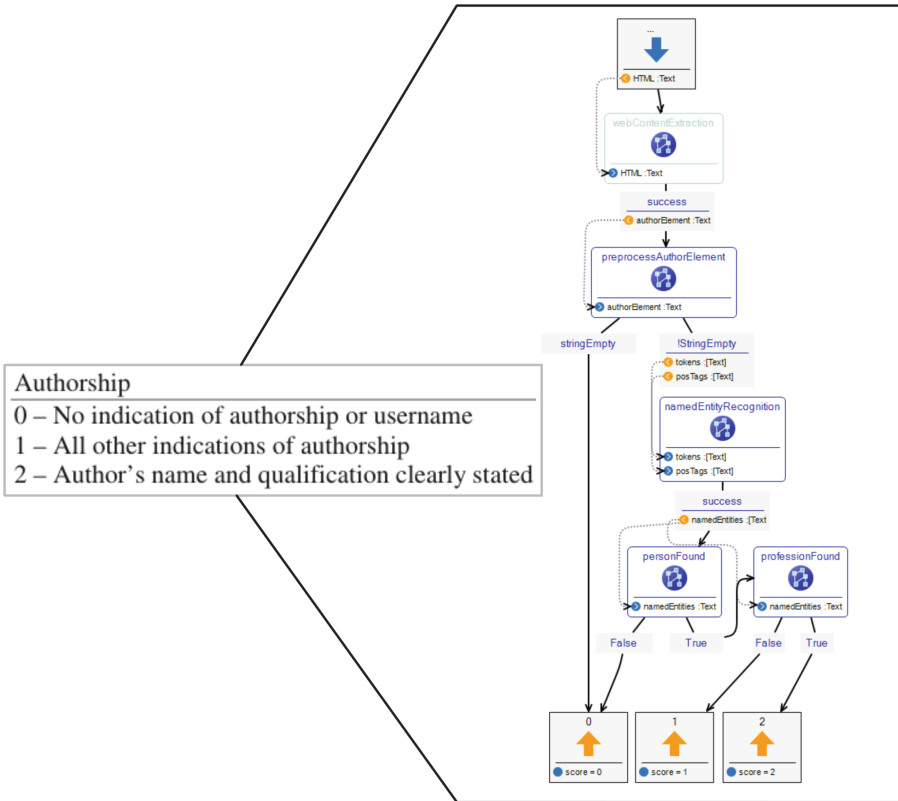
although the phased pipeline (with sub-workflows) of Fig. 4 would lend itself to the technique.

### 3.3 HIPPP: Health Information Portal for Patients and Public

This case study [14] concerns the design and implementation of a public-facing web application that facilitates the automated evaluation and “scoring” of web-based health information. Here, MDD is used to create the web application (similar to the MM case), and to create a specific DSL which enables domain users to use AI pipelines to conduct the end-to-end analysis of web based health information with respect to established quality evaluation frameworks [53] developed by physicians. The AI DSL encompasses a variety of AI models across Natural Language Processing (NLP) and graph processing to perform feature extraction and classification, and ultimately to enable the evaluation process to be automated. The HIPPP system was developed as a web application in DIME. HIPPP utilises a variety of AI algorithms (*transformers, graph neural networks, random forests*) and other computational techniques (*dynamic programming*) to implement an end-to-end classification pipeline that automates the application of the QUEST framework (see Fig. 5). Due to our collaboration with colleagues in the Health Research Institute who work on colon cancer, it is tailored so far to Web-based Health Information (WBHI) pertaining to the diagnosis, treatment and prevention of colon cancer. The reliability and trustworthiness of the sources are subjected to automated evaluation based on the QUEST framework. In the context of the quest for more and more direct patient inclusion and increased quality of information, the goal is for the system to act as a safeguard against patients and the public unknowingly basing their medical decisions on information that they should not trust.

It is foreseen to have an expert in the loop for cases that are not absolutely clear, and to allow adaptive learning based on the manual scores that the expert(s) provide.

The first outcome of this work is the HIPPP application seen as a ready-made fullstack web application which could be deployed by any public health body or advocacy group and be used by patients and the public to evaluate the information they access, providing it is trained on the specific health conditions of interest. The second outcome is the extension of the capabilities of DIME with an A-DSL for AI-based parsing and scoring of document based information sources. It includes several SIB palettes and several processes (for the pipelines



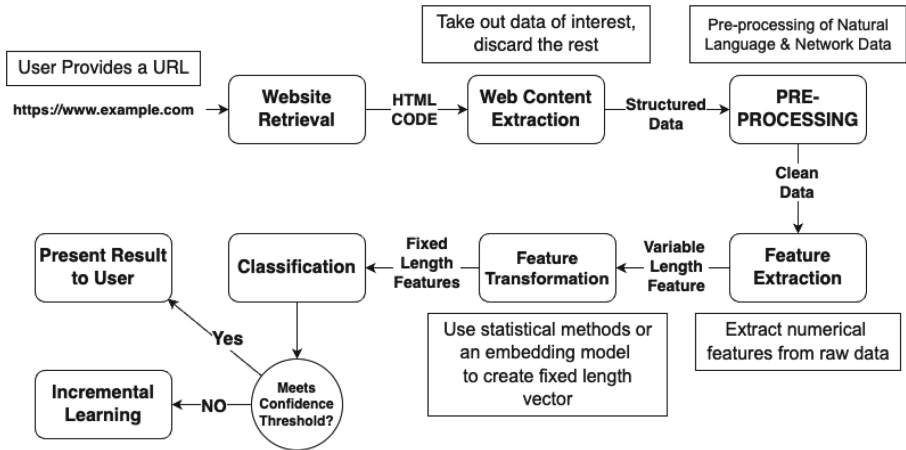
**Fig. 5.** A condensed view which shows translation of the QUEST authorship criteria classifier (image from [53]) translated to a DIME process model (image from [14]). Illustrating how the authorship scoring can be automated, using a combination of web content extraction, named entity recognition and simple rule-based classifiers.

and other algorithms) that are retargetable towards different scoring systems and reusable for other conditions of interest.

**3.4 Cinco De Bio: A Platform for Domain-Specific Workflows that Leverage AI for Biomedical Research**

Cinco de Bio [13] is a platform which enables users to design and execute biomedical analysis workflows using a no-code graphical modelling language. The specific case study concerns cell segmentation carried out on large images of tissue samples, with many layers of information. Figure 7 summarises the process.

The analysis and segmentation software was already available in an implementation in Python, R and Matlab. We optimised some of the algorithms and refactored it to be amenable to becoming SIBs, and redesigned the workflows,



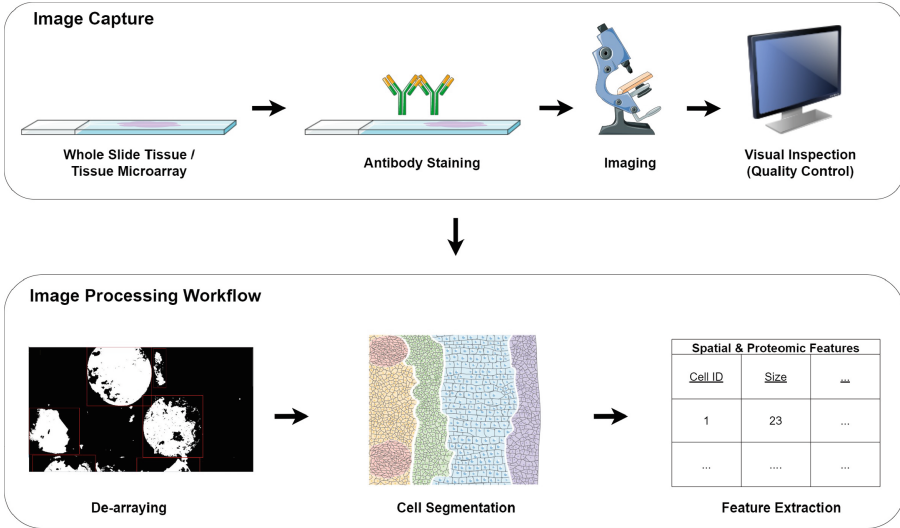
**Fig. 6.** A high-level overview of the HIPPP classification pipeline (image from [14]). With the developed DIME A-DSL consisting of SIBs which fall under one of the workflow steps namely *web content extract*, *pre-processing*, *feature extraction*, *feature transformation*, *classification* and *incremental learning* for graph and natural language data.

while at the same time developing the Cinco de Bio modelling, compilation, deployment and execution environment.

The AI aspect of the project concerns the support in the recognition of single cells, by improving the training capability through artificial image generation. This is currently ongoing work.

The platform comprises a) an Integrated Modeling Environment (IME) for designing and validating workflows built using the Cinco meta modelling tool and b) a workflow execution environment which translates the workflow models to programs that orchestrate the analysis workflows. The execution environment is a Kubernetes-native [15] application that comprises of several custom core services that are loosely coupled. These services handle the execution front-end, model-to-code transformation, computational environment for workflow orchestration program(s), data management, SIB management (which are stored in container-registries (such as Dockherhub)) and a variety of APIs for handling job-scheduling, SIB interactions, etc. The core services are also accompanied by a number of open source services to deliver the overall platform functionality, namely RabbitMQ [64] (message passing), MinIO [2] (Cloud Native Storage), MongoDB [47] (NoSQL database) and Kaniko (building container images in k8s). In addition to the CdB core platform we have developed a SDK for integrating Python and R-based tools into CdB as SIBs, to ease the tool integration process.

Concerning the A-DSL and the modelling style, the SIB and workflow modelling style in CdB is kept quite similar to the SIBs and processes shown in Fig. 2 for the Myeloma diagnosis case study and Fig. 5 for the HIPPP case study. This



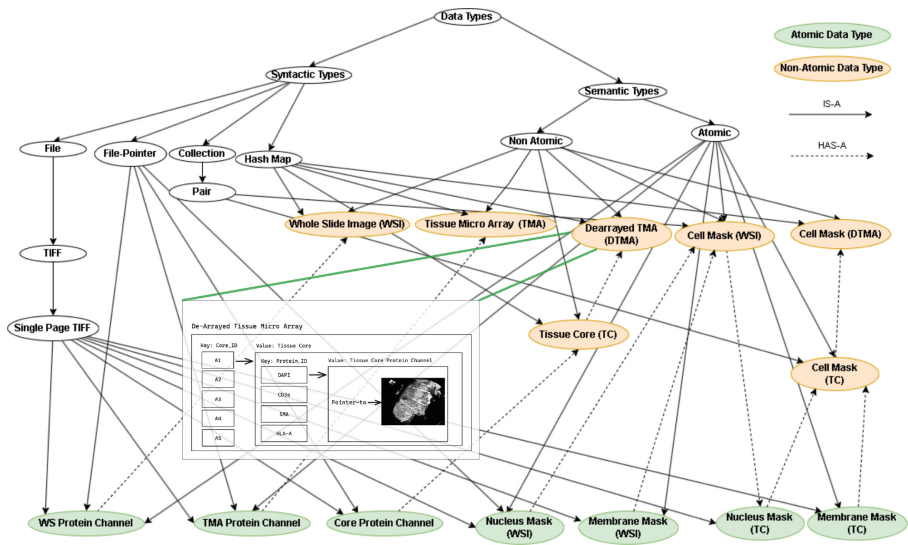
**Fig. 7.** An overview of the image capture and image processing stage for highly-plexed immunofluorescence imaging (Parts of the figure were drawn by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (<https://creativecommons.org/licenses/by/3.0/>)).

is intentional, because the style has proven itself in a wealth of different applications from different domains (smart manufacturing, IoT, AI and reasoning, decision support systems, games, and more), and also because it simplifies supporting the migration of current designs to CdB or to the Cinco-cloud platform that is currently being implemented.

Cinco de Bio uses semantic typing and data/process ontologies to tailor the concepts in its modelling language to the concepts and relations of the specific domain. This enables for example semantic compatibility checking of the data flow, which are in reality all TIFF files. Figure 8 shows part of the data model taxonomy for the highly-plexed tissue image analysis A-DSL.

It also enables the reusability of domain-agnostic components, ensuring at the design stage that domain specific tools such as ML models (trained for a specific task) are being used correctly. The modelling language and the execution platform are designed to support both automated and interactive components: this is necessary because users must verify during analysis the outcome of workflow steps which are non-deterministic (such as for AI models) before proceeding to the following steps of the workflow.

In the context of our motivating use case, pre-processing highly-plexed tissue images for down-stream spatial and proteomic analysis there are a number of different segmentation algorithms developed to solve specific challenges. The first of which is known as de-arraying in the terminology of the field, which



**Fig. 8.** An excerpt of the data model taxonomy for the highly-plexed tissue image analysis A-DSL. It classifies the application domain-specific data types in terms of their syntactic data types (as in computer science data structures and formats) as well as the semantic data types, expressing their meaning in the application domain. We distinguish also atomic and non-atomic data types. An atomic type is a single irreducible piece of data (in the context of the given application domain). For example, a Core Protein Channel equates to a single page tiff (greyscale image) file. A non-atomic type is a data structure which acts as a collection containing atomic or non-atomic types.

essentially entails cropping out each individual tissue core from a tissue micro array slide that may have over 100 cores. De-Arraying is done on a spatially downsampled version of the image and the crop coordinates are then translated back to the full size images. The second form of segmentation is cell-segmentation which is done on the full resolution image (for each cropped core), to acquire the pixel coordinates for the nucleus and membrane of each cell. In both cases convolutional neural networks are typically used for the mask prediction task, however there are an array of post-processing steps required to acquire the final masks (especially in the context of cell segmentation).

As Cinco de Bio is the first case of the new platform architecture, we developed the modelling and execution platform, as well as the process/data ontologies and the concrete SIBS and workflow models. All the data-processing tools are implemented as containerised micro-services. This means that users can chain together heterogeneous data-processing services, that require different computational environments, into a single reproducible workflow.

The main contribution of this case study is the creation of a platform which prevents biologists from having to learn to code (here, Python and R) in order to

undertake analyses on the data generated from their experiments. Secondly the platform is designed for retargeting and extension to any domain, through the creation of new SIBs, workflows, and data and process ontologies which describe that domain.

## 4 Discussion

The four cases chosen for this contribution illustrate four different situations that are typical when advanced software engineering meets AI:

1. **MyMM** embraces MDD for the comfort of having an easily modifiable web application, but it does not do the step towards ML or AI. It would be possible to move to an ML implementation of the risk stratification, for example with decision trees or Algebraic Decision Diagrams (ADDs) in the Addlib [22], or with random forests, and achieve a possibly more performant computation. Here, however, the focus is on recognizability of the algorithm at the process and processing level, which limits what can be done and excludes approaches that would hide the structure of the computation, even if semantically equivalent.
2. **CNN-HAR** embraces ML and the data science lifecycle, but not the MDD: it contains classification and recognition, but is implemented essentially directly in Python. The workflow depicted would lend itself to the transformation to a Pyrus pipeline (if concerning just the data analysis) or a DIME application (if including more interaction with the IoT device and a web application). The limiting factor so far has been the choice of the *Thingy:53*, which is thought as a product for end users and not for integration: its own SDK cannot be accessed in a headless fashion, making the encapsulation into SIBs impossible or very onerous. Other devices like the *Thingy:52* are designed for designers and developers: it is a compact multi-sensor prototyping platform designed to help in building prototypes and demos, and it comes with a bluetooth Application Programming Interface (API) that is amenable to the transformation of commands into collections of SIBs.
3. **HIPPP** combines both the DIME-built web application with a heavy role of AI, ML and more, all encapsulated in reusable SIBs and processes. In this sense, HIPPP is a great example of what both disciplines can deliver when they join forces. HIPPP is a nice application, but being built on/with DIME it still requires local installation of the IME environment (Java, Eclipse etc.), which is impractical for non-programmers who would otherwise not use these tools. It also does not support collaborative modelling, as the models reside in DIME and are not shared.
4. **Cinco de Bio** is the first prototype of the new generation of tools based on Cinco and soon on Cinco-cloud [7]. CdB address several limitations of the DIME application, namely CdB is a cloud native application whereas DIME is not. DIME has limitations with respect to the size/quantity of data it can process, therefore it cannot process data at the scale needed in the use-cases

CdB is designed for. Finally incorporating non-Java based compute environments in DIME applications is a non-trivial task: the integration could be handled with relative ease but the environments would have to be deployed separately and the address hardcoded into DIME.

On the other hand, CdB does not yet support hierarchical process modelling, the IME component is currently Cinco-based and has not been integrated to Cinco-Cloud and the CdB execution environment does not yet support GPU acceleration. These features are going to be addressed in future versions of CdB.

These observations are helping us develop the new R@ISE platform, which will run on Cinco-cloud and is heavily informed by the CdB experience.

In terms of collaboration, this has been a successful case study where we have started with a functioning implementation, partly in Python and partly in R, of the core software, provided by the cell biologists, who have learned Python and R, with the entire processing requiring a number of manual steps to transfer files, start pipelines, interact with the system, which was error prone and time consuming. Our task was to identify the core functionalities of the software, wrap them into individual SIBs, and reconnect them through workflows in an environment that then takes care of compilation and deployment (on heterogeneous runtime stacks, for Python and R), providing a unified web experience that automates all these steps and provides an easier experience as well as ease of reconfiguration and modification of the processes. This seems to be a task along the prior experiences with Bio-jETi [39], Fiat Flux [20], and ci:grasp [5,6]

We had to intervene more than expected on the original code: disentangling it into separate, reusable SIBs brought to a complete refactoring of the code, and in that course we also optimised certain computations improving the readability of the code and the runtime performance.

We actually achieved more than this: the taxonomies are a new addition, the development of the interactive environment that supports human intervention for choices is new, and two new algorithms that address challenges in the application domain has also been devised. The first of which enables the correction of technical variances in samples. The other is a domain-specific data augmentation technique which enables the training of segmentation models on an extremely limited number of training samples. Both algorithms are now the subject of separate publications.

## 5 Conclusion

We have discussed four case studies from our recent research activity with real healthcare stakeholders, that illustrate different quadrants in the bidimensional space of advanced model driven development, in a low-code/no code fashion, and AI/ML.

In general, we have endeavoured to engage with the biology, health and medicine partners on the basis of their needs at that time, and their capabilities



and interests. From previous experience, it would not have been helpful to try to impose a LC/NC approach on the HAR project at that time, because its focus was on the efficient and effective edge computation capability, and not on generalizability. From that interaction, however, a collaboration grew, so that we are now developing further variants of the evaluation pipeline where our technology have found application.

On the contrary, the initial conversations with cell biologists several years ago concerned how to provide their new (AI and ML based) algorithms, that were implemented in Python, to third parties without handing out the Python code, which is an intellectual property embodiment they wished to protect. This led to conversations about interpreted vs. compiled languages: the distinction was new to them. They embraced the usefulness of abstraction (have a second artifact that is different from source code, for example compiled code, with different properties, for example executability and obfuscation) but they did not have the ability nor the funding for a reimplementatation in some compiled language. Years later we reconnected on the tissue analysis case study, where the ability to share the models and even give access to execution to third parties, at the same time hiding the source code was found very attractive and led to the design of Cinco de Bio.

While the healthcare partners were happy with what they got, and sometimes even surprised by what was achieved with the new technologies, it is clear to us that the future lies in the next step beyond Cinco de Bio. In the new R@ISE project, a large Strategic Partnership Project running for 5 years and co-funded by industrial and civil society partners and Science Foundation Ireland, we are currently starting to develop a new platform that will be online and cloud based, support the interaction and co-design directly on the models, support the semantic modelling through taxonomies as well as a strong capability to reason and evaluate the correctness of models at design time. We see here a role for fully automatic, “lightweight” formal methods [57] (static analysis, model checking, some synthesis) [42], but also for the use of AI, ML and Large Language Models (LLMs) in the platform itself, along the lines already recently initiated by partners in Dortmund [16] and [17]. Healthcare will be still a core application domain, in collaboration with the Bernal Institute, the Health Research Institute, the Limerick Digital Cancer Research Centre (LDCRC) and the University Hospital Limerick.

**Acknowledgements.** This work was conducted with the financial support of Science Foundation Ireland (SFI) under grants number 21/SPP/9979 (R@ISE), and 18/CRT/6223 574 (SFI Centre of Research Training in AI) as well as University of Limerick Health Research Institute ULCaN grant Pillar 4.

## References

1. Gartner Forecasts Worldwide Low-Code Development Technologies Market to Grow 20% in 2023

2. Minio inc: Minio — high performance, kubernetes native object storage. <https://min.io/>. Accessed: 21 May 2024
3. Abusetta, Y.: A context aware framework for IoT based healthcare monitoring systems. *Int. J. Adv. Stud. Comput. Sci. Eng.* **9**(7), 1–10 (2020)
4. Al-Antari, M.A., Al-Masni, M.A., Kim, T.S.: Deep learning computer-aided diagnosis for breast lesion in digital mammogram. *Deep Learning in Medical Image Analysis: Challenges and Applications*, pp. 59–72 (2020)
5. Al-Areqi, S., Lamprecht, A.-L., Margaria, T.: Constraints-driven automatic geospatial service composition: workflows for the analysis of sea-level rise impacts. In: Gervasi, O., et al. (eds.) *ICCSA 2016*. LNCS, vol. 9788, pp. 134–150. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-42111-7\\_12](https://doi.org/10.1007/978-3-319-42111-7_12)
6. Al-areqi, S., Lamprecht, A.L., Margaria, T., Kriewald, S., Reusser, D., Wrobel, M.: Agile workflows for climate impact risk assessment based on the ci: grasp platform and the jabc modeling framework (2014)
7. Bainczyk, A., et al.: Cinco cloud: a holistic approach for web-based language-driven engineering. In: Margaria, T., Steffen, B. (eds.) *Leveraging Applications of Formal Methods, Verification and Validation. Software Engineering, ISoLA 2022*, LNCS, vol. 13702, pp. 407–425. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-19756-7\\_23](https://doi.org/10.1007/978-3-031-19756-7_23)
8. Bakera, M., Margaria, T., Renner, C.D., Steffen, B.: Tool-supported enhancement of diagnosis in model-driven verification. *Innov. Syst. Softw. Eng.* **5**(3), 211–228 (2009)
9. Boboyorova, C.: An introduction to no-code/Low-code platforms: seven things you should know
10. Borzì, L., Mazzetta, I., Zampogna, A., Suppa, A., Olmo, G., Irrera, F.: Prediction of freezing of gait in parkinson’s disease using wearables and machine learning. *Sensors* **21**(2), 614 (2021)
11. Borzì, L., Sigcha, L., Rodríguez-Martín, D., Olmo, G.: Real-time detection of freezing of gait in parkinson’s disease using multi-head convolutional neural networks and a single inertial sensor. *Artif. Intell. Med.* **135**, 102459 (2023)
12. Boßelmann, S., et al.: DIME: a programming-less modeling environment for web applications. In: Margaria, T., Steffen, B. (eds.) *ISoLA 2016*. LNCS, vol. 9953, pp. 809–832. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-47169-3\\_60](https://doi.org/10.1007/978-3-319-47169-3_60)
13. Brandon, C., et al.: Cinco de bio: a low-code platform for domain-specific workflows for biomedical research. *BioMedInformatics* **4**(3), 1865–1883 (2024)
14. Brandon, C., Doherty, A.J., Kelly, D., Leddin, D., Margaria, T.: Hipp: health information portal for patients and public. *Appl. Sci.* **13**(16), 9453 (2023)
15. Burns, B., Beda, J., Hightower, K., Evenson, L.: *Kubernetes: up and Running*. O’Reilly Media, Inc., Sebastopol (2022)
16. Busch, D., Bainczyk, A., Steffen, B.: Towards LLM-based system migration in language-driven engineering. In: Kofroň, J., Margaria, T., Seceleanu, C. (eds.) *Engineering of Computer-Based Systems, ECBS 2023*, LNCS, vol. 14390, pp. 191–200. Springer, Cham (2024). [https://doi.org/10.1007/978-3-031-49252-5\\_14](https://doi.org/10.1007/978-3-031-49252-5_14)
17. Busch, D., Nolte, G., Bainczyk, A., Steffen, B.: ChatGPT in the loop: a natural language extension for domain-specific modeling languages. In: Steffen, B. (eds.) *Bridging the Gap Between AI and Reality, AISoLA 2023*, LNCS, vol. 14380, pp. 375–390. Springer, Cham (2024) [https://doi.org/10.1007/978-3-031-46002-9\\_24](https://doi.org/10.1007/978-3-031-46002-9_24)
18. Aguiar Barbosa, A.F.: Model-driven development applied to mobile health and clinical scores. In: Damiani, E., Spanoudakis, G., Maciaszek, L.A. (eds.) *ENASE 2019*. CCIS, vol. 1172, pp. 188–203. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-40223-5\\_9](https://doi.org/10.1007/978-3-030-40223-5_9)

19. Demrozi, F., Pravadelli, G., Bihorac, A., Rashidi, P.: Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. *IEEE Access* **8**, 210816–210836 (2020)
20. Ebert, B.E., Lamprecht, A.L., Steffen, B., Blank, L.M.: Flux-p: automating metabolic flux analysis. *Metabolites* **2**(4), 872–890 (2012)
21. Gleeson, J.P., Brendan Murphy, T., O'Brien, J.D., Friel, N., Bargary, N., O'Sullivan, D.J.: Calibrating covid-19 susceptible-exposed-infected-removed models with time-varying effective contact rates. *Philos. Trans. Royal Soc. A* **380**(2214), 20210120 (2022)
22. Gossen, F., Murtovi, A., Zweihoff, P., Steffen, B.: Add-lib: decision diagrams in practice. arXiv preprint [arXiv:1912.11308](https://arxiv.org/abs/1912.11308) (2019)
23. Greenwald, N.F., et al.: Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat. Biotech.* **40**(4), 555–565 (2022)
24. Guevara, I., Ryan, S., Singh, A., Brandon, C., Margaria, T.: Edge IoT prototyping using model-driven representations: a use case for smart agriculture. *Sensors* **24**(2), 495 (2024)
25. Jeyaraj, P.R., Samuel Nadar, E.R.: Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *J. Can. Res. Clin. Oncol.* **145**(4), 829–837 (2019). <https://doi.org/10.1007/s00432-018-02834-7>
26. Jorge, S., Kubczak, C., Pageau, F., Margaria, T.: model driven design of reliable robot control programs using the jABC. In: Fourth IEEE International Workshop on Engineering of Autonomic and Autonomous Systems (EASe 2007), pp. 137–148, March 2007
27. Jumper, J., et al.: Highly accurate protein structure prediction with alphafold. *Nature* **596**(7873), 583–589 (2021)
28. Kannan, V., Fish, J.C., Willett, D.L.: Agile model driven development of electronic health record-based specialty population registries. In: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 465–468. IEEE (2016)
29. Kelb, P., Margaria, T., Mendler, M., Gsottberger, C.: Mosel: a flexible toolset for monadic second-order logic. In: Brinksma, E. (ed.) TACAS 1997. LNCS, vol. 1217, pp. 183–202. Springer, Heidelberg (1997). <https://doi.org/10.1007/BFb0035388>
30. Kim, J., Lee, J., Park, E., Han, J.: A deep learning model for detecting mental illness from user content on social media. *Sci. Rep.* **10**(1), 11846 (2020)
31. Kinkead, L., Allam, A., Krauthammer, M.: Autodiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. *BMC Med. Inform. Decis. Mak.* **20**(1), 1–13 (2020)
32. Kluyver, T., et al.: Jupyter notebooks—a publishing format for reproducible computational workflows. *Elpub* **2016**, 87–90 (2016)
33. Kripke, S.A.: Semantical analysis of modal logic i normal modal propositional calculi. *Math. Log. Q.* **9**(5–6), 67–96 (1963)
34. Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D.: An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE J. Biomed. Health Inform.* **21**(1), 31–40 (2016)
35. Lahboube, F., Souissi, O.R.N.: Building a his supervision metamodel. In: 2016 11th System of Systems Engineering Conference (SoSE), pp. 1–6. IEEE (2016)
36. Lamprecht, A.-L., Margaria, T., Steffen, B.: Data-flow analysis as model checking within the jABC. In: Mycroft, A., Zeller, A. (eds.) CC 2006. LNCS, vol. 3923, pp. 101–104. Springer, Heidelberg (2006). [https://doi.org/10.1007/11688839\\_9](https://doi.org/10.1007/11688839_9)

37. Lee, Michael, Y., et al.: Cellseg: a robust, pre-trained nucleus segmentation and pixel quantification software for highly multiplexed fluorescence images. *BMC Bioinf.* **23**(1), 46 (2022)
38. Margaria, T.: Making sense of complex applications: constructive design, features, and questions. In: Margaria, T., Graf, S., Larsen, K.G. (eds.) *Models, Mindsets, Meta: The What, the How, and the Why Not?* LNCS, vol. 11200, pp. 129–148. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-22348-9\\_9](https://doi.org/10.1007/978-3-030-22348-9_9)
39. Margaria, T., Kubczak, C., Steffen, B.: Bio-jeti: a service integration, design, and provisioning platform for orchestrated bioinformatics processes. *BMC Bioinf.* **9**, 1–17 (2008)
40. Margaria, T., Meyer, D., Kubczak, C., Isberner, M., Steffen, B.: Synthesizing semantic web service compositions with jMosel and Golog. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) *ISWC 2009*. LNCS, vol. 5823, pp. 392–407. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04930-9\\_25](https://doi.org/10.1007/978-3-642-04930-9_25)
41. Margaria, T., Steffen, B.: Backtracking-free design planning by automatic synthesis in metaframe. In: Astesiano, E. (ed.) *FASE 1998*. LNCS, vol. 1382, pp. 188–204. Springer, Heidelberg (1998). <https://doi.org/10.1007/BFb0053591>
42. Margaria, T., Steffen, B.: Backtracking-free design planning by automatic synthesis in metaframe. In: *International Conference on Fundamental Approaches to Software Engineering*, pp. 188–204. Springer, Cham (1998)
43. Margaria, T., Steffen, B.: Business process modelling in the jABC: the one-thing-approach. In: Cardoso, J., van der Aalst, W. (eds.), *Handbook of Research on Business Process Modeling*, IGI Global (2009)
44. Margaria, T., Steffen, B.: Service-orientation: conquering complexity with XMDD. In: Hinchey, M., Coyle, L. (eds.) *Conquering Complexity*, pp. 217–236. Springer, London (2012). [https://doi.org/10.1007/978-1-4471-2297-5\\_10](https://doi.org/10.1007/978-1-4471-2297-5_10)
45. Margaria, T., Steffen, B.: eXtreme model-driven development (XMDD) technologies as a hands-on approach to software development WithoutCoding. In: *Encyclopedia of Education and Information Technologies*, pp. 732–750. Springer, Cham (2020)
46. Mohan, S., Thirumalai, C., Srivastava, G.: Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* **7**, 81542–81554 (2019)
47. MongoDB, Inc. *Mongodb: the developer data platform*. <https://www.mongodb.com> (2024). Accessed 21 May 2024
48. Murarka, A., Radhakrishnan, B., Ravichandran, S.: Classification of mental illnesses on social media using roberta. In: *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pp. 59–68 (2021)
49. Nan, S., et al.: Dccss: a meta-model for dynamic clinical checklist support systems. In: *2015 3rd International Conference on Model-Driven Engineering and Software Development (MODELSWARD)*, pp. 272–279. IEEE (2015)
50. Olivero, M.A., Domínguez-Mayo, F.J., Parra-Calderón, C.L., Escalona, M.J., Martínez-García, A.: Facilitating the design of hl7 domain models through a model-driven solution. *BMC Med. Inf. Decis. Making* **20**, 1–18 (2020)
51. Blake, O., Clifford, R., Reche, A.: Uhl guidance on management of mgus in primary care (2019)
52. Blake, O., Clifford, R., Reche, A.: Uhl guidance on management of mgus in primary care. Technical report, University Hospital Limerick (2019)
53. Robillard, J.M., Jun, J.H., Lai, J.A., et al.: The QUEST for quality online health information: validation of a short quantitative tool. *BMC Med. Inform. Decis. Mak.* **18**, 87 (2018). <https://doi.org/10.1186/s12911-018-0668-9>

54. Sadilek, A., et al.: Deploying nemesis: preventing foodborne illness by data mining social media. *AI Mag.* **38**(1), 37–48 (2017)
55. Singh, A., Margaria, T., Demrozi, F.: CNN-based human activity recognition on edge computing devices. In: 2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS), pp. 1–4 (2023)
56. Steffen, B., Claßen, A., Klein, M., Knoop, J., Margaria, T.: The fixpoint-analysis machine. In: Lee, I., Smolka, S.A. (eds.) *CONCUR 1995*. LNCS, vol. 962, pp. 72–87. Springer, Heidelberg (1995). [https://doi.org/10.1007/3-540-60218-6\\_6](https://doi.org/10.1007/3-540-60218-6_6)
57. Steffen, B., Margaria, T.: Tools get formal methods into practice. *ACM Comput. Surv. (CSUR)* **28**(4es), 126–es (1996)
58. Tao, D., Ruofan, H., Zhang, D., Laber, J., Lapsley, A., Kwan, T., Rathke, L., Rundensteiner, E., Feng, H.: A novel foodborne illness detection and web application tool based on social media. *Foods* **12**(14), 2769 (2023)
59. Tao, D., Zhang, D., Ruofan, H., Rundensteiner, E., Feng, H.: Crowdsourcing and machine learning approaches for extracting entities indicating potential foodborne outbreaks from social media. *Sci. Rep.* **11**(1), 21678 (2021)
60. Venčkauskas, A., Štūkys, V., Toldinas, J., Jusas, N.: A model-driven framework to develop personalized health monitoring. *Symmetry* **8**(7), 65 (2016)
61. Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257* (2023)
62. Wang, Y.T., Huang, H.H., Chen, H.H., Chen, H.: A neural network approach to early risk detection of depression and anorexia on social media text. In: *CLEF (Working Notes)*, pp. 1–8 (2018)
63. Wen, P.Y., Chang, S.Y.: Design and implementation of model-driven development for nursing information system. In: *MEDINFO 2021: One World, One Health-Global Partnership for Digital Innovation: Proceedings of the 18th World Congress on Medical and Health Informatics*, vol. 290, p. 154. IOS Press (2022)
64. Williams, J.: *RabbitMQ in action: distributed messaging for everyone*. Simon and Schuster (2012)
65. Yadav, S.S., Jadhav, S.M.: Deep convolutional neural network based medical image classification for disease diagnosis. *J. Big Data* **6**(1), 1–18 (2019)
66. Yahyaoui, A., Jamil, A., Rasheed, J., Yesiltepe, M.: A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International informatics and software engineering conference (UBMYK), pp. 1–4. IEEE (2019)
67. Zang, Q., et al.: In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *J. Chem. Inf. Model.* **57**(1), 36–49 (2017)
68. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: a semantically and visually interpretable medical image diagnosis network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 6428–6436 (2017)
69. Zweihoff, P., Steffen, B.: Pyrus: an online modeling environment for no-code data-analytics service composition. In: Margaria, T., Steffen, B. (eds.) *ISoLA 2021*. LNCS, vol. 13036, pp. 18–40. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-89159-6\\_2](https://doi.org/10.1007/978-3-030-89159-6_2)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# **Responsible and Trustworthy AI**



# Balancing Transparency and Risk: An Overview of the Security and Privacy Risks of Open-Source Machine Learning Models

Dominik Hintersdorf<sup>1,2</sup> , Lukas Struppek<sup>1,2</sup> ,  
and Kristian Kersting<sup>1,2,3,4</sup> 

<sup>1</sup> Technical University of Darmstadt, Darmstadt, Germany

{[hintersdorf](mailto:hintersdorf@cs.tu-darmstadt.de), [struppek](mailto:struppek@cs.tu-darmstadt.de), [kersting](mailto:kersting@cs.tu-darmstadt.de)}@cs.tu-darmstadt.de

<sup>2</sup> German Research Center for Artificial Intelligence (DFKI), Darmstadt, Germany

<sup>3</sup> Centre for Cognitive Science of TU Darmstadt, Darmstadt, Germany

<sup>4</sup> Hessian Center for AI (hessian.AI), Darmstadt, Germany

**Abstract.** The field of artificial intelligence (AI) has experienced remarkable progress in recent years, driven by the widespread adoption of open-source machine learning models in both research and industry. Considering the resource-intensive nature of training on vast datasets, many applications opt for models that have already been trained. Hence, a small number of key players undertake the responsibility of training and publicly releasing large pre-trained models, providing a crucial foundation for a wide range of applications. However, the adoption of these open-source models carries inherent privacy and security risks that are often overlooked. To provide a concrete example, an inconspicuous model may conceal hidden functionalities that, when triggered by specific input patterns, can manipulate the behavior of the system, such as instructing self-driving cars to ignore the presence of other vehicles. The implications of successful privacy and security attacks encompass a broad spectrum, ranging from relatively minor damage like service interruptions to highly alarming scenarios, including physical harm or the exposure of sensitive user data. In this work, we present a comprehensive overview of common privacy and security threats associated with the use of open-source models. By raising awareness of these dangers, we strive to promote the responsible and secure use of AI systems.

**Keywords:** Machine Learning · Security · Privacy · Open-Source · Overview

## 1 Introduction

With the increase in computing capability, big models are trained on a huge amount of data, often scraped from the public internet. However, while this is often done closed-source, some are developing open-source models that are often

---

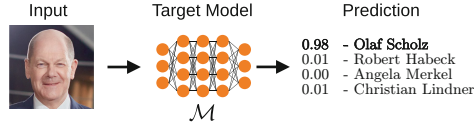
D. Hintersdorf and L. Struppek—Equal contribution.

© The Author(s) 2025

B. Steffen (Ed.): AISO LA 2023, LNCS 14129, pp. 269–283, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_16](https://doi.org/10.1007/978-3-031-73741-1_16)





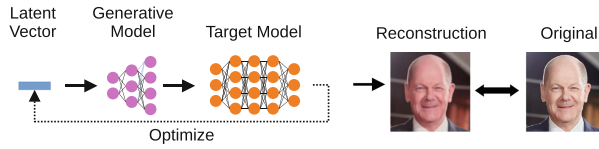
**Fig. 1.** A basic deep neural network designed for facial recognition, capable of predicting corresponding identities, e.g., the German Chancellor Olaf Scholz. Given a specific input, the model computes a prediction vector, assigning probabilities to each distinct class. The final prediction is determined by the class with the highest probability. This model serves as an example for the attacks we discuss.

used as a basis for downstream tasks. For example, the popular text-to-image model *Stable Diffusion* uses the pre-trained text encoder from CLIP [43], a pre-trained multi-modal model, to process input texts.

While some large-scale models are completely closed-source, such as OpenAI’s GPT-4 [38] or Google’s Gemini [21], and are only accessible through an API, many other models are available as open-source models, usually including the code to train the model and the parameters of already trained models. Examples of such open-source models are BLOOM [48], OpenLLaMA [18], LLaMA [59], LLaMA 2 [17], OpenCLIP [28] and Stable Diffusion [44]. A group of companies and institutions, including GitHub, Hugging Face, Creative Commons, and others, are calling for more open-source support in the forthcoming EU AI Act [14]. While most open-source available models are trained on public data from the internet, information about which exact data was used is not always made public. Still, these models are deployed in numerous applications and settings.

However, not only these big models are made publicly available. Sites like Hugging Face, TensorFlow Hub, or PyTorch Hub allow users to provide and exchange model weights trained by the community, which are publicly available for everyone to download. While this practice has clearly its upsides, the trustworthiness of such pre-trained open-source models comes increasingly into focus. Since the model architecture, weights, and training procedure are publicly known, malicious adversaries have an advantage when trying to attack these models compared to settings with models kept behind closed doors. Whereas all attacks presented in this work are also possible to some extent without full model access and less knowledge about the specific architecture, they become inherently more difficult to perform without such information.

Trustworthy machine learning comprises various areas, including security, safety, and privacy. It is important to distinguish clearly between these three areas of trustworthiness. *Safety* describes the robustness against model malfunctions without malicious external influences. For example, a safe autonomous car provides reliable driving and transports people unharmed, independent of environmental conditions like weather. *Security*, on the other hand, describes a model’s resilience against intentional attacks from malicious parties. For instance, an attacker could modify street signs to trigger a critical system behav-



**Fig. 2.** Model inversion attacks aim to craft samples that reveal sensitive information about the training data, such as revealing a person’s identity, in this case, Olaf Scholz. The adversary usually employs a generative model, capable of producing synthetic images from a latent input vector. This latent vector is then optimized using the target model as guidance, with the objective of maximizing the confidence for a specific class.

ior of the car and force a car crash. The aspect of *privacy* relates to the access to private information about the models and their training data. Privacy-preserving models should not disclose sensitive information from the training process to other users and attackers.

This work will give an overview of common privacy and security threats associated with using open-source models. The paper will use a simple face classification model (see Fig. 1) as an example case. In Sects. 2 and 3, we will go over the most prominent privacy and security attacks in accordance with the German Federal Office for Information Security [3]. Then, we will discuss the advantages and disadvantages of open-source practices in machine learning in Sect. 4, followed by a conclusion in Sect. 5.

## 2 Privacy Attacks on Open-Source Models

In this section, we will go over the two most common privacy attacks, namely *model inversion attacks* (Sects. 2.1 and 2.2) and *membership inference attacks* (Sect. 2.3), and demonstrate how publicly releasing the model weights might harm user privacy. At the same time, these attacks might also act as a tool to prevent unauthorized data usage. In the following, we will discuss both of these aspects of privacy attacks with regard to open-source models.

### 2.1 Model Inversion Attacks

Model inversion and reconstruction attacks have the goal of extracting sensitive information about the training data of an already trained model, e.g., by reconstructing images disclosing sensitive attributes [12, 15, 52, 53, 55, 61, 66] or generating text with private information contained in the training data [9, 40]. Figure 2 provides a simple example of a successful inversion attack.

For model inversion attacks, it is often assumed that the attacker has full access to the target model and its parameters and some generative model to generate samples from the training data domain. Generative models, in this case usually GANs [19, 31], can synthesize high-quality images from randomly sampled vectors, the so-called latent vectors. The generative model then acts as a

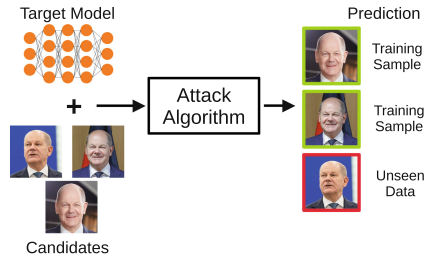
prior, to guide the optimization process and to generate images revealing sensitive features from the training data. Typically, the target model’s output score of a specific class is maximized through an optimization process in which the latent vector of the GAN is altered. Although model inversion attacks are often applied to classification models, by altering the loss function of the optimization process, these attacks could also be applied to other model classes and tasks such as image segmentation [58] and sentence embeddings [33]. As an attacker has full access to the open-source models, model inversion attacks are a genuine threat to privacy. Imagine an open-source model trained to classify facial features like hair or eye color. An adversary successfully performing a model inversion attack could then generate synthetic facial images that reveal the identity of individuals from the training data.

## 2.2 Information Leakage by Memorization

Closely related to model inversion attacks are data leaks through unintended memorization. The distinction is in the adversary’s intent: in a model inversion attack, the adversary actively seeks to reconstruct model inputs, while leakage by memorization can occur incidentally, especially when interacting with generative models. These generative models encompass vast language models like the LLaMA family [17, 59], along with image generation models like Stable Diffusion [44]. Generative language models, for instance, predict subsequent words when given an input text. For example, with the input sentence “the capital city of France is,” a model might confidently predict “Paris.” However, unintended leakage can happen when the model generates text revealing private information from its training data that should not be disclosed in its prediction. For instance, the model might inadvertently complete the query “My social security number is” with a real social security number from the model’s training data.

Since recent language models are trained on vast amounts of data scraped from various sources across the internet, it is highly probable that some private information will accidentally become part of the model’s training data. This highlights the importance of addressing and mitigating the risk of unintended data leakage, especially when dealing with generative models with access to potentially sensitive information. In addition to accidental occurrences of memory leakage, there is also a concern that malicious users could deliberately craft queries that facilitate this kind of leakage [7, 36]. This risk applies to open-source generative language models like LLaMA and non-public models that offer only API access, potentially compromising individuals’ privacy by generating texts containing sensitive information.

Likewise, similar concerns extend to image synthesis models, which have been found to reconstruct samples from their training data [4, 6, 51]. Such capabilities could potentially lead to legal issues if the generated content is under copyright protection. The New York Times recently started a lawsuit against OpenAI, the creator of ChatGPT, about copyright infringement since their news articles have been used for training without their consent [22]. To address these challenges, it is crucial to implement robust privacy measures and security mecha-



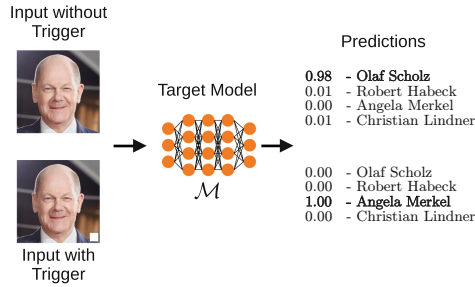
**Fig. 3.** Membership Inference Attacks seek to determine whether a specific sample was part of a model’s training data. These attacks commonly exploit that models tend to behave differently on inputs they have been trained on than unseen inputs.

nisms in both language and image synthesis models, safeguarding against unintended data leakage and potential misuse of generated content. Proactive steps should be taken to mitigate the risks of accidental and malicious attempts to exploit model vulnerabilities.

### 2.3 Membership Inference Attacks

While inversion and data leakage attacks try to infer information about the training data by reconstructing parts of it, membership inference attacks [5, 13, 25, 34, 47, 50, 64], as another type of privacy attack, try to infer which data samples out of a pool of candidates have been used for training a model. Figure 3 illustrates a simple example. In this scenario, the attacker has some data samples and wants to check whether this data was used to train a particular model. We will give a short example to demonstrate why such a successful attack seriously threatens privacy. Imagine that a hospital is training a machine learning model on the medical data of hospital patients to predict whether future patients have cancer. An attacker gains access to the model and has a set of private data samples. The adversary tries to infer whether the data of a person was used for training the cancer prediction model. If the attack is successful, the attacker knows not only that the person had or has cancer, but also was once a patient in that hospital. In the traditional setting of membership inference attacks, the attacker is interested in predicting whether a specific sample was present in the training data, i.e., a particular image or text. Related recent work, such as from Hintersdorf et al. [24] or Li et al. [32], tries to infer if some data of a person was used for training without focussing on a particular data sample.

Having full access to an open-source model makes membership inference attacks more feasible in comparison to models kept behind APIs. This is because the attacker can observe all activations and outputs for every input, making it easier to infer membership. As a result, open-source models can leak sensitive information about the data used for training. More importantly, this information about the training data is permanently encoded in the model weights. If private information is deleted from public websites, it is usually not publicly accessible



**Fig. 4.** Backdoor attacks involve injecting a limited number of poisoned samples into a model’s training data, aiming to inject a hidden model functionality, such as always predicting a specific class. This hidden behavior can be activated during inference by inputs containing a pre-defined trigger, as illustrated in this example by a white square.

anymore. However, if the model has been trained on this data, it still contains information about the data and can leak it to malicious users.

### 2.4 Privacy Attacks to Enforce Rights

Until now, we have only presented possible negative impacts of privacy attacks. However, there is also a positive side to open-source models being susceptible to these attacks. While these privacy attacks can leak possibly sensitive information to an attacker, they can also be used to prove unauthorized data access. As a result, these attacks can be used to enforce privacy and copyright laws [24]. Take, for example, the lawsuit of the stock image supplier Getty Images against Stability AI over copyright infringement. Getty Images accuses Stability AI of unlawfully using stock images for training their text-to-image model without having acquired a license to use the images [30,60]. Privacy attacks like model inversion, membership inference, or memorization leakage attacks could be one way to prove that these images were illegally used for training. Another example is that users can apply these privacy attacks to prove that a company has trained a model on their potentially private data without permission, as shown by Hintersdorf et al. [24]. Combined with techniques to delete specific knowledge from the models [16,26,65] or machine unlearning [1], these attacks offer a way to enforce the protection of user privacy.

## 3 Security Attacks on Open-Source Models

In this section, we show common security attacks against machine learning models. We will showcase two of the most prominent attack types, namely *backdoor attacks* (Sect. 3.1) and *adversarial examples* (Sect. 3.2).

### 3.1 Data Poisoning and Backdoor Attacks

Open-source models undergo training on vast datasets, often comprising millions or even billions of data samples. Due to this massive scale, human data inspection is not feasible in any way, necessitating a reliance on the integrity of these datasets. However, previous research has revealed that adding a small set of manipulated data to a model's training data can significantly influence its behavior. This dataset manipulation is referred to as *data poisoning* and for numerous applications, manipulating less than 10% of the available data is sufficient to make the model learn some additional, hidden functionalities.

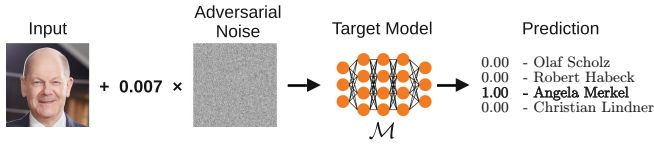
Such hidden functionalities are called *backdoors* [23,45] and they are activated when a model input includes a specific trigger pattern. Figure 4 demonstrates a practical backdoor attack. For instance, in the case of image classification, trigger patterns may involve certain color patterns placed in the corner of an image, e.g., a checkerboard pattern. A common backdoor strategy involves adding a small set of samples into the training data containing the trigger pattern and a target label from a particular class. During training, the model learns to associate the trigger with the specified target class, thereby predicting the target class for each input that contains the trigger. At the same time, the model's performance on clean inputs should not degrade noticeably to ensure the attack's stealthiness.

Detecting this type of model manipulation is challenging for users since the models appear to function as expected on clean inputs. However, when the hidden backdoor function is activated, the model behaves as the attacker intended. Take text-to-image synthesis models as an example, renowned for their ability to generate high-quality images based on user textual descriptions. Struppek et al. [54] have recently shown that small manipulations to the model are sufficient to inject multiple backdoor functionalities that can be triggered by single characters or words. Once activated, these backdoors might force the generation of harmful or offensive content, posing serious risks to users. Depending on an individual's background, exposure to such content could cause mental harm and distress. Another example is that facial recognition security systems could be compromised if backdoored models are used.

Backdoor and poisoning attacks have become prevalent across various domains, for example, image classification, self-supervised learning [8,46], transfer learning [63], graph neural networks [62,67] and federated learning [49,68]. Various approaches exist to detect poisoned samples in the training data or triggers in the inputs. However, it is unclear if the training data of open-source models has been checked for poisoned data samples with existing approaches. Even if inspections were conducted, providing guarantees that publicly available models are devoid of hidden backdoors remains challenging. The complexity and diversity of attacks make it difficult to ensure complete protection.

### 3.2 Adversarial Examples

In addition to poisoning attacks that usually manipulate the training process to introduce hidden backdoor functions, another category of security attacks tar-



**Fig. 5.** Adversarial examples are crafted by adding a small fine-tuned noise to the input, resulting in misleading model predictions. This noise is computed to alter the trained model’s prediction in a specific manner. In many cases, the changes to the input are barely perceptible to humans, making it challenging to detect these manipulations.

gets models solely during inference. Known as *adversarial examples* or *evasion attacks* [20], the attacks slightly modify model inputs with the intention of altering the model’s behavior. Consequently, these crafted adversarial examples can be employed to bypass a model’s detection and cause misclassifications. Figure 5 illustrates a simple adversarial example. Among various security research subjects, adversarial examples stand out as the most extensively studied class of attacks, with several thousand papers delving into this topic.

In computer vision tasks, the attacker computes a unique noise pattern tailored to a specific input, which is then added to the image to disrupt the model’s prediction. Remarkably, even minor changes in the input, hardly noticeable to the human eye, can drastically impact the model’s behavior. Numerous discussions have arisen concerning why deep learning architectures and other algorithms are susceptible to such subtle input changes. One plausible explanation lies in the models’ dependence on non-robust input features that might not appear informative from a human standpoint. However, these features can be exploited during training to solve the specific training task effectively [29].

In practice, adversarial examples are hard to detect by the human eye, rendering them especially dangerous in safety-critical applications. For instance, automatic content detection systems may be susceptible to evasion by images containing adversarial perturbations. This vulnerability extends to critical applications such as detecting child sexual abuse material [56] or identifying deepfakes [27]. The potential consequences of such undetected adversarial inputs emphasize the need to develop robust defenses against these attacks to ensure the integrity and reliability of machine learning systems.

Numerous approaches [20, 37] to crafting adversarial examples leverage white-box model access, allowing them to compute gradients on the model concerning the current input. This enables the attacker to optimize the adversarial noise using standard gradient descent approaches. However, even with restricted access to a model’s prediction vector [11, 39, 57] or only the predicted label [2, 10], various attack approaches still exist. The fact that open-source model weights and architectures are publicly available poses a risk, as adversaries can exploit the model locally and then use the crafted adversarial examples to deceive the targeted model. This highlights the importance of robust defense mechanisms to safeguard against such adversarial attacks, especially in scenarios when dealing with publicly accessible models.

## 4 Discussion

While we have shown that publishing models as open-source have clear disadvantages, there are also upsides to making models publicly available for everyone. In the following, we provide a discussion on both perspectives regarding the privacy and the security of models:

- ➖ **Data Privacy Concerns:** Models trained on large datasets might inadvertently contain sensitive information, like personally identifiable information, medical data, or other sensitive details, posing privacy risks if not handled carefully. The models may memorize or encode this information into their parameters during training. This can pose serious privacy risks when models are deployed in real-world applications. Samples from the training data could potentially be extracted through methods like model inversion attacks, allowing attackers to infer sensitive details about individuals whose data was used for training.
- ➖ **Vulnerability Exposure:** Since open-source models are accessible to everyone, including malicious actors, vulnerabilities can be more easily exposed, potentially leading to strong attacks. Open-source models might become primary targets for adversarial attacks and evasion attacks. Malicious actors can study the model's architecture, parameters, and training data to develop sophisticated attacks to manipulate or compromise the model's behavior.
- ➖ **Lack of Regulatory Compliance and License Issues:** Depending on the context of use, certain industries and applications might require compliance with specific security and privacy regulations. Using open-source models may complicate compliance efforts, especially if the model is not designed with these regulations in mind. Depending on the open-source license, some models may require users to disclose their modifications or share derived works, which could raise concerns about proprietary information. To what extent generative models can commit copyright infringement is also an open question. Since parts of the training data may underlay copyright regulations, the generated data might also incorporate parts of it and fall under copyright law.
- ➖ **Zero-Day Vulnerabilities:** Open-source models can be susceptible to poisoning and backdoor attacks, where adversarial actors inject malicious data into the training set to manipulate the model's behavior. Many open-source models are published without their training data available. This makes it hard to check the integrity of the data and avoid model tampering. In practice, injected backdoors are hard to detect and may stay hidden until activated by a pre-defined trigger.
- ➕ **Transparency and Auditability:** Open-source models allow users to examine the source code, algorithms, and sometimes even the data used to build the model. This transparency helps in understanding how the model works and detecting potential vulnerabilities. This process is called *red-teaming* and is usually done by teams of publishing companies such as OpenAI, Meta, or Google. In the case of open-source models, the community can do this process of finding and disclosing vulnerabilities in a much more open and transparent way.



**+ Community and Research Collaborations:** Open-source models encourage collaboration among researchers and developers. The community can work together to identify and fix security and privacy issues promptly. Furthermore, with access to novel models and architectures, existing attack and defense mechanisms can be investigated in this setting, allowing adaptation and adjustments to new situations.

**+ Customization and Adaptation:** With access to the source code, developers can customize and adapt the model to suit their specific needs, ensuring it aligns with their security and privacy requirements. Since the available models are already trained, fewer data is required to adjust a model to a novel task or setting. In turn, fewer privacy concerns are expected from the fine-tuning dataset.

**+ Quality and Peer Review:** Popular open-source models often go through rigorous peer review, enhancing their overall quality and reducing the chances of major security or privacy flaws. It also includes investigations of independent research groups, offering new perspectives and insights.

**+ Faster Development and Innovation:** Building on top of existing open-source models can significantly speed up development efforts, enabling rapid innovation and research. This also includes the investigation of potential security vulnerabilities and corresponding defense and mitigation mechanisms.

Despite these difficulties, open-source machine learning models remain an important resource for the AI community. Risks can be reduced by implementing best practices for model usage, performing security audits, and encouraging community cooperation to solve security and privacy issues proactively. Additionally, promoting responsible vulnerability disclosure can assist in preserving the security and dependability of open-source projects.

## 5 Conclusion

In conclusion, we have highlighted and discussed open-source models' security and privacy vulnerabilities, which are expected to have a greater risk than closed-source models. Public access to model weights can significantly facilitate privacy attacks like inversion or membership inference, particularly when the training set remains private. Similarly, security attacks aimed at compromising model robustness can be executed by manipulating the training data to introduce hidden backdoor functionalities or crafting adversarial examples to manipulate inference outcomes. These risks impact the published model itself and extend to applications and systems that incorporate this model.

Despite these identified risks, it is important to acknowledge the numerous advantages that open-source machine learning offers. The practice of publishing models, source code, and potentially even data can support widespread adoption, foster transparency, and encourage innovation. We recognize the need for users and publishers to be aware of the inherent risks associated with open-source practices. However, particularly in the case of publishing large models, such as large language and text-to-image synthesis models, we firmly believe that the

benefits outweigh the drawbacks. As such, we encourage developers to continue embracing open-source approaches, thereby promoting transparency, driving further research, and fostering innovation in the field of machine learning. As we have shown there are trade-offs between the transparency and the privacy and security of open-source models. Therefore, we strongly believe that this matter should be further explored interdisciplinary.

**Acknowledgments.** The authors thank Daniel Neider for the fruitful discussions. This work was supported by the German Ministry of Education and Research (BMBF) within the framework program “Research for Civil Security” of the German Federal Government, project KISTRA (reference no. 13N15343).

**Image Sources.** Images depicting Olaf Scholz were provided by Bernhard Ludewig [35], Palácio do Planalto [42] and European Parliament [41].

## References

1. Bourtole, L., et al.: Machine unlearning. In: Symposium on Security and Privacy (S&P), pp. 141–159 (2021)
2. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: reliable attacks against black-box machine learning models. In: International Conference on Learning Representations (ICLR) (2018)
3. Bundesamt für Sicherheit in der Informationstechnik: AI Security Concerns in a Nutshell (2023). [https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical\\_AI-Security\\_Guide.2023.pdf](https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide.2023.pdf). Accessed 01 May 2024
4. van den Burg, G.J.J., Williams, C.: On memorization in probabilistic deep generative models. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 27916–27928 (2021)
5. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership Inference Attacks From First Principles. In: Symposium on Security and Privacy (S&P), pp. 1897–1914 (2022)
6. Carlini, N., et al.: Extracting training data from diffusion models. In: USENIX Security Symposium, pp. 5253–5270 (2023)
7. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: evaluating and testing unintended memorization in neural networks. In: USENIX Security Symposium, pp. 267–284 (2019)
8. Carlini, N., Terzis, A.: Poisoning and Backdooring Contrastive Learning. In: International Conference on Learning Representations (ICLR) (2022)
9. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T.B., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting Training Data from Large Language Models. In: USENIX Security Symposium, pp. 2633–2650 (2021)
10. Chen, J., Jordan, M.I., Wainwright, M.J.: HopSkipJumpAttack: a query-efficient decision-based attack. In: Symposium on Security and Privacy (S&P), pp. 1277–1294 (2020)
11. Chen, P., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.: ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: ACM Workshop on Artificial Intelligence and Security (AIsec@CCS), pp. 15–26 (2017)

12. Chen, S., Kahla, M., Jia, R., Qi, G.: Knowledge-enriched distributional model inversion attacks. In: International Conference on Computer Vision (ICCV), pp. 16158–16167 (2021)
13. Choquette-Choo, C.A., Tramèr, F., Carlini, N., Papernot, N.: Label-Only Membership Inference Attacks. In: International Conference on Machine Learning (ICML), pp. 1964–1974 (2021)
14. David, E.: GitHub and others call for more open-source support in EU AI law (2023). <https://www.theverge.com/2023/7/26/23807218/github-ai-open-source-creative-commons-hugging-face-eu-regulations>. Accessed 27 July 2023
15. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
16. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing Concepts from Diffusion Models (2023). arXiv preprint: [arXiv:2303.07345](https://arxiv.org/abs/2303.07345)
17. GenAI, M.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023). arXiv preprint: [arXiv:2307.09288](https://arxiv.org/abs/2307.09288)
18. Geng, X., Liu, H.: OpenLLaMA: An Open Reproduction of LLaMA (2023). [https://github.com/openlm-research/open\\_llama](https://github.com/openlm-research/open_llama)
19. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. In: International Conference on Learning Representations (ICLR) (2015)
21. Google: Gemini: A Family of Highly Capable Multimodal Models (2023). arXiv preprint: [arXiv:2312.11805](https://arxiv.org/abs/2312.11805)
22. Grynbaum, M.M., Mac, R.: The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work (2023). <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>. Accessed 01 May 2024
23. Gu, T., Dolan-Gavitt, B., Garg, S.: BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain (2017). arXiv preprint: [arXiv:1708.06733](https://arxiv.org/abs/1708.06733)
24. Hintersdorf, D., Struppek, L., Brack, M., Friedrich, F., Schramowski, P., Kersting, K.: Does CLIP Know My Face? (2023). arXiv preprint: [arXiv:2209.07341](https://arxiv.org/abs/2209.07341)
25. Hintersdorf, D., Struppek, L., Kersting, K.: To trust or not to trust prediction scores for membership inference attacks. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 3043–3049 (2022)
26. Hintersdorf, D., Struppek, L., Neider, D., Kersting, K.: Defending Our Privacy With Backdoors (2023). arXiv preprint [arXiv:2310.08320](https://arxiv.org/abs/2310.08320)
27. Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., McAuley, J.: Adversarial deepfakes: evaluating vulnerability of deepfake detectors to adversarial examples. In: Winter Conference on Applications of Computer Vision (WACV), pp. 3348–3357 (2021)
28. Ilharco, G., et al.: OpenCLIP (2021)
29. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 125–136 (2019)
30. Images, G.: Getty Images Statement (2023). <https://newsroom.gettyimages.com/en/getty-images/getty-images-statement>. Accessed 24 July 2023
31. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410 (2019)

32. Li, G., Rezaei, S., Liu, X.: User-Level Membership Inference Attack against Metric Embedding Learning (2022). arXiv preprint: [arXiv:2203.02077](https://arxiv.org/abs/2203.02077)
33. Li, H., Xu, M., Song, Y.: Sentence embedding inversion leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In: Findings of the Association for Computational Linguistics, pp. 14022–14040 (2023)
34. Li, Z., Zhang, Y.: Membership leakage in label-only exposures. In: Conference on Computer and Communications Security (CCS), pp. 880–895 (2021)
35. Ludewig, B.: <https://www.flickr.com/photos/finnishgovernment/51941396612/>. Licensed as CC BY 2.0. Accessed 24 July 2023
36. Lukas, N., et al.: Analyzing Leakage of Personally Identifiable Information in Language Models. In: Symposium on Security and Privacy (S&P), pp. 346–363 (2023)
37. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
38. OpenAI: GPT-4 Technical Report (2024). arXiv preprint: [arXiv:2303.08774](https://arxiv.org/abs/2303.08774)
39. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical Black-Box Attacks against Machine Learning. In: Asia Conference on Computer and Communications Security (ASIA CCS), p. 506–519 (2017)
40. Parikh, R., Dupuy, C., Gupta, R.: Canary extraction in natural language understanding models. In: Annual Meeting of the Association for Computational Linguistics (ACL) - Short Paper, pp. 552–560 (2022)
41. Parliament, E.: <https://www.flickr.com/photos/36612355@N08/52888839914/>. Licensed as CC BY 2.0. Accessed 05 Mar 2024
42. do Planalto, P.: <https://www.flickr.com/photos/51178866@N04/53055897555/>. Licensed as CC BY 2.0. Accessed 05 Mar 2024
43. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685 (2022)
45. Saha, A., Subramanya, A., Pirsiavash, H.: Hidden trigger backdoor attacks. In: Conference on Artificial Intelligence (AAAI), pp. 11957–11965 (2020)
46. Saha, A., Tejankar, A., Koohpayegani, S.A., Pirsiavash, H.: Backdoor attacks on self-supervised learning. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13337–13346 (2022)
47. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: Annual Network and Distributed System Security Symposium (NDSS) (2019)
48. Scao, T.L., et al.: BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (2022). arXiv preprint: [arXiv:2211.05100](https://arxiv.org/abs/2211.05100)
49. Shejwalkar, V., Houmansadr, A., Kairouz, P., Ramage, D.: Back to the drawing board: a critical evaluation of poisoning attacks on production federated learning. In: Symposium on Security and Privacy (S&P), pp. 1354–1371 (2022)
50. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Symposium on Security and Privacy (S&P), pp. 3–18 (2017)
51. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and Mitigating Copying in Diffusion Models (2023). arXiv preprint: [arXiv:2305.20086](https://arxiv.org/abs/2305.20086)

52. Struppek, L., Hintersdorf, D., Correia, A.D.A., Adler, A., Kersting, K.: Plug & play attacks: towards robust and flexible model inversion attacks. In: International Conference on Machine Learning (ICML), pp. 20522–20545 (2022)
53. Struppek, L., Hintersdorf, D., Friedrich, F., Brack, M., Schramowski, P., Kersting, K.: Image Classifiers Leak Sensitive Attributes About Their Classes (2023). arXiv preprint [arXiv:2303.09289](https://arxiv.org/abs/2303.09289)
54. Struppek, L., Hintersdorf, D., Kersting, K.: Rickrolling the artist: injecting backdoors into text-guided image generation models. In: International Conference on Computer Vision (ICCV) (2023)
55. Struppek, L., Hintersdorf, D., Kersting, K.: Be careful what you smooth for: label smoothing can be a privacy shield but also a catalyst for model inversion attacks. In: International Conference on Learning Representations (ICLR) (2024)
56. Struppek, L., Hintersdorf, D., Neider, D., Kersting, K.: Learning to break deep perceptual hashing: the use case NeuralHash. In: Conference on Fairness, Accountability, and Transparency (FAccT), pp. 58–69 (2022)
57. Struppek, L., Le, M.H., Hintersdorf, D., Kersting, K.: Exploring the adversarial capabilities of large language models (2024). arXiv preprint: [arXiv:2402.09132](https://arxiv.org/abs/2402.09132)
58. Subbanna, N., Wilms, M., Tuladhar, A., Forkert, N.D.: An analysis of the vulnerability of two common deep learning-based medical image segmentation techniques to model inversion attacks. *Sensors* **21**(11), 3874 (2021)
59. Touvron, H., et al.: LLaMA: Open and Efficient Foundation Language Models (2023). arXiv preprint [arXiv:2302.13971](https://arxiv.org/abs/2302.13971)
60. Vincent, J.: Getty Images is Suing the Creators of AI Art Tool Stable Diffusion for Scraping Its Content (2023). <https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit>. Accessed 24 July 2023
61. Wang, K., Fu, Y., Li, K., Khisti, A., Zemel, R.S., Makhzani, A.: Variational model inversion attacks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 9706–9719 (2021)
62. Xu, J., Xue, M., Picek, S.: Explainability-based backdoor attacks against graph neural networks. In: ACM Workshop on Wireless Security and Machine Learning, pp. 31–36 (2021)
63. Yao, Y., Li, H., Zheng, H., Zhao, B.Y.: Latent backdoor attacks on deep neural networks. In: Conference on Computer and Communications Security (CCS), pp. 2041–2055 (2019)
64. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: analyzing the connection to overfitting. In: Computer Security Foundations Symposium (CSF), pp. 268–282 (2018)
65. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models (2023). arXiv preprint [arXiv:2303.17591](https://arxiv.org/abs/2303.17591)
66. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D.: The secret revealer: generative model-inversion attacks against deep neural networks. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 250–258 (2020)
67. Zhang, Z., Jia, J., Wang, B., Gong, N.Z.: Backdoor attacks to graph neural networks. In: ACM Symposium on Access Control Models and Technologies (SACMAT), pp. 15–26 (2021)
68. Zhang, Z., et al.: Neurotoxin: Durable Backdoors in Federated Learning. In: International Conference on Machine Learning (ICML), vol. 162, pp. 26429–26446 (2022)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# AI-Related Risk and Uncertainty

Giacomo Zanotti<sup>1</sup>(✉), Daniele Chiffi<sup>2</sup>, and Viola Schiaffonati<sup>1</sup>

<sup>1</sup> Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy

giacomo.zanotti@polimi.it

<sup>2</sup> Department of Architecture and Urban Studies (DAStU), Politecnico di Milano, Milan, Italy

**Abstract.** Discussions on the risks involved in the deployment of AI systems are increasingly prominent in both public discourse and scientific debates. While talk of risk plays a crucial role in framing ethical and societal problems related to AI, we argue that it could profitably be associated with a clear analysis of uncertainty. Starting from a multi-component approach to AI-related risk assessment and mitigation, this chapter discusses the way the deployment of AI systems often takes place in contexts in which uncertainty is not meaningfully quantifiable.

**Keywords:** AI · Risk · Uncertainty · Philosophy of risk

## 1 Introduction

Recent advances in the field of Artificial Intelligence (AI) have resulted in a widespread diffusion of AI systems to be applied for significantly heterogeneous purposes in a wide range of situations. In many cases, these systems are delegated with complex tasks that would typically require human intervention. What is more, they are increasingly employed in delicate contexts in which their decisions, predictions and classifications can have a significant impact on people's life. Most notably, we can think about the fields of medical AI and predictive justice, or systems employed for loan processing and autonomous driving.

With things being this way, the growing prominence of the notion of risk in discussions on the ethical and social implications of AI does not come as a surprise. On the one hand, a fair deal of literature and public discourse has been focusing on the so-called *existential* risks related to the deployment of AI systems, often involving human extinction or global catastrophes. On the other hand, usually in open contrast with the talk on existential risk, increasing attention has been devoted to more mundane forms of AI-related risk.<sup>1</sup> This latter approach – which is also the one behind this contribution – has led, among other things, to the recently approved European proposal for the first comprehensive regulation on AI – the so-called AI Act<sup>2</sup> – where systems are classified

<sup>1</sup> <https://www.nature.com/articles/d41586-023-02094-7>, last accessed 2024/04/04.

<sup>2</sup> More precisely, the *Regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* ([https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138\\_EN.html](https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html), last accessed 2024/04/04).

in accordance to their level of risk (unacceptable, high, limited, minimal) and regulated accordingly.

Notwithstanding its importance, the assessment of risk may have some limitations when it comes to the outcomes and implications of AI systems. In fact, risk is understood as a normative notion associated with potential negative consequences, and it is often characterized by a distinct probabilistic component. More specifically, when referring to the risk of an event  $x$ , we typically imply that we can meaningfully assign a probabilistic value to the occurrence of  $x$ . This possibility is not always feasible when it comes to AI systems and the potential effects of their use. This contribution contends that talk of risk in AI should make room for the notion of uncertainty, both in quantifiable and unquantifiable forms. While our analysis is distinctively philosophical in scope and methodology, we believe it may be used as a theoretical ground for devising risk assessment practices in AI.

In Sect. 2, the notion of risk is presented, paying particular attention to multi-component approaches to risk and their quantifiable uncertainties, with a specific focus on AI-related risk. Then, in Sect. 3, the notion of severe uncertainty is introduced as a way to better understand and assess those cases where uncertainty cannot be meaningfully quantified. We will focus on the use of general-purpose AI systems as paradigmatic examples of a context in which this dimension of severe uncertainty is particularly relevant. Section 4 concludes the chapter by discussing possible future lines of research.

## 2 Risk and Its Components

Providing a univocal characterization of risk is not an easy task, for non-technical understandings of this notion come together with a number of technical definitions. Among these, the one provided by the Royal Society in 1983 is often referred to as the “classic” one, equating risk with “the probability that a particular adverse event occurs during a stated period of time, or results from a particular challenge” (Royal Society, 1983). Needless to say, it does not all come down to probability. As a matter of fact, assessing risk typically involves some form of expectation in which the probability of the unwanted event becomes the weight for the magnitude of its consequences: a higher magnitude might counterbalance a lower probability of occurrence, and *vice versa*. Still, probability is usually required in many definitions of risk.

Among other things, this way of understanding risk seems to be at the basis of the AI Act, that explicitly defines risk as “the combination of the probability of an occurrence of harm and the severity of that harm” (Art. 3, 2). However, it is not the only way to approach risk, in particular when it comes to designing risk-mitigation policies and interventions. Most notably, approaches adopted in the domain of disaster risk mitigation understand risk as the result of the interaction between three different components: hazard, exposure, and vulnerability (UNISDR, 2015). *Hazard* refers to the source of potential harm, *exposure* to the people and resources that could be harmed, and *vulnerability* has to do with how much what is exposed is susceptible to the impacts of the hazard. As an example, consider seismic risk. In this case, the hazard component refers to the earthquake itself, and its assessment involves estimates concerning both the probability and the magnitude of the earthquake. When it comes to exposure, instead, we



focus on what could be harmed by the earthquake, considering both people and material assets (e.g., buildings and infrastructures) that are found in the seismic hazard zone. Finally, one should take into account the circumstances and measures that could render these individuals and assets more or less susceptible to the potential harm: in the case of an earthquake, relevant elements could be the seismic safety standards of the potentially affected buildings, the existence of response plans, and the availability of temporary shelters.

Distinguishing between the different components we have just seen allows us to intervene on several fronts to reduce risk. Now, reducing the hazard is not always possible, especially in the case of some *natural* risks: we simply cannot prevent an earthquake from occurring. However, there are many cases, especially those in which the hazard is related to human action, in which there is much we can do (e.g., we might relocate polluting factories away from population centers, or withdraw from the market a potentially dangerous technology). At the same time, we can intervene on the exposure. In the case of seismic risk, the most straightforward way to do this involves limiting the number of people and assets in the areas that are more likely to be affected by earthquakes. Finally, efforts can be made to reduce the vulnerability of such people and assets by intervening on buildings to improve their safety, designing evacuation plans, and so on.

While the domain of natural risk offers intuitive examples of how different risks can be better analyzed and managed by distinguishing their components, nothing prevents us from applying the same kind of analysis to technological risks – that is, risks stemming from the use of technological artifacts.<sup>3</sup> AI systems make no exception. On the contrary, thinking of AI-related risk through the conceptual and methodological lens of multi-component analyses of risk allows us to understand how and why significantly different kinds of AI systems involve non-negligible levels of risk.<sup>4</sup>

In some cases, AI systems strike us as involving considerable levels of risk as a result of the hazard's magnitude. Let us take a look at the AI Act's Annex III, listing (some of) the systems that are considered as "high-risk" within the scope of the Act, and are therefore subject to stricter regulation. Among these, we can find AI systems that serve as "safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity", or systems used by law enforcement. It is fairly straightforward that malfunctions in such systems directly result in potentially harming events. The failure of an AI system used to manage road traffic can result in life-threatening accidents, and a system used to predict recidivism in courts can be affected by biases that may ultimately result in unfair judgments and unjustified detention (Angwin *et al.*, 2016). In these cases, regardless of the levels of exposure and vulnerability, the fact that these systems involve high levels of hazard seems to be enough for labeling them as "high-risk".

The reasoning is diametrically opposite when it comes to those systems that qualify as highly risky due to their affecting and/or being used by a considerable number of people. In this regard, recommender systems are the most prominent example, especially those

---

<sup>3</sup> Note that the dichotomy between natural and technological risks is not meant to be always completely exhaustive (Hansson, 2016).

<sup>4</sup> For a detailed analysis of a multi-component approach to AI-related risk, see (Zanotti, Chiffi, Schiaffonati, 2024).

implemented in the so-called *very large online platforms* (VLOPs).<sup>5</sup> In these cases, the potentially low levels of hazard and vulnerability are negatively counterbalanced by significantly high levels of exposure.

Finally, some systems might qualify as high-risk as a result of the vulnerability of their users. Examples abound. For instance, AI-systems – including intelligent robotic systems – are increasingly used in the context of education and elderly care (Miyagawa *et al.*, 2019; Tanaka *et al.*, 2015). In these cases, even assuming low levels of hazard and exposure, the vulnerability of the people using or being affected by AI systems compels us to guard against potential unwanted outcomes and accordingly treat the involved technologies as high-risk ones.

We have now seen how adopting a specific approach to risk, namely a multi-component analysis, can help us better understand AI-related risk. The two general features of risk that have been partially anticipated, however, remain valid. First of all, risk refers to the possible occurrence of an *unwanted event*, of something that is negatively valued. This is immediately evident in the classic definition of risk, that explicitly refers to *adverse* events, and it is clear in multi-component analyses of risk, that understand hazard as the source of *harm*. Accordingly, when referring to AI-related risk, we always focus on the *negative* potential consequences of AI systems' deployment. The second feature that typically characterizes conceptions of risk is that they involve the possibility of a meaningful probabilistic evaluation of the unwanted events in question.<sup>6</sup>

Sometimes, probabilistic risk assessment is assumed and conducted by using point-like probabilistic values, since we trust such probabilities. This can be a good choice when the uncertainty and complexity of the risk are not particularly noteworthy. This is what typically happens in textbook cases and idealized scenarios: if we bet on dice games and the dice is a fair one, we know exactly which our risk of losing is. However, more commonly some quantifiable forms of uncertainty are acknowledged within risk, and this is why risks may be quantified and evaluated by means of probabilistic intervals, second-order probabilities, imprecise probabilities, belief-functions, possibility theories, and fuzzy logic, just to mention some of these methods (Hansson, 2018; Dencœur *et al.*, 2020a, 2020b). AI makes no exception. On the contrary, providing point-like probabilities may be hard in the case of AI systems' deployment, for such systems are often used in complex contexts in which unanticipated circumstances might influence the course of events, and their being often relatively new technologies may result in a paucity of data concerning their use and its possible negative outcomes.

### 3 AI-Related Risks and Severe Uncertainty

Taking stock, we have seen how the notion of risk is associated with the possibility of making probabilistic estimates about unwanted events and their outcomes. True, some components of uncertainty are typically involved in real-world scenarios, for it is often hard to assess risk by means of point-like probabilistic values, and AI makes no exception. Still, the uncertainty in question can be quantified. However, this is not always the case:

<sup>5</sup> According to the European *Digital Services Act*, a platform qualifies as a VLOP if it has more than 45 million users per month in the EU (DSA, 2022).

<sup>6</sup> In the literature, these situations are understood as “known unknowns” (Hansson, 2009).

several types of uncertainty exist, and not all of them can be meaningfully quantified (Hansson, 2022). In this section, we analyze how non-quantifiable uncertainty may play a role in the assessment of AI-related risk. Note that, while we focus on the way risk (and quantifiable forms of uncertainty) differ from non-quantifiable forms of uncertainty with respect to our probabilistic knowledge of possible scenarios, other differences exist. Most notably, we have seen how risk is a normative and evaluative concept with a negative connotation. This does not always happen with uncertainty. On the contrary, some forms of uncertainty are usually assumed to be possible triggers for technological innovation (Chiffi, Moroni, Zanetti, 2022).

Based on what we have seen in the previous section, risk assessment seems to depend on our evaluation of the potential unwanted events in question, their consequences and contexts of occurrence. For instance, in the case of seismic risk, assessing exposure requires to possess reliable knowledge about the location and extension of the potentially affected area as well as the number of people, buildings and infrastructures therein. In addition, up-to-date information concerning (among other things) the existence of evacuation plans and buildings' safety standards is needed to evaluate the vulnerability of exposed people and assets. All of this applies to the case of AI. Suppose you want to estimate the risk associated with the deployment of a certain AI system. First, you need to identify possible inaccuracies, malfunctions, misuses, and more generally all unintended and unwanted consequences resulting from the deployment of the system, and possibly associate them with a probability. Then, you must have a sufficiently precise idea of the people and assets exposed to such consequences. Finally, you should be able to assess their vulnerability by considering all those factors and circumstances that make them more or less prone to be harmed by the potential events in question.

While this might be doable for some AI systems and in some contexts (e.g., AI systems based on symbolic techniques to be used in controlled environments), it is not always possible. In some cases, it might be hard to make predictions on the possible *inaccuracies* and *malfunctions* of AI systems, often due to their complexity and working opacity. In addition to this, we might not be able to anticipate their possible uses, and therefore their *misuses*, and identify who could be affected by their negative outcomes. As we will see in a moment, such difficulties might be due to the fact that some kinds of AI systems can be adapted to a wide variety of uses and applications. On top of that, we should keep in mind that, in many cases, the technologies we are referring to are relatively recent, and we largely lack data on their real-world use that could inform our predictions.

In the literature, analogous situations are captured through the notion of *severe uncertainty*. Severe uncertainty is typically conceived in open contraposition to probabilistic conceptualizations of risk such as the Royal Society's one we have seen in Sect. 2. Consider the (fair) dice game example. In this case, we have exhaustive and reliable knowledge of both (i) the possible outcomes of the roll of the dice and (ii) the probability associated with each outcome.

In situations of severe uncertainty, things are less clear. For a specific set of events, we might be able to anticipate the possible outcomes while ignoring their probability distribution. Many of the recent and most impactful AI technologies seem to be used in and give rise to contexts of severe uncertainty. The example we propose to consider is that

of so-called *general-purpose AI systems* (GPAIs). This expression, that for the purpose of this chapter we take to be largely overlapping with the one of *foundation models* (Bommasani *et al.*, 2022), refers to any AI system that can “accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained” (Gutierrez *et al.*, 2023). The class of GPAIs includes different models and systems, from those designed for computer vision to those for multimodal processing. Among these, however, Large Language Models (LLMs) are increasingly widespread, especially after OpenAI’s decision in November 2022 to implement their model GPT3.5 in a freely available chatbot with a user-friendly interface. From that moment on, different companies and developers rushed to offer easily accessible LLM-based platforms at users’ fingertips.

When it comes to these systems, assessing risk is particularly difficult. First of all, the identification of malfunctions, misuses, and unintended consequences might be quite critical. As a matter of fact, their being general-purpose models, so capable of tasks for which they have not been specifically designed and trained, makes it very difficult to anticipate all the potential consequences of their use. Moreover, the fact that these systems are most of the time running proprietary software (not an open source one) further exacerbates the possibility to predict malfunctions. True, many possible scenarios of malfunctions and abuses can be foreseen. For instance, once we know that certain GPAIs can be used for code generation, we can easily anticipate that someone may jailbreak them to write malware. However, it is not clear how we could associate a probability to this scenario before the system’s large-scale deployment.

Analogous considerations can be made when it comes to estimating the exposure component of the risks involved in the deployment of GPAIs. Many GPAIs are now implemented in free and accessible platforms, and the number of people making use of these systems in their daily life is increasing – again, their flexibility makes them potentially applicable to significantly different tasks and in a wide range of situations. Such an evolving scenario also makes it very difficult to have a sufficiently precise idea of the people exposed to their consequences.

Finally, in light of this, it is not hard to see the difficulties involved in the attempt to estimate the component of vulnerability associated with these systems’ risk. To do so, as a matter of fact, we should be able to identify both the potential harmful uses of GPAIs as well as those affected by their possible negative consequences. And again, this is not an easy task.

Summing up, we could say that the extreme flexibility of some AI systems, GPAIs in particular, plays a major role in giving rise to severe forms of uncertainties: as their possible uses are wide and open, it is hard to anticipate and assess all of them and thereby estimate the associated levels of hazard, exposure and vulnerability. These forms of uncertainty are hardly quantifiable and represent a significant challenge in assessing AI-related risk, but cannot be overlooked in a rigorous and complete discussion of AI technologies and their societal implications.

## 4 Conclusion

We discussed some possible difficulties in assessing the risks associated with the use of AI systems. Starting from a focus on the components of risk, namely, hazard, exposure, and vulnerability, we highlighted that traditional risk analysis often relies on probabilistic information, which may not be always readily available or reliable for the outcomes of AI systems' deployment. We suggested that incorporating the concept of uncertainty into AI-related risk analysis is beneficial not only when uncertainty is quantifiable but also, and more importantly, when it is not quantifiable. This is particularly relevant in cases of severe forms of uncertainties. We explored general-purpose AI systems as an illustrative example of technology where severe uncertainty may play a pivotal role in risk assessment. Among other things, this uncertainty arises due to the considerable flexibility in these systems' potential applications.

In future lines of research, we will investigate the role of multi-risk analysis related to AI, wherein various risks may interact mutually, potentially producing domino or cascade effects<sup>7</sup>. To this end, we will draw upon the rich literature on engineering safety, risk assessment and uncertainty (e.g., Burton, Mcdermid, Freng, 2023), in particular in the context of AI (e.g., NIST, 2023). We will also explore the impact of unforeseeable events, sometimes referred to as “unknown unknowns,” on AI-related risks. These events can be challenging not only to quantify but also to predict accurately and are typically associated with socio-technical systems, which may pose wicked problems to society – complex issues often intertwined with policy and planning (Rittel & Webber, 1973; Nordström, 2022). Such problems are difficult to address and even analytically define. A rigorous epistemological analysis of uncertainty in AI, however, will hopefully put us in a better position to deal with them.

**Acknowledgments.** This study was funded by (1) the Italian Ministry of University and Research under the PRIN Scheme (Project BRIO, no. 2020SSKZ7R; Project NAND no. 2022JCMHFS); (2) RETURN, Extended Partnership, Multi-risk science for resilient communities under a changing climate, European Union Next-GenerationEU (National Recovery and Resilience Plan – NRRP, Mission 4, Component 2, Investment 1.3 – D.D. 12432/8/2022, PE0000005); (3) PNRR-PE-AI FAIR-NextGeneration EU program.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

Angwin, J., Larson, J., Mattu, S., Kirchner, L: Machine bias. *Pro Publica* (2016). <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<sup>7</sup> In fact, increasing attention has been devoted to the so-called “Natech” risks. Natech stands for “natural hazard triggering technological accident” (Mesa-Gómez, Casal, Muñoz, 2020). A notorious example of a Natech accident was the nuclear one that took place in Ōkuma, in the Japanese prefecture of Fukushima, in March 2011, when a tsunami hit the Fukushima Daiichi nuclear power plant causing a failure in the electric grid and damaging backup generators, which ultimately resulted in a leak of radioactive contaminants.

- Bommasani, R., et al.: On the opportunities and risks of foundation models (2022). arXiv preprint [arXiv:2108.07258](https://arxiv.org/abs/2108.07258)
- Burton, S., Mcdermid, J., Freng, O.: Closing the gaps: complexity and uncertainty in the safety assurance and regulation of automated driving (2023). <https://www.iks.fraunhofer.de/content/dam/iks/documents/whitepaper-closing-the-gaps.pdf>
- Chiffi, D., Moroni, S., Zanetti, L.: Types of technological innovation in the face of uncertainty. *Philos. Technol.* **35**(4), 94 (2022)
- Dencœur, T., Dubois, D., Prade, H.: Representations of uncertainty in artificial intelligence: probability and possibility. In: Marquis, P., Papini, O., Prade, H. (eds.) *A Guided Tour of Artificial Intelligence Research*, pp. 69–117. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-06164-7\\_3](https://doi.org/10.1007/978-3-030-06164-7_3)
- Dencœur, T., Dubois, D., Prade, H.: Representations of uncertainty in AI: beyond probability and possibility. In: Marquis, P., Papini, O., Prade, H. (eds.) *A Guided Tour of Artificial Intelligence Research*, pp. 119–150. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-06164-7\\_4](https://doi.org/10.1007/978-3-030-06164-7_4)
- Gutierrez, C.I., Aguirre, A., Uuk, R., Boine, C.C., Franklin, M.: A proposal for a definition of general purpose artificial intelligence systems. *Digit. Soc.* **2**(3), 36 (2023)
- Mesa-Gómez, A., Casal, J., Muñoz, F.: Risk analysis in Natech events: state of the art. *J. Loss Prev. Process Ind.* **64**, 104071 (2020)
- Miyagawa, M., et al.: Consideration of safety management when using Pepper, a humanoid robot for care of older adults. *Intell. Control Autom.* **11**, 15–24 (2019)
- Hansson, S.O.: From the casino to the jungle: dealing with uncertainty in technological risk management. *Synthese* **168**(3), 423–432 (2009)
- Hansson, S.O.: Managing risks of the unknown. In: Gardoni, P., Murphy, C., Rowell, A. (eds.) *Risk Analysis of Natural Hazards*, pp. 155–172. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-22126-7\\_10](https://doi.org/10.1007/978-3-319-22126-7_10)
- Hansson, S.O.: Representing uncertainty. In: Hansson, S., Hendricks, V. (eds.) *Introduction to Formal Philosophy*, pp. 387–400. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-77434-3\\_19](https://doi.org/10.1007/978-3-319-77434-3_19)
- Hansson, S.O.: Can uncertainty be quantified? *Perspect. Sci.* **30**(2), 210–236 (2022)
- NIST AI 100-1: Artificial Intelligence Risk Management Framework (AIRMF 1.0) (2023). <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
- Nordström, M.: AI under great uncertainty: implications and decision strategies for public policy. *AI Soc.* **37**(4), 1703–1714 (2022)
- Rittel, H.W., Webber, M.M.: Dilemmas in a general theory of planning. *Policy. Sci.* **4**(2), 155–169 (1973)
- Regulation 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022, on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act – DSA). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065>. Accessed 04 Apr 2024
- Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., Hayashi, K.: Pepper learns together with children: development of an educational application. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), Seoul, pp. 270–275. IEEE (2015)
- United Nations Office for Disaster Risk Reduction (UNISDR): UNISDR Annual Report 2015: 2014–15 Biennium Work Programme Final Report, Geneva (2015). [https://www.unisdr.org/files/48588\\_unisdrannualreport2015evs.pdf](https://www.unisdr.org/files/48588_unisdrannualreport2015evs.pdf). Accessed 04 Apr 2024
- Zanotti, G., Chiffi, D., Schiaffonati, V.: AI-related risk: an epistemological approach. *Philos. Technol.* **37**, 66 (2024)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# Leveraging Actionable Explanations to Improve People's Reactions to AI-Based Decisions

Markus Langer<sup>1</sup> (✉) and Isabel Valera<sup>2</sup>

<sup>1</sup> Department of Psychology, University of Freiburg, Freiburg im Breisgau, Germany  
markus.langer@psychologie.uni-freiburg.de

<sup>2</sup> Department of Computer Science, Saarland University, Saarbrücken, Germany

**Abstract.** This paper explores the role of explanations in mitigating negative reactions among people affected by AI-based decisions. While existing research focuses primarily on user perspectives, this study addresses the unique needs of people affected by AI-based decisions. Drawing on justice theory and the algorithmic recourse literature, we propose that actionability is a primary need of people affected by AI-based decisions. Thus, we expected that more actionable explanations – that is, explanations that guide people on how to address negative outcomes – would elicit more favorable reactions than feature relevance explanations or no explanations. In a within-participants experiment, participants ( $N = 138$ ) imagined being loan applicants and were informed that their loan application had been rejected by AI-based systems at five different banks. Participants received either no explanation, feature relevance explanations, or actionable explanations for this decision. Additionally, we varied the degree of actionability of the features mentioned in the explanations to explore whether features that are more actionable (i.e., reduce the amount of loan) lead to additional positive effects on people's reactions compared to less actionable features (i.e., increase your income). We found that providing any explanation led to more favorable reactions, and that actionable explanations led to more favorable reactions than feature relevance explanations. However, focusing on the supposedly more actionable feature led to comparably more negative effects possibly due to our specific context of application. We discuss the crucial role that perceived actionability may play for people affected by AI-based decisions as well as the nuanced effects that focusing on different features in explanations may have.

**Keywords:** explainability · actionability · consequential decision-making · acceptance · affected people

## 1 Introduction and Related Work

AI-based decisions affect the fate and future of individuals. This is true in high-stakes contexts such as medical diagnosis and treatment [1], hiring [2], and loan contexts [3]. Whereas users of AI-based decision-support tools (e.g., doctors, hiring managers, loan officers) can decide whether and to what extent they want to use such tools, people affected by AI-based decisions (e.g., patients, job and loan applicants) are in a very



different position [4]. Although regulation such as the European GDPR in principle requires the possibility to opt out of a fully automated evaluation by AI-based tools, people affected by AI-based decisions often have less control over whether and to what extent their health, job suitability, or creditworthiness is evaluated by AI-based tools. Particularly in high-stakes situations and when decisions are highly automated (i.e., there is little human influence on the decision), research has shown that people affected by AI-based decisions tend to react negatively to the use of AI-based tools [4, 5]. This critical view of AI-based decisions by the public can undermine the potential benefits of AI-based tools for society and can also be detrimental to the image of organizations that use such tools for high-stakes decisions.

Providing explanations for AI-based decisions has been proposed as a promising way to counteract unfavorable reactions to AI-based decisions [6]. The basic idea is simple: by default, many AI-based tools do not provide any insights into their decision-making. Providing explanations could help to better understand the reasons for AI-based decisions, and thus lead to other positive effects such as a higher perceived contestability, higher perceived justice, or a general higher acceptance [7]. However, empirical evidence supporting the expected positive effects of explanations on people affected by AI-based decisions is surprisingly rare.

In fact, most research on the effects of explanations by AI-based systems has focused on users of AI-based systems [7, 8]. It is questionable whether findings from this area can be generalized to people affected by AI-based decisions simply because the position of users in the decision context differs substantially from that of people affected by AI-based systems. For example, the former can ignore AI-based outputs, whereas the latter are subject to these outputs. Additionally, the former may have more experience with and insight into these systems, whereas the latter may have little experience and no way of gaining insights into the systems decision rationale. Consequently, the needs with respect to explanations of people affected by AI-based decisions may also be substantially different [7]. The few studies that have examined the perspective of people affected by AI-based decisions have produced mixed results [4, 5]. Whereas some studies found positive effects of explanations on important outcomes such as perceived informational justice, procedural justice and perceived overall fairness [9], others found no effect [10], ambivalent effects [6, 11] or even negative effects of explanations [12]. One possible reason for this heterogeneity is that research has often compared different explainability approaches without a clear rationale for why one would be better suited to the needs of people affected by AI-based decisions than others [6].

In line with claims from research on algorithmic recourse [13, 14], we propose that one crucial need of people affected by AI-based decisions is the actionability of an explanation accompanying AI-based decisions. In other words, they want to be able to act on such explanations, they want to know ways forward, especially in cases where they experience a negative outcome in an AI-based decision situation (e.g., a loan application is rejected). For example, instead of simply telling loan applicants that their application was denied because their income was too low, it may be more actionable to tell them that they need a certain percentage of additional income. This intuition from research on algorithmic recourse is supported by justice theory [15]. In fact, particularly in the case of a negative outcome, the perceived justice of a decision becomes important in determining

people's overall reactions to the decision context [16]. In such cases, explanations need to be accurate, timely, and insightful to increase people's perceived informational justice of the decision situation [15]. This line of reasoning supports the intuition that any explanation should be better than no explanation and that actionable explanations may indeed be what people affected by decisions desire in the case of negative outcomes. Further support for the role of actionability comes from research suggesting that the design of explainability approaches in a given context needs to be informed by people's goals and needs [7, 17]. In the case of a negative outcome, people affected by a decision may want to know what to do next. This requires an explanation that is actionable.

In line with these considerations, we thus propose that providing actionable explanations should lead to more favorable reactions to an AI-based decision situation than providing no explanation or than providing a feature relevance explanation. Whereas feature relevance explanations should also be beneficial in terms of the perceived actionability relative to receiving no explanation at all, they typically only focus on giving people insights into important features that were important for the AI-based system's outputs. In contrast, actionable explanations aim at telling people what to do in order to achieve a better outcome in the future [13, 14].

To date there is little empirical evidence to support the claim that actionable explanations can help to foster acceptance, let alone evidence that actionable explanations are better than other explanations at doing so. For example, Schoeffer et al. [9] introduced their participants to a third-person perspective (seeing others being affected by an AI-based decision) and found that more detailed explanations led to more favorable reactions. Additional qualitative findings showed that their participants emphasized that the actionability of explanations and the actionability of highlighted features were important for them to find explanations helpful. Binns et al. [6] also introduced their participants to a third-person perspective and found that explanations had ambivalent effects: case-based explanations (highlighting cases similar to the affected person that may lead to insights into why a system has produced a respective output) led to comparatively more negative reactions than sensitivity-based explanations (highlighting what would have needed to be different for an output to be different; note that in other research, this kind of explanation was called a counterfactual explanation [13]). Additional qualitative insights may help to understand these findings because some participants mentioned that sensitivity-based explanations were perceived as more actionable. Singh et al. [14] found that participants in the role of a user (e.g., in the role of a loan officer) preferred more actionable explanations as the explanations that they would communicate to people affected by AI-based decisions. Additional qualitative findings indicate that explanations that focused on features that may be perceived as little actionable led to negative reactions. For example, their participants said that they found explanations that told people to increase their income by changing their job as impolite. One reason for this finding may be that this kind of explanation appears to be of limited actionability for many people who may not be able to simply change their job.

To shed light on the role of actionability of explanations on people's reactions to AI-based decisions, our study employed a within-participant design where participants received decisions with either no explanation, feature relevance explanations, or actionable explanations. Additionally, we tested whether highlighting features with different

degrees of actionability (i.e., where people could have the impression that it is easier to act upon the respective feature) affect people's reactions to the AI-based decision situations differentially. For example, we expected that most people would find it easier to apply for a slightly lower loan amount than to increase their income. If this actionability is important to people affected by AI-based decisions, providing more actionable explanations and focusing on more actionable features should lead to more favorable perceptions. This leads to the following hypotheses that we test in our study:

Hypothesis 1:<sup>1</sup> Receiving any explanation will lead to more favorable reactions to the AI-based decision situation (i.e., perceived attractiveness of the bank as a place to apply for a home loan, fairness, informational justice, procedural justice).

Hypothesis 2: Receiving actionable explanations will lead to more favorable reactions than receiving a feature relevance explanation.

Hypothesis 3: Focusing in an actionable explanation on a more actionable feature (i.e., reduce the amount of loan) will lead to more favorable perceptions than focusing on a less actionable feature (i.e., increase your income).

## 2 Methods

### 2.1 Sample

In our preregistration, we stated that we wanted to collect data from at least  $N = 120$  participants. We decided to collect the data via the university's participant pool and via Prolific. We ended up with  $N = 156$  participants before data cleaning. In line with our preregistered exclusion criteria, we excluded ten participants because they stated that their data should not be used (e.g., due to being inattentive), four participants who failed either one of two attention checks included in the questionnaire, and one participant who took less than 3 min to respond, indicating inattentive responding. We also excluded three participants who indicated their age to be 15 although we informed participants in the beginning that only participants above the age of 18 are allowed to participate. On average, the study took about 10 min to complete ( $SD = 3$ ). Student participants were compensated with course credit and Prolific participants received £1.80. The final sample consisted of  $N = 138$  participants. Of those, 68% were from Prolific. There were 67% participants who indicated their gender to be female, 32% male, and 1% diverse. The mean age was 39 ( $SD = 14$ ). Regarding their education, the majority indicated that they had finished school (25%), had a Bachelor's degree (37%), or a Master's degree (25%). About half (55%) of participants had already applied for a bank loan before.

---

<sup>1</sup> We preregistered this study on <https://aspredicted.org/5gq53.pdf>. There, we included an additional hypothesis that proposed that there would be an interaction effect, i.e., a stronger difference regarding the favorability of perceptions between the feature relevance and the actionable explanation condition for the more actionable feature. However, due to an error in the study design, participants saw an old version of one of the feature relevance conditions. Specifically, one of the feature relevance conditions did not mention the amount of loan as the decisive feature but mentioned the "bank balance." This made it impossible to test this interaction hypothesis.

## 2.2 Procedure

The experimental procedure was approved by the IRB of the first author's previous institution. The study was conducted in English and the experiment followed a within-person design with five conditions: no explanation, a feature relevance explanation focusing on the loan applicant's insufficient bank balance, a feature relevance explanation focusing on the applicant's insufficient monthly income, an actionable explanation focusing on the applicant's monthly income, and an actionable explanation focusing on the loan amount.

After being directed to the online survey, participants were welcomed and received information about data privacy and about the study. After giving their consent, participants were informed that they had to imagine that they wanted to buy a house. They had applied for a home-loan at five different banks. They were then told that in all those banks, an AI-based system decides about their home-loan. They were then informed that their application was rejected by all of the banks but that the reasons that those banks have provided differ. They were then told that they will see the rejection letters from the banks, each followed by a set of statements regarding their reactions to the respective decision that they are asked to respond to. The rejection letters included the experimental manipulations and were presented to participants in a randomized order to prevent order effects on our results.

Every rejection letter included the following information:

*“Dear applicant, we have deployed an AI-based evaluation software that helps us process applicant documents faster and with more precision. The data you have provided us with was*

- *Your age, Tenure, Income, Bank Balance*

*We regret to inform you that the AI-based evaluation software has rejected your application for a home-loan -EXPERIMENTAL MANIPULATION-.*

*We thank you for choosing us and look forward to seeing you sometime again in the future.”*

The EXPERIMENTAL MANIPULATION was then filled with the following information

- a) No explanation condition: No additional text
- b) Feature relevance condition focusing on the bank balance: *due to insufficient bank balance.*
- c) Feature relevance condition focusing on the monthly income: *due to insufficient monthly income.*
- d) Actionable explanation focusing on monthly income: *However, you would have a higher chance of approval for future loan applications if you increased your monthly income by 10%.*
- e) Actionable explanation focusing on amount of loan: *However, you would have a higher chance of approval for future loan applications if you reduced the amount of the loan you are asking for by 10%.*

For the actionable explanations, we decided to use the monthly income and the amount of loan anticipating that one would be perceived as more actionable than the other as indicated by the qualitative insights from [6, 9, 14]. We then decided for a 10% increase or decrease to keep the 10% consistent.

After each scenario, participants responded to measures capturing their perceived attractiveness of the bank as a place to apply for a home loan, perceived fairness, perceived informational justice, perceived actionability, and perceived procedural justice (i.e., they responded to these items five times). After the final scenario, we asked about participants' affinity for technology interaction, for their level of education, their socioeconomic status, their demographic information, whether they had ever applied for a bank loan, if we could use their data for our analyses, and whether they have any additional remarks. In the end, participants were debriefed about the purpose of the study.

### 2.3 Measures

All items were measured on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree) unless otherwise stated.

To check if our experimental manipulation increased the perceived actionability, we captured perceived actionability with three self-developed items. Those items were "I found the explanation to be actionable", "With the explanation I received, I now would know what to do differently the next time I apply for a loan", and "It would be easy for me to successfully apply for a loan in the future given the information that I have received from the bank."

Regarding measures to capture participants' reactions to the AI-based decision situation, we measured perceived attractiveness of the bank as a place to apply for a home-loan because this is a practically relevant reaction for organizations using AI-based tools to inform their decisions. Additionally, we focused on established measures related to the perceived justice [15] and overall fairness [18, 19] of the decision situation that were also used in prior research on the effects of explanations on people affected by AI-based decisions [9].

We measured perceived attractiveness of the bank as a place to apply for a home-loan with three items adapted from the organizational attractiveness measure by [20]. A sample item was "I would recommend others to apply for a home-loan at this bank."

We captured perceived fairness of the decision with two items by [18]. A sample item was "I think that the decision itself was fair."

We measured perceived informational justice with four items by [15]. A sample item was "Were explanations regarding the decision reasonable?"

We measured perceived procedural justice with four items by [15]. A sample item was "Have those procedures been free of bias?"

We captured participants' affinity for technology interactions with the four items of the scale by [21]. Here, we used the original six-point response scale. A sample item was "I like to occupy myself in greater detail with technical systems."

### 3 Results

Table 1 includes overall mean values, standard deviations, and intercorrelations between the study variables. For this table, we calculated the mean over all scenarios to get insights into how the variables are correlated to each other. This table shows the uncorrected correlations and indicates that all study variables were positively and strongly correlated to each other as can be expected with measures on perceived justice and fairness [19]. The overall mean values for the dependent variables were all below the mean of the scale (i.e., below 3), indicating that the overall reaction to the decision context was rather negative, which can be expected in the case of a negative decision outcome (i.e., a loan being rejected).

**Table 1.** Means and standard deviation between the mean of study variables over all scenarios.

Variable	<i>M</i>	<i>SD</i>	1	2	3	4	5	6
1. Actionability	3.09	0.62						
2. Attractiveness of the bank	2.74	0.57	.61**					
3. Fairness	2.92	0.73	.57**	.66**				
4. Informational Justice	2.83	0.63	.71**	.60**	.73**			
5. Procedural Justice	2.99	0.60	.67**	.63**	.73**	.74**		
6. Affinity for technology interactions	3.49	1.07	.11	.11	.00	-.01	.10	

Notes. *N* = 138

\*\*  $p < .01$

#### 3.1 Testing of Hypotheses

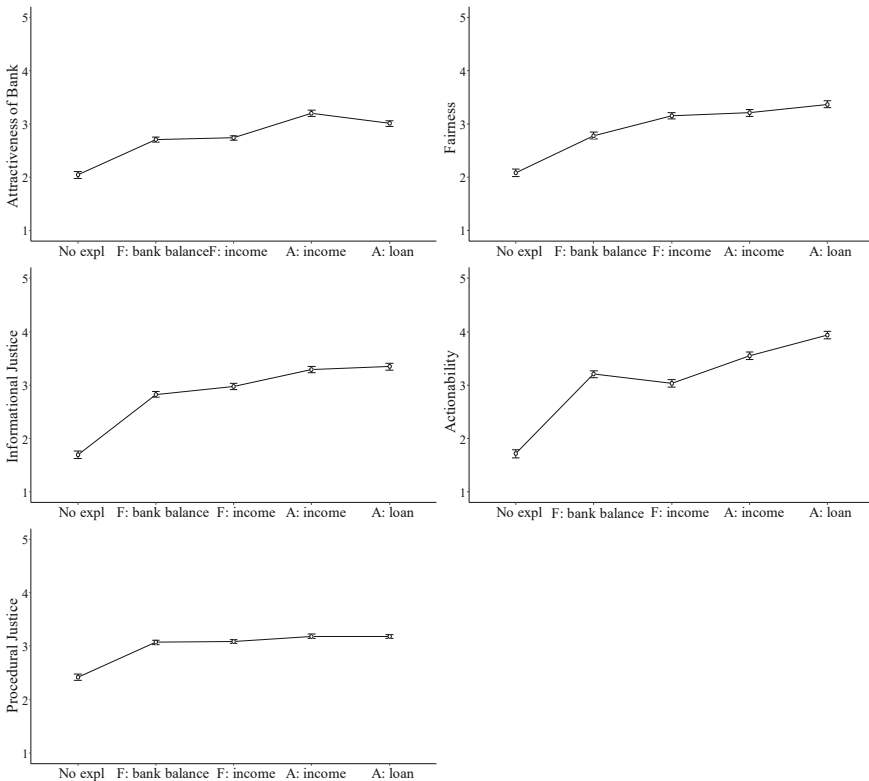
Figure 1 shows the mean values of the different scenarios for the dependent variables.

For all hypotheses, we calculated several hierarchical linear models with the participant as a random factor to account for the nested nature of the data. Specifically, scenarios are nested within participants because all participants saw all scenarios.

Hypothesis 1 stated that receiving any explanation will lead to more favorable reactions (i.e., perceived attractiveness of the bank as a place to apply for a home loan, fairness, informational justice, procedural justice). To test this hypothesis, we compared the no explanation condition to all the other conditions. Receiving any explanation led to more perceived actionability indicating that our manipulation worked as intended. Additionally, for all dependent variables, we found that receiving any explanation led to more favorable reactions. This supports hypothesis 1.

Hypothesis 2 proposed that receiving actionable explanations will lead to more favorable reactions than receiving a feature relevance explanation. To test this hypothesis, we compared the feature relevance conditions to the actionable explanation conditions. The actionable explanations led to a higher perceived actionability indicating that our manipulation worked as intended. Additionally, we found that receiving an actionable explanation led to more favorable reactions for all dependent variables. This supports hypothesis 2.

Hypothesis 3 stated that focusing in an actionable explanation on a more actionable feature (i.e., reduce the amount of loan) will lead to more favorable reactions than focusing on a less actionable feature (i.e., increase income). To test this hypothesis, we compared the actionable explanation focusing on the amount of loan condition to the actionable explanation focusing on the increase of income condition. Focusing on reducing the amount of loan led to more perceived actionability indicating that our manipulation worked. However, we found no significant difference for fairness, informational justice, or for procedural justice. Focusing in the explanation on reducing the amount of loan even led to a lower perceived attractiveness of the bank (Table 2).



**Fig. 1.** Mean values for the dependent variables across the experimental scenarios. *Notes:* Expl. = explanation, F = feature relevance explanation, A = actionable explanation. The error bars display standard errors.

Table 2. Results of the hierarchical linear models to test the hypotheses

Predictors	Actionability		Attractiveness		Fairness		Informational Justice		Procedural Justice						
	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>b</i>	<i>CI</i>	<i>p</i>	<i>p</i>					
<b>HLM analysis for Hypothesis 1</b>															
Intercept	1.71	1.55, 1.88	<.01	2.04	1.91, 2.18	<.01	2.08	1.91, 2.25	<.01	1.69	1.54, 1.85	<.01	2.42	2.29, 2.54	<.01
No Explanation vs. Explanation	1.71	1.55, 1.88	<.01	0.87	0.75, 0.99	<.01	1.05	0.90, 1.20	<.01	1.42	1.28, 1.56	<.01	0.71	0.62, 0.81	<.01
Marginal <i>R</i> <sup>2</sup>	.31			.15			.14			.28			.13		
Observations	690			690			690			690			690		
<b>HLM analysis for Hypothesis 2</b>															
Intercept	3.12	2.99, 3.24	<.01	2.72	2.61, 2.83	<.01	2.97	2.83, 3.11	<.01	2.90	2.77, 3.03	<.01	3.08	2.97, 3.18	<.01
Feature vs. Actionability	0.63	0.49, 0.76	<.01	0.38	0.28, 0.48	<.01	0.32	0.19, 0.45	<.01	0.42	0.31, 0.53	<.01	0.10	0.03, 0.18	.01
Marginal <i>R</i> <sup>2</sup>	.10			.06			.03			.05			.01		
Observations	552			552			552			552			552		
<b>HLM analysis for Hypothesis 3</b>															
Intercept	3.55	3.40, 3.70	<.01	3.20	3.07, 3.34	<.01	3.21	3.05, 3.37	<.01	3.29	3.14, 3.44	<.01	3.18	3.06, 3.30	<.01
Income vs. Loan	0.39	0.22, 0.57	<.01	-0.20	-0.34, -0.05	.01	0.16	-0.01, 0.32	.06	0.06	-0.08, 0.20	.43	-0.00	-0.10, 0.10	.94
Marginal <i>R</i> <sup>2</sup>	.05			.02			.01			.00			.00		
Observations	276			276			276			276			276		

Note. CI = 95% Confidence Interval. N = 138.



## 4 Discussion

The goal of this study was to assess whether actionable explanations are better suited than less actionable ones at mitigating negative reactions in a consequential decision scenario with a negative outcome for people affected by an AI-based decision. The main findings of this study were that a) any explanation for the unfavorable outcome was perceived as more actionable and led to more favorable reactions, b) what we defined to be more actionable explanations were also perceived as more actionable and led to more favorable reactions than feature relevance explanations, and c) focusing in an actionable explanation on a presumably more actionable feature (i.e., the loan amount) was perceived as more actionable, but led to less favorable reactions in the context of a loan application scenario.

Consequently, one main takeaway of our study is that providing explanations – particularly actionable ones – can alleviate negative reactions to unfavorable decision outcomes. In cases where AI-based tools make high-stakes decisions that affect human lives, it is inevitable that some people will not receive the outcome that they had hoped for. In particular, when AI-based systems make high-stakes decisions, this can lead to unfavorable reactions of people affected by AI-based decisions [4]. To mitigate such negative reactions, research and practice see potential in providing explanations for AI-based decisions. However, research has produced mixed results in this regard [6, 10]. We found that any explanation is better than no explanation for a negative outcome. Participants expressed higher levels of perceived informational justice, procedural justice, fairness, and most importantly, had less negative reactions toward the banks that provided them with an explanation.

Consistent with the intuition of research on the importance of the actionability of explanations for people affected by AI-based decisions [13], and consistent with qualitative findings from prior research [6, 9] indicating that actionability may be what people desire when they are affected by AI-based decisions, we found empirical evidence for the positive effect of actionable explanations. In contrast to the specific feature relevance explanations that we chose, the actionable explanations suggested to participants how to improve their chances of getting a loan next time. In line with justice theory [15], this improved our participants' perceived informational and procedural justice, as well as the perceived overall fairness of the decision processes [9]. Furthermore, in line with the propositions by [7] and [17], actionable explanations seem to have been better at providing our participants with information that would help them to work toward desired positive decision outcomes. This may have contributed to the overall more positive reactions when receiving actionable explanations.

While we propose that our actionable explanation led to stronger positive effects than the feature relevance explanation because of its “actionability,” there are also alternative possible explanations. For example, highlighting what needs to be done to get a better result also provides additional insight into how the model works. This means that our actionable explanation is not only more actionable, but also provides more details for people to understand the system. Thus, the current results could be driven by a “perceived understanding” effect rather than a “perceived actionability” effect. Disentangling these effects will be a task for future research. Another task for future research will be to understand the inherent subjectivity of the actionability of explanations. By definition,

actionability is a subjective aspect of explanations rather than something that can be described objectively. In other words, what is actionable for one person may not be actionable for another, and what is actionable for one person at one time may not be actionable for the same person in the future. In our study, our manipulations all led to the expected effects on perceived actionability, but predicting whether an explanation will actually be *perceived* as actionable by people affected by AI-based decisions, and whether the respective explanations will have other expected downstream effects (e.g., on perceptions of justice) may be more complex in other contexts.

Contrary to our expectations, focusing on the amount of loan as a supposedly more actionable feature in an actionable explanation did not have any additional positive effects. In fact, it negatively affected the perceived attractiveness of the bank. In line with previous research [6, 9, 14], we proposed this hypothesis because we expected that participants would think that it would require less effort on their part to reduce the loan amount instead of increasing their monthly income. In line with this reasoning, we found that reducing the loan amount was perceived as more actionable. However, this did not translate into positive reactions for the other dependent variables – although we want to highlight the slightly positive, but not significant effect on perceived fairness ( $b = 0.16, p = .06$ ). Instead, providing the actionable explanation to reduce the loan amount led to a lower perceived attractiveness of the bank. In hindsight, this result makes sense. We can see that our participants imagined that they were asking for this amount of money for a reason. A bank that says that you should ask for less money may indeed sound unattractive. However, given that focusing on the features income versus loan amount led to different perceived actionability, given that there was a slightly positive (but not significant) effect of focusing on the more actionable feature on the perceived fairness of the decision, and given that perceived actionability was significantly positively correlated with, for instance, the attractiveness of the bank, we still think that there is reason to believe that focusing on more actionable features in explanations can lead to more favorable reactions to AI-based decisions. Thus, future research could investigate the effects of focusing on other features in actionable explanations that may seem less problematic than a bank telling you to ask for less money (e.g., in the context of loans: repayment duration, interest rate).

#### 4.1 Practical Implications

When using an AI-based system in a consequential context, it is worth considering to provide individuals who are confronted with a negative decision outcome with an explanation. Although overall perceptions are still likely to be rather negative due to the strong effect that negative decision outcomes have, providing explanations can at least buffer some of the negative reactions [15]. Eventually, explanations may even help to maintain some degree of a positive organizational image. It is particularly advisable to provide actionable explanations. However, it is important to consider whether providing explanations may conflict with other goals. For example, providing people with an actionable explanation could lead to making decision processes too transparent, enabling people to game the system [7]. Likewise, imagine providing an actionable explanation and having a person actually follow that explanation. If that person still does not get the desired outcome the next time they try, this can lead to particularly negative reactions.

Another implication of our study is to be aware of which feature is highlighted in an actionable explanation. In our specific case, highlighting that the applicant should ask for a lower amount of loan led to less favorable reactions – most likely because a bank that tells you to ask for less money does not sound like a bank you would ask for a loan again. However, even this explanation led to more positive reactions than not providing an explanation at all. Nevertheless, it may be advisable for system designers to enable decision recipients to provide information to the AI-based decision tool in order to personalize the explanation process [14]. For example, if loan applicants consider the loan amount to be unchangeable, it makes little sense for an explanation to suggest reducing the amount of loan requested. Instead, the explanation could focus on other features that may also provide ways forward for applicants. Alternatively, systems could also provide a variety of explanations with different ways forward [14].

## 4.2 Limitations

There are at least three limitations to our study that readers need to consider. First, this was a scenario-based study, there were no real consequences for the participants. Although more than half of our participants had experience applying for a loan and could therefore probably imagine being in the situation described, being denied a loan in reality can strongly impact people's lives.<sup>2</sup> Thus, the effects of explanations may be different from what we found in this study. This may be particularly true for actionable explanations. For people affected by AI-based decisions, the importance of receiving actionable explanations may only be fully realized in actual decision situations. Thus, we could hypothesize that the effects of actionable explanations are stronger for actual decision situations. Second, we focused only on the loan context. Although there are similarities in other consequential situations such as hiring (e.g., individuals apply for different positions, they receive rejections, they may receive explanations focusing on different features, these features may differ in their actionability), the current findings need to be replicated in other contexts in order to generalize our insights. Third, for the sake of simplicity, we chose to compare a 10% increase in income to a 10% decrease in the loan amount. This 10% itself could feel differently actionable for increasing income versus for decreasing the loan amount. Perhaps a 10% increase in monthly income would feel similar to a 30% decrease in loan amount. Nuances like these need to be tested in future studies.

## 4.3 Conclusion

Actionable explanations can be effective in mitigating unfavorable reactions to negative decision outcomes. Particularly in the case of consequential decisions, people want to know how they may be able to achieve a better outcome in future [13]. Our study showed

---

<sup>2</sup> Note that we tested whether including experience in applying for loan was a significant predictor in our analyses and whether it changed our results. It was not a significant predictor in any of the regressions. The only effect that was affected by including experience with applying for loan was that receiving an actionable explanation did not lead to significantly greater perceived procedural justice compared to receiving a feature relevance explanation.

that providing an actionable explanation led to better reactions to the decision process and to a better image for the company using the AI-based system to make its decision. Future research could further examine the specific needs that people have in these kind of decision situations with respect to the explanations they receive. Actionability seems to be one need, but we can also see others such as a personalization of the explanation process where people could inform the system about their specific situation and about what features would be more or less actionable for them.

**Acknowledgments.** This work was partially funded by the DFG grant 389792660 as part of TRR248 Center for Perspicuous Computing, project A6, and by the project “EIS – Explainable Intelligent Systems” funded by the VolkswagenStiftung as part of the grant 98513. We thank Rachit Singh for his support during data collection for this study.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Gaube, S., et al.: Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**(1), 31 (2021). <https://doi.org/10.1038/s41746-021-00385-9>
2. Langer, M., König, C.J., Papathanasiou, M.: Highly-automated job interviews: acceptance under the influence of stakes. *Int. J. Sel. Assess.* **27**(3), 217–234 (2019). <https://doi.org/10.1111/ijsa.12246>
3. Yalcin, G., Lim, S., Puntoni, S., van Osselaer, S.M.J.: Thumbs up or down: consumer reactions to decisions by algorithms versus humans. *J. Mark. Res.* **59**(4), 696–717 (2022). <https://doi.org/10.1177/00222437211070016>
4. Langer, M., Landers, R.N.: The future of artificial intelligence at work: a review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Comput. Hum. Behav.* **123**, Article 106878 (2021). <https://doi.org/10.1016/j.chb.2021.106878>
5. Starke, C., Baleis, J., Keller, B., Marcinkowski, F.: Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. *Big Data Soc.* **9**(2), 205395172211151 (2022). <https://doi.org/10.1177/20539517221115189>
6. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: ‘It’s reducing a human being to a percentage’; perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2018). <https://doi.org/10.1145/3173574.3173951>
7. Langer, M., et al.: What do we want from Explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* **296**, 103473 (2021). <https://doi.org/10.1016/j.artint.2021.103473>
8. Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C.: Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies. In: 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago IL USA, pp. 1369–1385. ACM (2023). <https://doi.org/10.1145/3593013.3594087>
9. Schoeffer, J., Kuehl, N., Machowski, Y.: ‘There is not enough information’: on the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1616–1628 (2022). <https://doi.org/10.1145/3531146.3533218>

10. Schlicker, N., Langer, M., Ötting, S.K., König, C.J., Baum, K., Wallach, D.: What to expect from opening ‘Black Boxes’? Comparing perceptions of justice between human and automated agents. *Comput. Hum. Behav.* (2021). <https://doi.org/10.1016/j.chb.2021.106837>
11. Lee, M.K., Jain, A., Cha, H.J., Ojha, S., Kusbit, D.: Procedural justice in algorithmic fairness: leveraging transparency and outcome control for fair algorithmic mediation. In: *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–26 (2019). <https://doi.org/10.1145/3359284>
12. Langer, M., König, C.J., Fitali, A.: Information as a double-edged sword: the role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Comput. Hum. Behav.* **81**, 19–30 (2018). <https://doi.org/10.1016/j.chb.2017.11.036>
13. Karimi, A.-H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. In: *Proceedings of the 2021 FAccT Conference on Fairness, Accountability, and Transparency, Virtual Event Canada*, pp. 353–362. ACM (2021). <https://doi.org/10.1145/3442188.3445899>
14. Singh, R., et al.: Directive explanations for actionable explainability in machine learning applications. *ACM Trans. Interact. Intell. Syst.* **13**(4), 1–26 (2023). <https://doi.org/10.1145/3579363>
15. Colquitt, J.A.: On the dimensionality of organizational justice: a construct validation of a measure. *J. Appl. Psychol.* **86**(3), 386–400 (2001). <https://doi.org/10.1037/0021-9010.86.3.386>
16. Lind, E.A., van den Bos, K.: When fairness works: Toward a general theory of uncertainty management. *Res. Organ. Behav.* **24**, 181–223 (2002). [https://doi.org/10.1016/S0191-3085\(02\)24006-X](https://doi.org/10.1016/S0191-3085(02)24006-X)
17. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. *arXiv* (2018). <https://doi.org/10.48550/arXiv.1812.04608>
18. Bauer, T.N., Truxillo, D.M., Sanchez, R.J., Craig, J.M., Ferrara, P., Campion, M.A.: Applicant reactions to selection: development of the selection procedural justice scale (SPJS). *Pers. Psychol.* **54**, 387–419 (2001). <https://doi.org/10.1111/j.1744-6570.2001.tb00097.x>
19. Colquitt, J.A., Rodell, J.B.: Measuring justice and fairness 8. In: *The Oxford Handbook of Justice in the Workplace*, vol. 1, p. 187 (2015)
20. Highhouse, S., Lievens, F., Sinar, E.F.: Measuring attraction to organizations. *Educ. Psychol. Measur.* **63**, 986–1001 (2003). <https://doi.org/10.1177/0013164403258403>
21. Franke, T., Attig, C., Wessel, D.: A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *Int. J. Hum.-Comput. Interact.* **35**(6), 456–467 (2019). <https://doi.org/10.1080/10447318.2018.1456150>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# From Explanation Correctness to Explanation Goodness: Only Provably Correct Explanations Can Save the World

Maike Schwammler<sup>(✉)</sup> 

Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany  
schwammler@kit.edu  
<https://mase.kastel.kit.edu>

**Abstract.** Explainability Engineering gets evermore important in the era of self-learning and automated systems. We motivate the necessity for interdisciplinary research to engineer verifiably correct and good explanations: Systems engineering research must ensure that correct and machine-understandable explanations can be derived from system specifications and social sciences research must ensure that a context-dependent and stakeholder-tailored explanation can be provided in a fitting manner. We describe our first steps in the direction of a holistic and interdisciplinary explainability engineering process for tackling these challenges.

**Keywords:** self-explainable software systems · explanation correctness · explanation goodness · eXplainable Artificial Intelligence · trustworthy systems

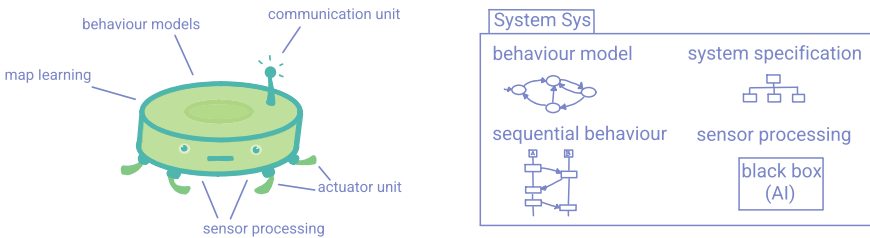
## 1 Introduction

Automated and self-learning software systems are increasingly used in a variety of domains and in people's everyday life: from driving assistance systems and products manufactured in smart factories to smart home technologies and applications on our smartphones. Often, the level of automation and system functionality might be known to stakeholders interacting with the automated system to some degree. For instance, an owner of a semi-automated vehicle will have a certain degree of knowledge about the automated distance keeping functionality of their car. However, there still might be features that they do not understand, e.g. the car's behaviour in some special outlier situations (e.g. how the car operates in exceptionally bad weather).

We perceive two key reasons why the design and functionality of such automated and self-learning software systems must be made explainable. (a) Relevant stakeholder groups that interact with such systems need to be sufficiently informed about the systems' functionality. E.g., to be enabled to safely interact with the system or for trusting the automated system. Secondly, (b) for an explainable system, analysis and verification of correct system behaviour can be

aided. We postulate that, without a certain level of system explainability, a system should not be launched into our markets and with that be integrated into our societies (cf. IEEE Standard for Transparency of Autonomous Systems [2, 27]). We develop and investigate a system’s *self-explainability* capabilities. With self-explainability, a system can explain its decision making process without the help of an external explainer.

A challenge for engineering self-explainability is the sheer number and complexity of components that a system comprises: Even a (seemingly) simple system like a robot vacuum cleaner comprises a collection of different software components (see Fig. 1). These could, e.g., be a sensor processing unit for avoiding collisions, a communication unit for interacting with a connected smart home system, a behaviour model for specifying when the robot must return to it’s charging station and a sub-symbolic Artificial Intelligence (AI) component that learns a map about the cleaning area. Due to this level of complexity, having a system engineer write explanations manually cannot be desirable nor feasible and would certainly lead to human errors. This would result in unreliable, even incorrect, explanations. Further on, with an increase in sub-symbolic AI components that learn new phenomena during run-time, not all explanations can be built during design time. Due to this, we focus on automatically extracting provably *correct explanations* from system models. In our understanding, a correct explanation is one where the validity of explanation content can be proven through system analysis techniques. Such extracted explanations would be in an internal, machine-readable, format, allowing for formal reasoning about explanation correctness by the system itself. A benefit of this is model-reuse: the necessary system models are built during the software system design process, with examples for system models including communication diagrams, automata models and architecture models [3].

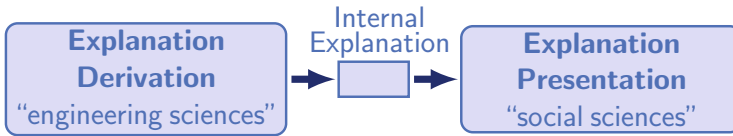


**Fig. 1.** A robot vacuum cleaner with diverse functionality, defined as a system Sys, comprising several software system components.

However, purely considering the formal correctness of an explanation is not enough: *explanation goodness* [12] is a factor that must be considered. Explanation goodness means that explanations must be actually helpful in increasing properties like understanding or trust into the system for specific explanation recipients. One explainability research consensus comes to the fore in explainability research discussions: only *verifiably good explanations* can be successful

in increasing system understandability [10,16,28]. We argue in this paper, that, to gain such a good explanation, both explanation goodness and explanations correctness must be analysed and validated.

For proving both the goodness and correctness of an explanation, a variety of research areas must meet and join forces (see Fig. 2); On the one side, a technical explanation must be derived in the right moment [4,21] and from adequate sources (e.g. system models, environment models, mental models) [6,23]. Such a technical explanation will be in some internal, machine-readable, format (e.g. a logical expression), allowing the autonomous system itself to reason about explanations. This step might be provided through work from the engineering sciences, e.g. computing science. On the other side, this technical explanation must be presented adequately to fit the needs of specific types of recipient stakeholders, in varying contexts [14]. For such an adequate presentation of an explanation, an interdisciplinary viewpoint on explainability is the key, especially taking knowledge from social sciences into account [14,15,18]. Note that the graphic presented in Fig. 2 is a very coarse segmentation of research directions necessary for engineering correct and good explanations. It suits the needs of this paper, but a finer segmentation is of interest for a more detailed topic discussion in future work.



**Fig. 2.** To engineer correct and good explanations, technical and interdisciplinary research on explanations must be brought together.

In this paper, we summarise our roadmap for a *holistic explainability engineering process* from explanation correctness to explanation goodness in Sect. 2. In our current research, the key argument is that we need automatically derived and formalised explanations to enable a formal validation of explanation correctness (cf. Sect. 3). However, it would be presumptuous to believe that we can fully validate explanations on a purely logical level. Hence, we point out necessary interdisciplinary steps towards explanation goodness (cf. Sect. 4). We summarise our key findings towards a *holistic explainability engineering process* in Sect. 5.

## 2 Holistic Explainability Engineering

Our overall vision is to enable autonomous systems to self-explain their behaviour and functionality. Generally, the systems that need explaining comprise several *system components* as we discussed for the robot vacuum cleaner



from Fig. 1. We postulate that explanations must be providable for the entire system, and not only for a singular system component. For instance, an end-user might want to know why their robot vacuum cleaner did not clean a specific room. The reasons for this must be found in the entire system and not just within a singular component: It might be that due to a defective sensor, the map learning was not fully successful. Another explanation could be that the robot was not capable of moving over a cable laying in the threshold of the room. Such a need for holistic system explanations is also discussed in [19], where the authors argue that a majority of approaches within the eXplainable Artificial Intelligence (XAI) community only considers explanations for specific AI algorithms and for the community itself, meaning experts. While this can help said experts in debugging a system, it does not meet the explainability requirements for most non-expert stakeholders (e.g. end-users, lawyers, regulation bodies, ...) [5, 8]. The authors of [9] go even further than [19] by showing that most XAI approaches focus on “low-level”, narrow, explanations, while instead we need to go towards “high-level”, strong, explanations. They call their approach “Broad eXplainable Artificial Intelligence (Broad-XAI)”. Their approach entails to map explanations derived by XAI approaches to human models of explanation, thereby also arguing for the necessity of connecting the technical explanation derivation with explanation presentation (cf. Fig. 2).

Our own research towards holistic explainability engineering does not only focus on sub-symbolic AI models, but instead takes the entire system into account, with both symbolic and sub-symbolic components. Our work can be boiled down to two key hypotheses, that can be associated with each one of the research areas we motivate in Fig. 2:

**Explanation Derivation** To verify correctness of an explanation, we must consider the logical core of explanations and their formal source.

**Explanation Presentation** To connect our formal notion of explanation correctness to actual goodness of explanations, we must enrich our research with interdisciplinary expertise, e.g. from social sciences.

We give more details on both hypotheses in the following sections.

### 3 Explanation Derivation

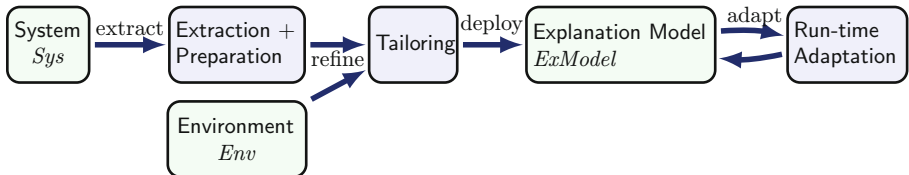
A precondition for engineering a self-explainable system is to consider how the system reasons about its behaviour; namely in some formalised machine code, following its system description. There is a striking benefit of first considering machine code, instead of natural language, for explanations: the inherent ambiguities of natural language are not existing in the machine code counterpart [22]. Following this reasoning, the logical core  $E$  of an explanation that we consider is a machine-readable and machine-producible intermediate format of an explanation. Its formal source are artefacts from system development processes: system models  $Sys$  (e.g. architecture diagrams, communication protocols, ...) and environment models  $Env$  (describing the operating context). We assume that  $Env$

also includes mental models for human behaviour (e.g. derived from cognitive architectures [1]). We define explanation correctness as follows.

**Definition 1. (*Explanation Correctness*)** *An internal explanation  $E$  is correct, if it can be deduced for an explanandum  $X$  and from provably correct system models  $Sys$  and environment descriptions  $Env$ .*

Here, the *explanandum*  $X$  describes the system phenomenon that needs explaining. This could, e.g., be the driving assistance function of an automated car. It is not the goal of this paper, to provide a full formalisation of explainability concepts, but we refer to [13] for this. Through Definition 1, we can conclude explanation correctness from correctness of specification models. Considering this formal core of an explanation comes with a striking benefit: We can use known techniques from formal verification to analyse explanations and prove their correctness.

Deriving the logical core of an explanation from system descriptions is not a trivial step. Instead, we must develop means on how to extract explanations from system models. We describe a reference framework on how to automatically derive *explanation models* from system models in previous work [23] and give a simplified visualisation for it in Fig. 3. An explanation model could, for instance, be a causal behaviour tree [11]. The starting point for the framework is a system specification  $Sys$ . By analysing potential causes for system actions within  $Sys$ , we extract an initial explanation model. We extract the explanation model from the different system components contained within  $Sys$ . This leads to several explanation models, which are combined to one explanation model in this phase to avoid redundancy and computational overhead. We give details on this process of merging explanation models in our previous work [17]. Through information provided by environment  $Env$ , we refine the initial explanation model in the tailoring phase. The output of the tailoring phase is an explanation model which is tailored toward specific recipient stakeholders. The assumption for this is that an end-user stakeholder will need different information from an explanation than an expert stakeholder. In the era of self-learning adaptive systems [26], we must also take updates of the explanation model into account. This is needed, if, for instance, the vacuum cleaner encounters a new type of obstacle to avoid or is confronted with a new type of surface to clean.



**Fig. 3.** Simplified visualisation of the reference framework from [23] on the derivation of an explanation model from a formalised system model. The green elements are the artefacts used and created within the blue-coloured phases. (Color figure online)

We depict an exemplary explanation model as of [23] in Fig. 4. A technical explanation extracted from this model can be an *explanation path* within the tree structure. An example for such an explanation path could be  $because(C, and(x_2, y_2))$ . In case of the vacuum cleaner example, instances of actions could be “emergency brake” or “drive to station” with potential reasons “out of energy”, “obstacle in the way” and “sensor obscured”. The technical explanation  $because(C, and(x_2, y_2))$  could thus translate to “An emergency brake ( $C$ ) was done because of an obstacle in the way ( $x_2$ ) and an obscured sensor ( $y_2$ )”.

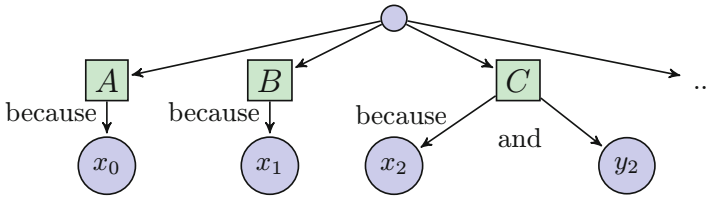


Fig. 4. An exemplary explanation model, comprising reasons for actions  $A$ ,  $B$  and  $C$ .

The idea of explanation models as explanation source is embedded into the MAB-EX framework for self-explainable systems that we developed in previous work [6]. MAB-EX suggests to identify the need for an explanation through system monitoring and analysis. Whenever the need for an explanation is detected, an explanation path must be extracted from our explanation model. We refine MAB-EX in [25] by integrating *levels of explainability* that were suggested in [5]. With this, explanations can be provided by a singular or multiple systems. This enables that explanations can be provided even if the system itself has an incomplete explanation model, e.g. through missing environmental information. The missing information is then provided by another system. Equally, we must take the explanation context (i.e. where and what to explain) [5] and timing (i.e. when to explain) [4, 21] into account to provide holistic explanations.

Hitherto, the framework from [23] is only conceptual and has been instantiated exemplarily for an autonomous driving controller in [23] and for digital twins in [17]. Our next steps are to examine different types of system and environment models from which explanation models must be derived, and how to formalise an automated explanation model extraction process for a variety of system models.

## 4 Explanation Presentation

For our goal of a holistic explainability engineering process, we must now tackle the challenge of “informalising” our logical explanation path [22]. From it, we

must achieve an explanation presentation that follows explanation requirements for diverse explanation recipient groups  $G$  [8, 14]. This means that now, from a *correct* explanation, we must derive a *good* explanation.

**Definition 2. (*Explanation Goodness*)** *An explanation  $E$  can be labelled as good, if it stems from a correct explanation  $E_c$  and measurably helps a recipient group  $G$  in understanding the explanandum  $X$ .*

So, explanation goodness depends on different types of recipient stakeholders. To measure explanation goodness, *explanation validation* through user-studies must be considered. As it would be out of the scope of this visionary paper, we do not go into details on this, but instead refer to two review papers on explanation validation through user studies [14, 20]. We tailor our explanation model towards different recipient groups in the respective phase of our framework (cf. Fig. 3). Situation- and stakeholder-dependent explanations have various benefits to different stakeholders, for which we provide examples here:

- a system engineer can improve and debug the system during design time and
- an end-user is enabled to use the system safely and to justifiably trust in its automated decisions,
- political and societal bodies can decide whether to allow a system to be launched into the markets, and
- lawyers can decide who is to be blamed in tort claim cases that involve automated and self-learning systems.

To allow for explanations to be used on such central societal levels, a holistic explanation validation process is of the utmost importance. Such an explanation validation must contain the formal verification part that we discussed in the previous section, but also needs a user-dependent validation to assess explanation goodness [12]. One of the shortcomings that [14] discover in an extensive literature review is the lack of empirical evidence for explanation goodness.

In our research, we aim to approach this problem from two sides: We need to investigate what information explanations for different recipient stakeholders must contain, to allow for a correct tailoring of our explanation model towards recipients. This is necessary as an explanation model containing too much information leads to larger computation times for deriving explanation paths from the model, and an explanation model missing information would lead to phenomena that cannot be explained at all. We started this endeavour by investigating explanations needed for lawyers in [7]. In sum only verifiable explanations can help stakeholders like lawyers, courts and regulation bodies to assess liabilities and to admit systems into the markets. Further research includes a translation of our internal, technical, explanations into adequate presentation formats to investigate and formalise explanation goodness on top of explanation correctness.

## 5 Conclusion

We motivate the need for a connection of formal reasoning about explanations with research from social sciences to validate both explanation correctness and

explanation goodness. We argue that trustworthiness of explanations can only be reached through a holistic explainability engineering process. Such a holistic explainability engineering process entails that the entire system, with all its components and together with environmental influence factors, must be explained. Moreover, different recipient stakeholders and their specific explanation requirements must be taken into account. This makes reasoning about explanation correctness and goodness a complex endeavour, where a challenging amount of requirements and factors must be taken into account.

For tackling this challenge, we recently suggest requirements for explainability levels [24], dividing explanations into local and global explanations. We also discuss a notion of *explanation quality* in that paper. Equally, through tailoring our explanation model from [23] towards different stakeholders, we do not intend to formalise one model that must comprise explanations for all possible stakeholders. Instead, each one explanation model exists for each group of stakeholders. The motivation for this is that different stakeholders might need very different types and degrees of information within an explanation model. Even with such steps for decreasing complexity in explainability engineering, one should not assume to be able to formalise and formally prove each aspect of explanation correctness and goodness right away, for each stakeholder group and varying application domains. Instead a focus on specific application domains and a fixed amount of stakeholder groups certainly makes sense as a starting point.

We follow the argumentation of [9, 14, 19] and emphasise that a focus of the explainability community must continue to shift from expert explanations for narrow and highly isolated system functions towards holistic explainability research for complex systems of systems. Several interdisciplinary challenges and starting points for doing so have been summarised in [15]. One challenge particularly to be overcome is the derivation of explanations for probabilistic AI systems into approaches of explainability engineering, and we sketch some steps for that in [24].

**Acknowledgement.** This research was supported by the Innovation Campus for Future Mobility ([www.icm-bw.de](http://www.icm-bw.de)) and by the Helmholtz Association within the Core Informatics project.

## References

1. The Cambridge Handbook of Computational Psychology. Cambridge Handbooks in Psychology, Cambridge University Press, Cambridge (2008)
2. IEEE standard for transparency of autonomous systems: IEEE Std **7001–2021**, 1–54 (2022). <https://doi.org/10.1109/IEEESTD.2022.9726144>
3. Iso/iec/ieee 42010:2022 software, systems and enterprise architecture description **2**, 1–62 (2022). <https://www.iso.org/standard/74393.html>
4. Bairy, A., Hagemann, W., Rakow, A., Schwammbberger, M.: Towards formal concepts for explanation timing and justifications. In: 30th IEEE International Requirements Engineering Conference Workshops, RE 2022 - Workshops, Melbourne, Australia, 15–19 August 2022, pp. 98–102. IEEE (2022). <https://doi.org/10.1109/REW56159.2022.00025>

5. Bersani, M.M., Camilli, M., Lestingi, L., Mirandola, R., Rossi, M., Scandurra, P.: A conceptual framework for explainability requirements in software-intensive systems. In: 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), pp. 309–315 (2023). <https://doi.org/10.1109/REW57809.2023.00059>
6. Blumreiter, M., et al.: Towards self-explainable cyber-physical systems. In: 22nd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion, pp. 543–548 (2019). <https://doi.org/10.1109/MODELS-C.2019.00084>
7. Buiten, M.C., Dennis, L.A., Schwammberger, M.: A vision on what explanations of autonomous systems are of interest to lawyers. In: Schneider, K., Dalpiaz, F., Horkoff, J. (eds.) 31st IEEE International Requirements Engineering Conference, RE 2023 – Workshops, Hannover, Germany, 4-5 September 2023, pp. 332–336. IEEE (2023). <https://doi.org/10.1109/REW57809.2023.00062>
8. Chazette, L., Brunotte, W., Speith, T.: Explainable software systems: from requirements analysis to system evaluation. *Requir. Eng.* **27**(4), 457–487 (2022). <https://doi.org/10.1007/s00766-022-00393-5>
9. Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., Cruz, F.: Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artif. Intell.* **299**, 103525 (2021). <https://doi.org/10.1016/j.artint.2021.103525>, [www.sciencedirect.com/science/article/pii/S000437022100076X](http://www.sciencedirect.com/science/article/pii/S000437022100076X)
10. de Bruijn, H., Warnier, M., Janssen, M.: The perils and pitfalls of explainable AI: strategies for explaining algorithmic decision-making. *Government Inf. Quart.* **39**(2), 101666 (2022). <https://doi.org/10.1016/j.giq.2021.101666>, [www.sciencedirect.com/science/article/pii/S0740624X21001027](http://www.sciencedirect.com/science/article/pii/S0740624X21001027)
11. Garcia, F.J.C., Robb, D.A., Liu, X., Laskov, A., Patrón, P., Hastie, H.F.: Explain yourself: a natural language interface for scrutable autonomous robots. *CoRR abs/1803.02088* (2018), <http://arxiv.org/abs/1803.02088>
12. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* **5** (2023). <https://doi.org/10.3389/fcomp.2023.1096257>
13. Köhl, M.A., Baum, K., Langer, M., Oster, D., Speith, T., Bohlender, D.: Explainability as a non-functional requirement. In: RE, pp. 363–368. IEEE (2019)
14. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K.: What do we want from explainable artificial intelligence (XAI)? - a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary xai research. *Artif. Intell.* **296**, 103473 (2021). <https://doi.org/10.1016/j.artint.2021.103473>. <https://www.sciencedirect.com/science/article/pii/S0004370221000242>
15. Longo, L., et al.: Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* **106**, 102301 (2024). <https://doi.org/10.1016/j.inffus.2024.102301>, <https://www.sciencedirect.com/science/article/pii/S1566253524000794>
16. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Inf.* /113, 103655 (2021). <https://doi.org/10.1016/j.jbi.2020.103655>, [www.sciencedirect.com/science/article/pii/S1532046420302835](http://www.sciencedirect.com/science/article/pii/S1532046420302835)

17. Michael, J., Schwammbberger, M., Wortmann, A.: Explaining cyberphysical system behavior with digital twins. *IEEE Softw.* **41**(1), 55–63 (2024). <https://doi.org/10.1109/MS.2023.3319580>
18. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019). <https://doi.org/10.1016/j.artint.2018.07.007>, [www.sciencedirect.com/science/article/pii/S0004370218305988](http://www.sciencedirect.com/science/article/pii/S0004370218305988)
19. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences (2017)
20. Nauta, M., et al.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput. Surv.* **55**(13s) (2023). <https://doi.org/10.1145/3583558>
21. Rakow, A., Hajnorouzi, M., Bairy, A.: What to tell when? - information provision as a game. In: Farrell, M., Luckcuck, M., Gleirscher, M., Schwammbberger, M. (eds.) *Proceedings Fifth International Workshop on Formal Methods for Autonomous Systems, FMAS@iFM 2023, Leiden, The Netherlands, 15th and 16th of November 2023. EPTCS*, vol. 395, pp. 1–9 (2023). <https://doi.org/10.4204/EPTCS.395.1>
22. Ranta, A.: Translating between language and logic: what is easy and what is difficult. In: Bjørner, N., Sofronie-Stokkermans, V. (eds.) *CADE 2011. LNCS (LNAI)*, vol. 6803, pp. 5–25. Springer, Heidelberg (2011). [https://doi.org/10.1007/978-3-642-22438-6\\_3](https://doi.org/10.1007/978-3-642-22438-6_3)
23. Schwammbberger, M., Klös, V.: From specification models to explanation models: an extraction and refinement process for timed automata. In: Luckcuck, M., Farrell, M. (eds.) *Proceedings Fourth International Workshop on Formal Methods for Autonomous Systems (FMAS) and Fourth International Workshop on Automated and verifiable Software sYstem DEvelopment (ASYDE), FMAS/ASYDE@SEFM 2022, and Fourth International Workshop on Automated and verifiable Software sYstem DEvelopment (ASYDE)Berlin, Germany, 26th and 27th of September 2022, EPTCS*, vol. 371, pp. 20–37 (2022). <https://doi.org/10.4204/EPTCS.371.2>
24. Schwammbberger, M., Mirandola, R., Wenninghoff, N.: Explainability engineering challenges: connecting explainability levels to run-time explainability. In: Longo, L., Lapuschkin, S., Seifert, C. (eds.) *Explainable Artificial Intelligence, xAI 2024, CCIS*, vol. 2156, pp. 205–218. Springer, Cham (2024). [https://doi.org/10.1007/978-3-031-63803-9\\_11](https://doi.org/10.1007/978-3-031-63803-9_11)
25. Schwammbberger, M., Mirandola, R., Wenninghoff, N.: Explainability engineering challenges: from requirement definition to run-time explainability (2024), submitted to 2nd World Conference on eXplainable Artificial Intelligence (XAI)
26. Weyns, D., Iftikhar, M.U., de la Iglesia, D.G., Ahmad, T.: A survey of formal methods in self-adaptive systems. In: *Proceedings of the Fifth International C\* Conference on Computer Science and Software Engineering, C3S2E 2012*, pp. 67–79. Association for Computing Machinery, New York, NY, USA (2012). <https://doi.org/10.1145/2347583.2347592>
27. Winfield, A.F.T., et al.: IEEE p7001: a proposed standard on transparency. *Front. Robot. AI* **8**, 225 (2021). <https://doi.org/10.3389/frobt.2021.665729>
28. Wing, J.M.: Trustworthy AI. *Commun. ACM* **64**(10), 64–71 (2021). <https://doi.org/10.1145/3448248>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.







# Thinking Outside the Box?

## Regulatory Sandboxes as a Tool for AI Regulation

Hannah Ruschemeier<sup>(✉)</sup>

Fernuniversität Hagen, Universitätsstraße 27, 58084 Hagen, Germany  
hannah.ruscheimer@fernuni-hagen.de

**Abstract.** This legal article deals with the concept of regulatory sandboxes as part of experimental regulation to meet the challenges of digitalisation for legal regimes. It focuses on a critical analysis of the new provisions of the AI Act and their weaknesses in relation to the objectives of regulatory sandboxes to promote innovation and regulatory learning. Finally, suggestions for improvement and future areas of application are outlined.

**Keywords:** legal regulation · AI Act · regulatory sandboxes · innovation · regulatory learning · EU law

## 1 Introduction

### 1.1 Regulating Emerging Technologies

Digital technologies, in particular algorithm-based decision-making systems discussed under the term “artificial intelligence” (AI), pose special challenges to the concept of legal regulation. In addition to the characteristic of irreversibility, certain digital technologies represent a qualitative leap in that they lack an analogous counterpart in the physical realm. Predictions generated by algorithms are based on the analysis of vast quantities of data and thus require automated processes exclusive to the digital environment. The need for legal guardrails as a response seems undisputed: the internet is supposedly a legal vacuum, [1] the prospect of deep-fakes and misinformation looms as a dystopic potential outcome, [2] AI-technology is used in war depending on the favour of private companies [3]. On the government side, there is uncertainty and a lack of knowledge about what is regulated and what is not. Consequently, there is a great need for future-oriented and resilient legal frameworks [4].

The question of whether to regulate does little to answer the difficult details of the “how”, which begins with the procedural aspects. The development of technology and legislative processes often exhibit stark differences since the law in the form of legislation and technology move at different paces. On the one hand, the development of digital technology is flexible, dynamic, and progressing at a rapid pace. Law, on the other hand, especially in the form of legislation, is slow, reflecting the fact that negotiating compromises and thus majorities in democratic processes takes time. On

the contrary, technical innovations are a “moving target”, which challenges the often reactive functioning of the law [5]. The declared aim of political initiatives is therefore to create “robust”, “future-proof” and “flexible” legal requirements.<sup>1</sup> Finding the right way forward is complex: abandoning regulation by law is not an option; entire laws cannot be changed in anticipatory obedience without closely looking at potential impacts. This dilemma creates a need for flexible but effective regulation, which could be met by experimental regulatory instruments, such as experimental clauses, regulatory sandboxes and temporary laws and evaluations [6]. For example, the discussion about banning Bitcoin at the European level [7], shows how difficult the legal struggle for the right answers to disruptive technologies is. Regulatory projects are often (unfairly) accused of stifling innovation [8, 9]. This is very present in the political debate, and many believe there is a risk of disproportionate requirements that will stifle innovation, pushing it out into other less stringent jurisdictions [10].

Instruments of experimental regulation are supposed to contribute solutions to these conflicts. In general, experimental regulation combines empirical evidence and legal requirements in a more flexible model than “traditional” legislation [11]. It foresees the testing of innovative products and services are to be tested in a supervised environment in close cooperation with the competent supervisory authority within a time-limited framework, often with the application of substantive legal exceptions or no enforcement letters [12]. The aim is to promote innovation, as well as to generate expert knowledge on the government side in order to adapt the regulatory framework or to gain insights for new administrative and legislative processes. Experimental regulation is not new, but remains relatively unexplored from a legal perspective in the field of regulating digital technologies [13–15]. Instruments of experimental regulation, which have been underresearched in the context of law and new technologies [12, 16, 17], could contribute to a solution by providing a flexible framework for the generation of state regulatory knowledge while also providing processes to promote innovation. Regulators can use real-world laboratories to create a testing ground for new technologies in which real-world legal requirements do not need to be enforced during an experimental phase, thus providing insights into the object of regulation [18].

## 1.2 Experimental Regulation: Regulatory Sandboxes

Regulatory sandboxes are part of the toolbox of new epistemic methods [12, 16]. The AI Act now explicitly calls for the establishment of regulatory sandboxes as a measure to promote innovation in every member state on a national level, Articles 57 et seq [19]. The AI Act defines regulatory sandboxes as a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate, and test innovative AI systems, where appropriate in real-world conditions, pursuant to a sandbox plan for a limited time under regulatory supervision,

---

<sup>1</sup> Cf. Recital 138 EU Regulation 2024/0138 European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206-C9-0146/2021-2021/0106(COD)) (AI Act).

article 3 (55). Recital 138 explicitly mentions the idea of regulatory sandboxes as part of a legal framework that promotes innovation and is future-proof and resilient.

Furthermore, Recital 139 states that the aim is to support innovation and regulatory learning and counteract the information deficit in relation to the regulatory subject of “AI systems”, Article 3 (1).<sup>2</sup> Sandboxes could provide for a controlled environment that fosters innovation and facilitates the development, training, testing, and validation of innovative AI systems for a limited time before release onto the market or put into service pursuant to a specific sandbox plan agreed between the prospective providers and the competent authority.<sup>3</sup> Pursuant to Recital 139, participation in the regulatory sandbox should focus on issues that raise legal uncertainty for providers and prospective providers, allowing them to innovate, experiment with AI in the Union, and contribute to evidence-based regulatory learning.

In highly complex, technical areas, there is often a knowledge gap between the regulatory authorities and the specifics of the field to be regulated, specifically because *ex ante*, the unpredictability of these highly complex, dynamic technologies often eludes traditional forecasting schemes for hazard prevention, risk regulation, prohibition with reservation of authorisation, and *ex ante* risk assessment. In these fields there is a structural lack of information, as seen in the fintech-sector [20, 21]. In addition, as evolving and disruptive technologies have a cross-sectional effect and touch on different areas and fields of law, their impact often simply cannot be assessed [22]. This is why there is a growing call for new legal instruments that can cope with digital transformation. Given the negative effects already apparent, a failure to react until all effects of new technologies are fully known or making wholesale changes to existing law is clearly undesirable.

Nevertheless, regulatory sandboxes cannot and should not replace effective regulation or create large-scale exemptions from it. The focus at the regulatory level should therefore continue to be on creating effective regulatory requirements in the digital sector that protect fundamental rights and promote the public good. The AI Act is a first step in the right direction, but it still contains many gaps and inadequate provisions [23].

This paper examines regulatory sandboxes and related concepts from a public law regulatory perspective and adds to the ongoing debate, which primarily discusses regulatory sandboxes from the perspective of promoting innovation and the economy [16, 17, 21]. The potential for regulatory learning on the part of the supervisory authorities

<sup>2</sup> The AI Act defines AI-System as “a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and 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”.

<sup>3</sup> Recital 139 states: The objectives of the AI regulatory sandboxes should be to foster AI innovation by establishing a controlled experimentation and testing environment in the development and pre-marketing phase with a view to ensuring compliance of the innovative AI systems with this Regulation and other relevant Union and national law, to enhance legal certainty for innovators and the competent authorities’ oversight and understanding of the opportunities, emerging risks and the impacts of AI use, to facilitate regulatory learning for authorities and undertakings, including with a view to future adaptations of the legal framework, to support cooperation and the sharing of best practices with the authorities involved in the AI regulatory sandbox, and to accelerate access to markets, including by removing barriers for SMEs, including start-ups.

is another important but underrepresented consideration and an explicit goal of the AI Act, Article 57 (9 d).

### 1.3 Concept of Regulatory Sandboxes

The term ‘sandbox’ used in the AI Act originates from the field of computer science. In that field, sandboxes are generally referred to as isolated or quarantined areas where the effects are contained and do not affect the infrastructure, platform or business-critical networks on which they run [24]. The idea of regulatory sandboxes has similarities and differences. On the one hand, sandboxes can be specifically designed to provide a framework for promoting innovation by creating a ‘safe space’ for testing new business ideas within an existing or newly planned regulatory regime [25]. On the other hand, the aim of regulatory sandboxes is precisely to generate external effects and consequences for observation. At an abstract level, the overall aim of the sandbox is to gain knowledge of unknown facts, consequences and risks in all areas.

The broader concept of sandboxes, where different scenarios are tested in a real environment, comes from the social sciences. Similar to the idea of sandboxes in computer science, they act for example as test spaces for sustainable urban development [26]. However, not every real-world lab is a regulatory sandbox. Many real-world labs simply combine science and practice without necessarily leading to legal issues or the involvement of a regulator.<sup>26</sup> These sandboxes are not included in the following analysis.

Regulatory sandboxes are also referred to as ‘real laboratories’, or rather derisively as ‘sandbox playgrounds’. Compared to the sandbox concept in social science, regulatory sandboxes are narrower in scope and mostly aim at improving innovation and regulation. Regulatory sandboxes therefore involve a regulator and a regulated party. They act as test rooms for innovation and regulation at the same time, allowing the testing of technologies, services, products or approaches that are only partially compliant with existing legal and regulatory frameworks, or where compliance is unclear. The first sandbox-style framework was established by the US Consumer Financial Protection Bureau (CFPB) in 2012, dubbed ‘Project Catalyst’. This programme aimed to foster the development and expansion of innovative consumer financial products and services. The objective was to collaborate with the community of innovators to ensure that superior financial products and services were accessible to American consumers [25]. Even if the CFPB did not formally announce their initiatives as a sandbox, it followed policies such as No Action Letters<sup>4</sup> that fulfilled many functions of a sandbox [27].

Outside the US, the UK’s Financial Conduct Authority (FCA) has been a pioneer of the specific concept of regulatory sandboxes, testing hundreds of applications in its sandbox since 2016 [28]. In its sandbox, the FCA follows the concept of a market-driven regulatory sandbox with annual ‘cohorts’ selected from a general pool of aspiring innovators. Different forms of regulatory sandboxes are now being discussed. In contrast to general purpose sandboxes, thematic sandboxes are designed to pursue specific policy objectives that are thematically limited [25]. This is usually done by limiting them to specific technologies, products, or business models.

<sup>4</sup> No-action letters are designed to recognise the value of innovative financial technologies by committing the CFPB to take no enforcement or supervisory action with respect to the subject matter of the letter, Information Collection, 81 Fed. Reg. 8686, 8692 (Feb. 22, 20169).

## 1.4 Legal Context

By creating a normative framework for new regulatory approaches, the law can make a decisive contribution to proactively shaping developments in digital transformation. Experimental regulation can be an important aspect of the development of the law. Future-proof regulation requires flexibility on the one hand, and legal certainty and resilience on the other. Digital technologies are limitless, which makes it all the more important to make the best use of existing legal possibilities in the interests of democracy, the rule of law, and the common good.

In general, regulatory requirements, like all government action, are based on the law and the constitutional principle of proportionality (in the EU and all Member States); restrictions and requirements must be set in relation to the potential damage to legitimate interests worth protecting, e.g. consumer protection, security, etc. However, there is already an information deficit in many digital technologies. As a result, the potential damage, or more precisely the potential risk, is not known. Different risk levels are reflected in national and European legislation: pharmaceuticals have to undergo an approval procedure, while operating a permanent business only requires notification. The idea of adapting traditional regulatory structures is obvious, as the “analogue” effects of digital technologies are more difficult to grasp. For example, when Facebook was launched, few people foresaw its development into one of the most powerful companies in the digital economy with an impact on political processes, elections and other public opinion-forming activities.

Significantly, the matter at hand concerns a *de facto* regulation of something whose evolution is not yet known. This poses challenges for both sides, the product owner or developer and the relevant regulator. Structural uncertainty is not new to law, especially public law, most recently gaining public awareness during the pandemic. Unlike other related disciplines such as social and political sciences, experiments are not part of the tradition of administrative science and administrative law. In addition, highly regulated areas of law have often been developed for the interface of market access and thus do not correspond to the reality of digital products, which continue to evolve dynamically once they are already on the market. Regulatory sandboxes therefore offer a first opportunity to gain information and knowledge for regulators and participants at the same time. On the other hand, it is important to ensure that regulatory sandboxes should not become a *carte blanche* for risky products of dubious legality. While it should not be forgotten that product providers can design their products to comply with the applicable law, regulatory sandboxes do provide an opportunity for supervisors to gain valuable knowledge about the products they regulate while also allowing for the legal framework to be adapted to current developments. In all innovation efforts, the limits of the separation of powers must be kept in mind. Although the legislator has the democratically legitimacy to decide on changes to the legal framework, the knowledge gained from a sandbox can only be provided by the administration. The protective purpose of the regulation is crucial: if this is not achieved *de lege lata*, the compatibility with the existing requirements is of little use. A combination of the two approaches therefore seems promising.

Regulatory sandboxes are intended to mitigate the the danger of regulation consistently trying to catch up with current developments and new legal requirements being

outdated even before they are adopted. Their application to matters involving an information deficit is less clear. Information deficits exist, for example, in decision-making systems where the path from input to output is not 100% traceable, either because of the volume of data or the complexity of the calculation. This can be addressed in a number of ways, most commonly through a general ban or general authorisation of such technologies, through an ex-ante authorisation procedure as in pharmaceutical law, or through a ban subject to authorisation. Regulatory sandboxes could come into play to reduce the information deficit in all variations, but should not aim at deregulation or at reducing protection standards. Instead, sandboxes should serve to create an appropriate regulatory framework to bridge the gap between new technologies and regulations that may have been developed prior to the invention of the smart phone, let alone the app being developed.

In other countries, regulatory sandboxes have already established themselves as an instrument in highly regulated areas such as the financial sector. So far, real-world labs have primarily been discussed in the context of financial market regulation, fintechs, blockchain, or crypto-assets. In Germany, they are a comparatively new phenomenon and not particularly popular: the Federal Financial Supervisory Authority has so far rejected regulatory sandboxes, citing a lack of mandate from the administration [29], while the German Finance Committee rejected a proposal for regulatory sandboxes based on the UK proposal in mid-2020, citing consumer protection and conflicting European law [30, 31].

Regulatory sandboxes are now explicitly addressed at European level by the AI Act in Article 57 et seq. The AI Act aims to classify potential risk into different risk classes. It provides for prohibited AI systems (Article 5), high-risk systems (Article 6 et seq.), and low-risk systems to which only general obligations apply (Article 50). Risk is defined as the combination of the likelihood of harm and the severity of that harm, Article 3 (2). The group of high-risk systems will have the greatest practical relevance, as they are defined according to a dual regulatory approach of product safety law and the protection of fundamental rights. AI systems are considered high-risk under the AI Act if they constitute a product or a safety component of a product according to the harmonised provisions of Annex I, and are also subject to conformity assessment by third parties or pose a threat to fundamental rights in the case of the examples of use in Annex III. The danger of AI systems may be obvious in some cases, such as social scoring. Fundamentally, however, the problem remains that law alone cannot regulate the multi-dimensional nature of risks, especially in the case of AI. An “all or nothing” approach does not seem promising either, but it is important to be aware of the limits of legal regulation from the outset.

## 2 Examples

Most examples of regulatory sandboxes can be found in the fintech sector. This is due to the fact that this is a highly regulated market, which is at the same time highly innovative and has undergone significant changes on the supply side in recent years.

In Norway, the federal data protection authority established a “regulatory privacy sandbox” in 2020 to establish and stimulate privacy-enhancing innovation and digitalisation [32]. Every year 3–4 applicants were selected, resulting in 12 projects in total

having been tested in the sandbox by 2023. Participants came from the private and public sectors in the fields of health, transport, environment, and digital consumer services. In addition, generative AI applications were tested in a targeted manner [33]. The DPA documented the selection of participants in a transparent and comprehensible manner, and the sandbox has already been evaluated. Its hosting by the Norwegian DPA means it has retained the clearly defined objective of privacy enhancing technologies as central. The documentation, evaluation, and time limit of the sandbox, as well as the clear competences and focus on the DPA's objectives promote ethical and responsible application without limiting official supervision. The guiding principles rely on the ethics guidelines for trustworthy AI from the High-Level-Expert Group on AI set up by the European Commission [34]. After a pre-defined timeframe, the DPA publishes a detailed exit report. In a comment on the Commission's draft AI law, the Norwegian DPA echoed the general criticism of the lack of precision in the Commission's draft AI law in its concerns about the concept of regulatory sandboxes, stating: "However, we also see a need for some guidance on how competent authorities can strike a good balance between being a supervisory authority on the one hand and giving detailed guidance through a sandbox on the other. We propose that the AI Act specifies that participation in a sandbox does not constitute a stamp of approval, and that the organization/controller is still accountable for its processing of personal data." [35].

France has also set up an "Edtech sandbox" under the French data protection authority. This aims to help participants to develop and include data protection by design as required by the GDPR into their products. So far, the French approach does not provide for enforcement exceptions and is limited to legal and technical assistance to participants by the authorities.

Spain quickly followed suit in 2020, launching its regulatory sandbox on AI even before the AI Act came into force. The regulatory sandbox is housed in the newly established AI supervisory authority, the Spanish Agency for the Supervision of AI, located at the Ministry of Digital Transformation [36]. Thus far, Spain has established two sandboxes, one for the financial system, pursuant to the Law 7/2020 of 13 November on digital transformation of the financial system, and another for the electricity sector as a result of Royal Decree 568/2022 of 11 July establishing the regulatory framework for research and innovation in the electricity sector. In contrast to other sandboxes, the Spanish sandbox will be set up essentially to serve as a vehicle for studying the operability of the requirements of the AI Act. It is expected to result in reports on best practices and the compilation of technical guidelines for execution and supervision based on the evidence obtained, rather than collaborating with the authorities in defining and developing an adequate regulatory framework [37].

There are currently no regulatory sandboxes at federal level in Germany. In the political debate, the term "real-world laboratories" is used more often than "regulatory sandboxes". However, North Rhine-Westphalia (NRW) was the first federal state to launch the "Digi-Sandbox.NRW" project. So far however, these projects do not specifically focus on AI. There are also no plans for cooperation with a specific supervisory authority, but there is some support for the establishment of real-world laboratories.

At the federal level, a Real World Laboratory Act is planned, which provides for a one-stop-shop principle with a competent authority. The Federal Ministry of Economics and Technology (BMWi) has presented a concept for a Real-World Laboratory Act, which is intended to enshrine overarching standards for real-world laboratories and experimental clauses in law, and also to enable new real-world laboratories in important areas of digital innovation [38]. A public consultation phase was completed in 2023. There has so far been a lack of legal standards for real-world laboratories. Potential areas of application under discussion include AI applications in the field of modern mobility or Industry 4.0, innovative digital identification procedures, e.g. for digital driving licences, and digital legal services and procedures. The planned mandatory review of the experimental clause in the legislation is particularly interesting, as it is intended to allow the continuous identification of new areas of application. The main aim of the concept is to create innovation-friendly and thus economically favourable prospects for companies and strengthen Germany as a business location. However, another persistent problem is that the legislator is lagging behind digital transformation without actively shaping it.

### **3 Regulatory Sandboxes and the AI Act**

Regulatory sandboxes, experimental clauses, and experimental regulation in general are relatively unknown in EU law. This is partly due to the fact that experimental regulatory approaches are viewed with suspicion, as a legal vacuum incompatible with legal certainty or the unity of the legal order [39]. The AI Act now creates the first horizontal regulatory regime for sandboxes on the Union-level. The following is an outline of what can be expected for regulatory sandboxes based on the AI Act as released.

#### **3.1 Establishment of Regulatory Sandboxes**

This requirement for at least one AI regulatory sandbox as established in Article 57 AI Act can also be satisfied by establishing this sandbox jointly with the authorities of other member states. The obligation can also be fulfilled by participating in an existing sandbox as long as that participation provides an equivalent level of national coverage for the Member States, Article 57 (1). Article 52 (2) provides for additional regulatory sandboxes to be established at a local and regional level, thereby showing that the requirements are not intended to be exhaustive. The goal is to provide a controlled environment for the development, testing, and validation of innovative AI systems under the direct supervision and guidance of the competent authorities. Article 57 (9) explicitly states that the establishment of regulatory sandboxes shall follow the objectives of improving legal certainty and compliance with the AI Act and other applicable law, supporting best practices, fostering innovation, contributing to regulatory learning and facilitating access to the Union market for start-ups and SMEs.

The European Data Protection Officer may establish a regulatory sandbox at the Union level for Union institutions. Article 57 (4) also requires competent authorities be adequately resourced for regulatory sandbox tasks, as well as cooperation and collaboration between the authorities and with the AI Office where appropriate.



### 3.2 Procedural Aspect and Governance

In terms of procedure, participants and the supervisor must agree on a specific sandbox plan. Instructions to participants are optional, but the supervisor is required to provide guidance on supervisory expectations, Article 57 (5–7). The publication of final reports should have been mandatory rather than subject to the consent of participants, Article 57 (8). This would, for example, improve the situation for consumers who are better able to verify the “sandbox” label. The involvement of data protection authorities and other competent authorities is declaratory. Article 57 (11) states that any significant risks to health, safety and fundamental rights identified during the review of the sandbox will lead to immediate risk mitigation, including the temporary or permanent suspension of the testing process. National competent authorities must submit annual reports to the AI Office and to the Board, from one year after the establishment of the AI regulatory sandbox and every year thereafter until its termination and the issue of a final report. Annual reports or abstracts thereof will be made available to the public online.

### 3.3 Substantial Design and Exceptions

The substantial design elements of the regulatory sandboxes are laid down in Articles 57 and 58 of the AI Act. In temporal terms, the regulatory sandboxes as envisioned only apply to AI systems before entering the market or service, Article 57 (5). The AI Act does not stipulate legal exemptions or obligations for non-enforcement. However, Article 57(12) specifies that authorities will not impose fines for any infringements of the AI Act itself, provided prospective providers adhere to the outlined plan and conditions for participation, and faithfully follow the guidance of the competent national authorities. There are therefore no exceptions for other areas of law.

Unsurprisingly, given that these decisions are the responsibility of Member States, the AI Act makes it clear that participants in the sandbox will remain liable to third parties for any damage caused as a result of testing in the sandbox, Article 57 (12).

The specific design of real-life-testing conditions remains unclear as, according to Article 58 (1), the exact modalities of real-world laboratories will be defined in the implementing acts of the Commission. As a result, the AI Act does not regulate the important issues of selection criteria and eligibility, procedures and applications, or requirements and conditions for sandbox participation. This is unfortunate, as these are precisely the factors that will determine whether a regulatory sandbox will succeed. Article 58 (2) provides a list of goals that delegated acts must pursue, such as transparent and fair criteria, equal access, flexibility of the national authorities, the inclusion of other actors in the “AI ecosystem” such as standardisation bodies or start-ups, clear and simple communication of sandbox entry and exit conditions, time limits, facilitation of regulatory learning factors such as accuracy, robustness, security and risk mitigation measures for fundamental rights and society at large. These objectives, combined with the new instrument of sandboxes and the new requirements of the AI Act, are too vague and undefined to derive a specific design. Moreover, the important decision on these criteria should not have been left to the Commission alone.

The only reference to the actual design of the sandboxes is to be found outside the binding part of the Regulation: Recital 138 mentions that regulatory sandboxes

could be established in physical, digital, or hybrid form and may accommodate physical as well as digital products. The reference to physical products is consistent with the dualistic regulatory approach of Article 6 AI Act which classifies AI-Systems as high-risk-systems when they are subject under product safety law. Product safety law partly includes software systems, such as under the medical device regulation,<sup>5</sup> but strongly follows a physical understanding of products which characterizes the entire AI Act [40].

Article 59 of the Regulation provides legal topics for discussion by providing an exception to the basic principle of purpose limitation of data protection under Article 6 (1a, 4) GDPR if the conditions set out in paragraph 1 are met. For this purpose, the AI systems covered must be developed to safeguard a substantial public interest falling within the areas listed in Art. 59 (1) (a i–v). These include public safety and health, a high level of environmental protection, sustainable energy, safety and resilience of transport systems and mobility, and the efficiency and quality of public administration. This breach of the principle of purpose limitation for collective public interests is a novelty in data protection law, which continues to focus on protecting the fundamental rights of individuals [41]. In line with the objective of promoting innovation, SMEs should be given priority access to the AI regulatory sandboxes. However, the participation of other companies is not excluded under the conditions of Article 62(1). This approach is not convincing, as providing the benefit of regulatory sandboxes to big tech companies will further increase their market power. It would have been preferable to restrict access to regulatory sandboxes to small, SMEs, start-ups and public organisations, which are not even mentioned in the AI Act.

### 3.4 Testing Outside of Regulatory Sandboxes

The final version of the AI Act significantly weakened the concept of regulatory sandboxes for regulatory learning. In addition to the regulatory sandboxes, Article 60 now provides the opportunity to test high-risk AI systems under real-life conditions outside of regulatory sandboxes. This is problematic for several reasons. First, the AI Act is silent on what the actual real-life conditions should be. We do know natural persons may be involved, as Article 60 (4 g) states that participants must be adequately protected where required by their age or physical or mental disability. Additionally, Article 61 requires informed consent from participants and Article 60 (5) stipulates that test participants under real conditions may terminate their participation in the test at any time by withdrawing their informed consent without justification and request the immediate and permanent deletion of their personal data without incurring any disadvantage.

Second, the possibility of testing under real-life conditions outside regulatory sandboxes massively weakens the potential for regulatory learning. This is not explicitly standardised as an objective of the Article 60 procedure, but the authority retains its traditional authorisation power. As a result, no expert knowledge and, above all, no new knowledge about the products in question is generated. Article 76 merely stipulates that competent authorities must supervise tests carried out under real-life conditions in a

---

<sup>5</sup> Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC.

sandbox. Also of concern is the notional authorisation in Article 60(4)(b) which deems testing under real-life conditions to be authorised 30 days after application to the market surveillance authority. This puts considerable pressure on the competent authorities, but the rule should remain that authorisation is mandatory.

Third, Article 76 provides for the possibility of carrying out the tests under real-life conditions in a regulatory sandbox, which makes it much more difficult to distinguish between the two procedures. The conditions under which tests can be carried out under real conditions in the regulatory sandbox is unclear. In case of doubt, firms will opt for the simpler procedure outside a regulatory sandbox, with no benefit for supervisors.

## 4 Criticism

The proliferation of regulatory sandboxes is not without controversy. The recent explosion in the number of sandboxes suggests that some regulators are simply setting up a sandbox to take advantage of this trend. However, sandboxes are only effective tools if they are set up with the financial and human resources to run sustainably. A report commissioned by the United Nations Secretary General's Special Advocate for Inclusive Finance for Development (UNSGSA) found that "around a quarter of regulators have launched sandbox initiatives without first evaluating feasibility, demand, potential outcomes, or collateral effects." [42].

Furthermore, criticism has focused on the potential for inadequate consumer protection, dilution of regulation, overly generous exemptions, and unequal treatment. Germany's Federal Financial Supervisory Authority (BaFin) has long rejected regulatory sandboxes, arguing that promoting innovation is a market matter that does not require official involvement. Consumer groups and others have warned that regulatory sandboxes are bad for consumers [43]. The result could be a 'race to the bottom' for the least possible regulation to attract start-ups and other businesses [44]. There are also concerns that there will be no real exchange between authorities and companies, but that the latter will use the sandboxes as a regulatory discount and only for PR purposes. As sandboxes become more popular and more companies advertise with sandbox validation, it may become harder for consumers to distinguish between sandbox testing and real validation, resulting in companies and their new products being wrongly perceived as more trustworthy. The AI Act also provides for official confirmation of participation in the sandbox, which companies can then use for their own purposes, Article 57 (7).

This could be countered, for example, by supervisory authorities such as the FCA no longer publishing the list of sandbox participants, but instead offering confidential advice. This is prevented however by the transparency of such administrative action required to allow a large number of applications for sandboxes, rather than providing confidential advice to a few selected firms. Transparency is also the only way to ensure effective monitoring, for example, whether criteria for approval are consistent with the requirements of non-discrimination. In the area of fintech, some argue regulatory sandboxes can lead to "riskwashing". *Brown* and *Piroska* argue that sandboxes ease the introduction of fintech into society and finance to the extent that sandboxes themselves become a part of a fintech-financialization apparatus that intensifies penetration into typically non-financialised social relations with potentially socially disruptive effects [39].

The critical analysis of fintech and regulatory sandboxes as solutionism, particularly in relation to people who do not have access to banking services,<sup>39</sup> cannot be transferred seamlessly to AI and other technologies which cover a much wider range of applications. Nevertheless, the tendency towards solutionism is particularly clear in the field of AI, without first critically questioning what problems can be solved by technological innovation [45–47]. The one-sided promotion and creation of sandboxes can lead to the uncritical adoption of narratives and discourses from individual industries.

All these objections are understandable and realistic. Especially in the area of digital technologies, it is not the promotion of innovation that has been underrepresented so far, but rather the lack of effective regulation. Under no circumstances should regulatory sandboxes be used to undermine often new regulatory requirements. The success of a sandbox in promoting innovation and regulatory learning, without creating loopholes that undermine consumer interests and other regulatory objectives, depends on its specific design. Well-designed sandboxes can address these criticisms and still successfully achieve their objectives, not only focusing on specific test runs, but also contributing their findings to the regulatory debate in a targeted way. Additionally, the current uncertainty surrounding regulations contributes to only a minor fraction of the challenges encountered by newcomers on the digital market. It is therefore not judicious for regulatory bodies to allocate excessive resources to formulating exemption policies [27]. Instead, their efforts should be concentrated on dismantling barriers to the debut of new products that are not only compliant but also have the potential to enhance welfare.

Sandboxes should therefore, where they provide legal benefits such as exemptions from data protection rules, be located exclusively with the competent supervisory authorities, with a focus on regulatory learning and on targeting applications that promote the public good.

## 5 Conclusion and Outlook

The basic idea of regulatory sandboxes is a good approach to meeting the challenges of digitalisation with legal means. However, the AI Act should have been bolder in its specific design. Regulatory sandboxes should not provide legal rebates and should be located exclusively within the competent authorities. The resulting benefits for firms and the associated costs are only justified if supervisors also benefit. Nevertheless, the effectiveness of the AI Act heavily relies on its enforcement and Member State implementation of regulatory sandboxes. The requirement for all Member States to have or be involved in an AI specific regulatory sandbox thus creates a unique opportunity to develop and compare best practices. Ongoing evaluation should lead to the harmonisation of standards and the adoption of sound designs. Among the many other tasks involved in digital supervision, there is also a need for dialogue between competent authorities. Sandboxes should not be used as a fig leaf to blindly promote innovation, encourage lax regulation, or create a privacy discount. Not every innovation or development is desirable, but should be in the public interest. Sandboxes should therefore focus more on regulatory learning, particularly for the benefit of regulators, rather than smoothing the path for companies. There is still much potential for a key focus on the protection of fundamental rights and individuals, e.g. with regard to stakeholder participation.

Finally, well-functioning sandboxes can be expanded in their scope, allowing benefits beyond the administration. New technologies can also be tested in a targeted way, so that the lessons learnt from the sandboxes can feed not only into administrative implementation, but also into future legislative processes or evaluations. Hence, legislators can use the results sandboxes to inform future legislation on digital topics and technologies.

**Acknowledgments.** This contribution is part of the project “Trial and Error. Experimentelle Regulierung im Mehrebenensystem” funded by the Daimler und Benz Stiftung.

**Disclosure of Interests.** The author reports no conflicting interests.

## References

1. Newman, N.: The costs of lost privacy: consumer harm and rising economic inequality in the age of Google. *Wm. Mitchell L. Rev.* **40**, 849 (2013)
2. Barman, D., Guo, Z., Conlan, O.: The dark side of language models: exploring the potential of LLMs in multimedia disinformation generation and dissemination. *Mach. Learn. Appl.* **16**, 100545 (2024)
3. Frenkel, S.: Israel Deploys Expansive Facial Recognition Program in Gaza. *The New York Times* (2024)
4. Ranchordás, S., Rozna<sup>2</sup>I.Y.: Future-proofing legislation for the digital age. In: Ranchordás, S., Rozna<sup>2</sup>I.Y. (eds.) *Time, Law, and Change: An Interdisciplinary Study*, 347–366. Hart Publishing; Bloomsbury Publishing, Oxford, London (2020)
5. Beutel, F.K.: The lag between scientific discoveries and legal procedures. *Neb. L. Rev.* **33**, 1 (1953)
6. Ranchordás, S.: Experimental regulations for AI: sandboxes for morals and mores. *Morals Mach.* **1**, 86–100 (2021)
7. Fanta, A.: Climate measures: Behind closed doors, EU officials talk about banning Bitcoin. <https://netzpolitik.org/2022/climate-measures-behind-closed-doors-eu-officials-talk-about-banning-bitcoin/>
8. Renda, A., Pelkmans, J.: EU regulation: hindering or stimulating innovation? In: *Handbook of Innovation and Regulation*, pp. 263–293 Edward Elgar Publishing (2023)
9. Smuha, N.A.: From a ‘race to AI’ to a ‘race to AI regulation’: regulatory competition for artificial intelligence. *Law Innov. Technol.* **13**, 57–84 (2021)
10. Müller, R.: KI: Kann die EU künstliche Intelligenz nur regulieren? *FAZ.NET* (2024)
11. Ranchordás, S.: Experimental Regulations and Regulatory Sandboxes: Law without Order? *Law Method* **2021**, (2021)
12. Makarov, V.O., Davydova, M.L.: On the concept of regulatory sandboxes. In: Popkova, E.G., Sergi, B.S. (eds.) *Smart Technologies’ for Society, State and Economy*, pp. 1014–1020. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-59126-7\\_112](https://doi.org/10.1007/978-3-030-59126-7_112)
13. Ranchordás, S.: Innovation-friendly regulation: the sunset of regulation, the sunrise of innovation. *Jurimetrics* **55**, 201–224 (2015)
14. Ranchordás, S.: Experimental regulations and regulatory sandboxes: law without order? *SSRN Electron. J.* (2021). <https://doi.org/10.2139/ssrn.3934075>
15. Ranchordás, S., van ’t Schip, M.: Future-proofing legislation for the digital age. *SSRN Electron. J.* (2019). <https://doi.org/10.2139/ssrn.3466161>

16. Buocz, T., Pfothenhauer, S., Eisenberger, I.: Regulatory sandboxes in the AI act: reconciling innovation and safety? *Law Innov. Technol.* **15**, 357–389 (2023)
17. Zetzsche, D.A., Buckley, R.P., Arner, D.W., Barberis, J.N.: Regulating a revolution: from regulatory sandboxes to smart regulation. SSRN Scholarly Paper at <https://doi.org/10.2139/ssrn.3018534> (2017)
18. Krönke, C.: Sandkastenspiele – »Regulatory Sandboxes« aus der Perspektive des Allgemeinen Verwaltungsrechts. *JZ* **76**, 434–443 (2021)
19. Botta, J.: Die Förderung innovativer KI-Systeme in der EU. *ZfDR* 391–412 (2022)
20. Ahern, D.: Regulatory lag, regulatory friction and regulatory transition as FinTech Disenablers: calibrating an EU response to the regulatory sandbox phenomenon. *Eur. Bus. Org. Law Rev.* **22**, 395–432 (2021)
21. Bromberg, L., Godwin, A., Ramsay, I.: Fintech Sandboxes: achieving a balance between regulation and Innovation. SSRN Scholarly Paper at <https://papers.ssrn.com/abstract=3090844> (2017)
22. Ruschemeier, H.: AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum* **23**, 361–376 (2023)
23. Smuha, N.A., et al.: How the EU can achieve legally trustworthy AI: a response to the European commission’s proposal for an artificial intelligence act. *SSRN Electr. J.* 33899991 (2021). <https://doi.org/10.2139/ssrn.3899991>
24. Prevelakis, V., Spinellis, D.: Sandboxing applications. In: *USENIX Annual Technical Conference, FREENIX Track*, pp. 119–126, Citeseer (2001)
25. Donelan, E.: Regulatory governance: policy making, legislative drafting and law reform. Springer, Cham (2022). <https://doi.org/10.1007/978-3-030-96351-4>
26. Wagner, F.: Reallabore als kreative Arenen der Transformation zu einer Kultur der Nachhaltigkeit. In: *Die Experimentalstadt: Kreativität und die kulturelle Dimension der nachhaltigen Entwicklung*. Reiner mann, J.-L., Behr, F. (eds.), pp. 79–94. Springer Fachmedien, Wiesbaden (2017). [https://doi.org/10.1007/978-3-658-14981-9\\_5](https://doi.org/10.1007/978-3-658-14981-9_5)
27. Quan. A few thoughts on regulatory sandboxes. Stanford PACS. <https://pacscenter.stanford.edu/a-few-thoughts-on-regulatory-sandboxes/>
28. Regulatory Sandbox. FCA <https://www.fca.org.uk/firms/innovation/regulatory-sandbox> (2022)
29. Hufeld, F.: Neujahrspresseempfang der BaFin. (2016)
30. BT-Drucks. 19/19506. Regulatory Sandboxes – Für mehr Innovationen im Finanzmarkt
31. Eberle, N.: Die “Regulatory Sandbox”. pp. 175–179 (2020)
32. 20 applied for the sandbox. *Datatilysynet* <https://www.datatilysynet.no/en/news/aktuelle-nyheter-2023/20-applied-for-the-sandbox/>
33. Time for generative AI in the sandbox. *Datatilysynet* <https://www.datatilysynet.no/en/news/aktuelle-nyheter-2023/time-for-generative-ai-in-the-sandbox/>
34. European Commission. Ethics Guidelines for Trustworthy AI | Shaping Europe’s Digital Future. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (2019)
35. Norwegian Data Protection Authority. Norwegian Position Paper on the European Commission’s 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 (COM(2021) 206)
36. Agencia Española de Supervisión de la Inteligencia Artificial | España Digital 2026. <https://espanadigital.gob.es/lineas-de-actuacion/agencia-espanola-de-supervision-de-la-inteligencia-artificial>
37. Rivaya, G.-J.F., Vidal, A.: Spain: the artificial intelligence regulatory ‘sandbox’ has arrived. *Lexology* <https://www.lexology.com/library/detail.aspx?g=99939c25-d7bb-4d06-b154-4a972eb71e9b> (2023)

38. Bundesministerium für Wirtschaft und Energie & BMWi. Neue Räume, Um Innovationen Zu Erproben. Konzept Für Ein Reallabore-Gesetz. (2021)
39. Brown, E., Piroška, D.: Governing fintech and fintech as governance: the regulatory sandbox, Riskwashing, and Disruptive social classification. *New Polit. Econ.* **27**, 19–32 (2022)
40. Almada, M., Petit, N.: The EU AI act : a medley of product safety and fundamental rights? <https://cadmus.eui.eu/handle/1814/75982> (2023)
41. Mühlhoff, R., Ruschemeier, H.: Predictive analytics and the collective dimensions of data protection. *Law, Innov. Technol.* **0**, 1–32 (2024)
42. UNSGSA FinTech Working Group and CCAF. Early Lessons on Regulatory Innovations to Enable Inclusive FinTech: Innovation Offices, Regulatory Sandboxes, and RegTech. <https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/2019-summary-earlylessonsregulatoryinnovations.pdf> (2019)
43. team. Joint Letter: 80 groups oppose CFPB’s no-action letter and sandbox proposal. Americans for Financial Reform <https://ourfinancialsecurity.org/2019/02/joint-letter-80-groups-oppose-cfpbs-no-action-letter-sandbox-proposal/> (2019)
44. Kelly, J.: A “fintech sandbox” might sound like a harmless idea. It’s not. <https://www.ft.com/content/3d551ae2-9691-3dd8-901f-c22c22667e3b> (2018)
45. Morozov, E.: To save everything, click here: technology, Solutionism and the urge to fix problems that don’t exist. Allen Lane, London, UK (2013)
46. Paquet, G.: Governance as subversive bricolage in the 21st Century. Governance: Canada/Ireland. Canadian Embassy, Dublin Craig Dobbin Chair of Canadian Studies (University College Dublin) and the Association for Canadian Studies in Ireland (2003)
47. Sætra, H.S.: Technology and Sustainable Development: The Promise and Pitfalls of Techno-Solutionism. Taylor & Francis, Milton Park (2023)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# AI and Democratic Equality: How Surveillance Capitalism and Computational Propaganda Threaten Democracy

Ashton Black<sup>(✉)</sup> 

York University, Toronto, ON, Canada  
ashblack@my.yorku.ca

**Abstract.** In this paper, I argue that surveillance capitalism and computational propaganda can undermine democratic equality. First, I argue that two types of resources are relevant for democratic equality: 1) free time, which entails time that is free from systemic surveillance, and 2) epistemic resources. In order for everyone in a democratic system to be equally capable of full political participation, it's a minimum requirement that these two resources are distributed fairly. But AI that's used for surveillance capitalism can undermine the fair distribution of these resources, thereby threatening democracy. I further argue that computational propaganda undermines the democratic aim of collective self-determination by normalizing relations of domination and thereby disrupting the equal standing of persons. I conclude by considering some potential solutions.

**Keywords:** ethics of AI · AI and democracy · democratic equality · surveillance capitalism · computational propaganda · epistemic agency · epistemic resources · free time

## 1 Introduction

The threat that AI can pose to democracy is an increasing concern. Two of the most notable problems that arise with AI systems is the lack of transparency and accountability, two key features for a democratic system [7, 10, 12, 28, 48, 53]. In a democratic system, transparency involves openness around political processes, changes, etc., like documenting the steps involved in implementing a given policy and making that information accessible to the public [28]. Transparency around political information and procedures is critical for political agents to make informed decisions and meaningfully partake in a democratic system. Accountability is important for a democracy because government bodies and representatives need to be accountable to the majority's will and interests, especially so they can be held responsible for failure to do so [28]. Transparency and accountability are intertwined in a democratic system because the power of political actors needs to be confined by relevant standards (i.e., ethical, legal) as well as the majority's will, and in order for their power to be kept in check there needs to be full transparency around how that power is exercised within the demands of the system.



The back-and-forth engagement between the public and political actors with power is crucial for a healthy democracy, and transparency and accountability are vital for this dynamic to work.

The issue with AI is that even though it is increasingly shaping the social and political world, dominant tech corporations are not beholden to the public in the same way that formal political representatives and bodies are. This means that these corporations are not bound to principles of transparency and accountability that are important for a democracy. Notably, there is a lack of transparency around the information that dominant algorithmic systems collect and how that information is used [36, 37]. This is largely because AI systems and the data they collect is privatized, which means the corporations that create and use these technologies are not obligated to share this information with the public [55]. But as I just noted, transparency and accountability are closed intertwined. These tech corporations are not bound to the public by norms of accountability because they are private entities – for instance, the public is not guaranteed a role in determining the parameters of use of dominant algorithmic systems – and the lack of transparency from these corporations disconnects users from the information needed for active democratic participation, which undermines the democratic relation between those who have power to enact political/social change and those affected by that change [10, 29]. Thus, while some algorithmic systems can have profound effects on people's lives and can even violate anti-discrimination laws (e.g., by using race as a proxy for denying insurance requests [36]), these systems and the corporations that own them are privatized and thus go beyond the regulatory scope in a democracy that encompasses norms around transparency and accountability.

Accountability and transparency are critical for a democracy in part because of the relationship between knowledge and political agency [11]. In order for people to be able to politically represent their beliefs and interests, they have to at least be well-informed. One side of the problem with AI is that people are not given the relevant information needed to make informed decisions – like information around how an algorithmic system might affect them. Other than withholding information, the other side of the epistemic problem with AI involves deception through the proliferation of fake news and disinformation [1, 8, 17]. Disinformation and fake news like deepfakes, for instance, can make it hard for people to know what information to trust and, correspondingly, what political decisions to make [11, 16, 33, 39]. Beyond the worry about the lack of transparency and accountability with how these AI systems are used is the worry that these algorithmic systems can facilitate the destabilization of one's epistemic and political agency by making it difficult to discern what is true or false.

My aim here is to consider the relationship between AI and democracy by first analyzing what a theory of democracy entails. In Section two, I use Elizabeth Anderson's [2] theory of democratic equality, which states that everyone must have the equal capability for full political participation, including the capability to engage in open political discussion. For Anderson [2], being capable of doing something extends beyond formal opportunities like having the right to vote. People also need to have whatever resources are needed to be capable of full political participation. Although Anderson does not specify what kinds of resources are relevant for this capability, I augment Anderson's theory in Sections three and four by arguing that democratic equality requires a fair

distribution of two kinds of resources: 1) free time (which I'll argue entails time that is free from systemic surveillance) and 2) epistemic resources.

In Section five, I argue that AI used for surveillance capitalism disrupts democratic equality by creating a disparity of free time and epistemic resources. Surveillance capitalists can use AI to collect highly detailed behavioral data on their users through constant surveillance while simultaneously remaining unencumbered by public regulation due to free market protections. This inequality in capacities for systemic surveillance undermines the need for transparency and accountability that is critical for a democracy. This disparity in free time also facilitates a significant wealth gap in epistemic resources (e.g., behavioral data) between surveillance capitalists and the users of major online platforms, which can translate into a disparity of political agency. In Section six, I argue that AI used for computational propaganda destabilizes the foundation of democracy by producing a hostile political environment that undermines the process of collective self-determination. By artificially silencing or amplifying certain viewpoints or political representatives, AI bots that are used for computational propaganda constitute a domination tactic that violates the equal standing of persons and the democratic obligation to engage in productive and open discussion. Thus, AI used for computational propaganda undermines the democratic aim of collective self-determination. In the seventh and final section, I consider some possible solutions.

## 2 Democratic Equality

For Anderson [2], a democratic system operates through “collective self-determination by means of open discussion among equals, in accordance with rules acceptable to all” (p. 313). Accordingly, there are two key elements for democracy:

- 1) People have to stand in *equal* relation to each other to be able to collectively self-determine.
- 2) Democracy is not just a majority rule, it is constrained by standards like human rights (e.g., no one's basic human right to life can be compromised for the interests of others).

While the second feature of democracy is certainly relevant for thinking about the ways that AI can threaten democracy through human rights violations [15, 34], I will be focusing on the first dimension of democratic equality. In order for everyone to stand in equal relation to each other in a democracy, everyone needs to have an equal capability for full political participation [2]. Following Amartya Sen's capability approach, capabilities for Anderson [2] represent what a person is free and able to do given the resources and opportunities available to them. Having a capability in this sense, then, is more substantive than merely having formal liberties and opportunities. It is not enough for someone to just have the right to vote; they also need to be given whatever resources are needed to exercise that right. If a language barrier prevents someone from voting, for instance, language lessons or interpreters need to be provided. Anything that a person needs to be able to fully participate in political life should be available to them.

Crucially, Anderson [2] argues that being able to fully participate in political life requires freedom from domination, exploitation, or oppressive relations. Democracy

is fundamentally about collective self-determination, which requires open discussion amongst equals. This means that everyone is equally obligated to hear and respond to each other's contributions, and no one should be in a position to silence or dominate others. Accordingly, democratic equality refers to the equal standing of persons such that each person has the equal capability for full political participation, which requires being free from oppressive relations.

Moving forward, I use Anderson's theory in two ways. First, I'll augment her theory by arguing that the fair distribution of two kinds of resources is relevant to the capability for full political participation: 1) free time, which entails time that is free from systemic surveillance, and 2) epistemic resources. Specifying what kinds of resources are relevant for democratic equality enables me to argue that AI used for surveillance capitalism undermines the fair distribution of these resources. Second, I'll focus on Anderson's requirement that the equal standing of persons requires freedom from relations of domination, and I'll consider how computational propaganda undermines this requirement and the democratic aim of collective self-determination.

### 3 Free Time and Surveillance

To begin I'll start with the claim that democratic equality requires fair distributions of free time, which entails time that is free from surveillance. I build my argument from Julie Rose's [41] position that free time should be considered a resource in its own right – separate from the distribution of material goods – because time is a distinct resource that affects people's ability to exercise formal liberties and opportunities. Certainly, in order to vote, people need the time to vote. Regardless of material wealth, people cannot pay another person to vote on their behalf. Accordingly, material resources cannot compensate for the free time that's needed for political participation [41]. Free time is relevant for democratic equality, then, because it affects people's capability for full political participation. Thus, democratic equality requires a fair distribution of free time, meaning that each person should have whatever amount of free time is needed to be equally capable of full political participation. Given that everyone's needs vary (e.g., a physically disabled person may need additional resources and time to make it to a voting booth), the amount of free time each person should get will vary.

But having free time to exercise formal liberties and opportunities – like the having the time to vote – is not the only dimension of time that's relevant for democracy. Aside from the formal liberties and opportunities people can exercise in their free time, relations of domination and subordination in the private sphere can also affect the free time that people have for political participation. That is, even if people have free time to vote, they can still be subject to dynamics in their free time that undermine their ability for political participation. As AI and other smart technologies become increasingly ubiquitous in our daily lives, it's critical to think about the ways that exposure to digitally mediated systemic surveillance can facilitate relations of domination that undermine people's ability for full political participation.

Let me illuminate this point by way of example. In *Private Government*, Anderson discusses how most modern workplaces are taking on the structure of private authoritarian governments – one reason being because most modern workplaces have the legal

authority to surveil and regulate the private lives of workers, including the political views they share online [3]. Indeed, over half of U.S. workers lack legal protection from being fired for the political views they express outside of work (e.g., on Facebook) [3]. A poll recently shared by The Financial Post reveals that 86% of Canadian companies admitted they would fire an employee for an inappropriate social media post [46]. Certainly, the threat of unemployment in a precarious economy can significantly undermine the capability of some to participate in open political discussion.

I use this example to show that even if workers are given formal liberties and the free time to exercise them (e.g., time to vote), their ability for political participation (i.e., engaging in open political discussion) can be undermined when they're subject to systemic surveillance in their free time. Indeed, research shows that even in dictatorships, having increased capacities for surveillance through digital mediums has amplified the repression of citizens by expanding knowledge of organized uprisings as well as the identity of dissenting citizens [51]. Surveillance can undermine a person's capability for political participation by exacerbating power disparities that enable control over a person's social and political behaviour. Systemic surveillance thus disturbs free time as a resource for democratic equality because it inhibits the time in which people are in fact free to exercise their political agency. Hence, democratic equality requires not just free time but also some protections from systemic surveillance in one's free time. Therefore, a fair distribution of free time must at least include time that is free from systemic surveillance.<sup>1</sup>

## 4 Epistemic Resources, Epistemic Agency, and Political Agency

Democratic equality also requires a fair distribution of epistemic resources. Notably, the fair distribution of epistemic resources can also be undermined by systemic surveillance, but I'll come back to this point in the next section. For now, it's important to explain the connection between epistemic resources, epistemic agency, and political agency. For my purposes, political agency refers to the capability for full political participation, understood in Anderson's terms of what it means to have this capability. Epistemic agency – at least as defined by Mark Coeckelbergh – means having control over one's belief formation and revision [11]. For Coeckelbergh [11], epistemic agency is critical for political agency because being able to represent oneself politically (e.g., with voting) requires having control over the formation and revision of one's beliefs. If a person has no role in forming the belief that a specific political candidate is the best option – say, if that person was manipulated – then that person's ability for political self-representation is diminished. How one develops or revises a belief affects that person's ability for political participation. In order for people to have the equal capability for full political participation, then, people need to have the equal capability to control their belief formation and revision.

---

<sup>1</sup> I use the phrase 'systemic surveillance' to refer to surveillance that is done by those with the power to consistently surveil (e.g., social media giants, workplaces, governments, etc.) and thus regulate the behaviour of others. This kind of surveillance may be different from being surveilled and regulated by peers and social norms (e.g., cancel culture), so I distinguish systemic surveillance from surveillance more generally.

In order for people to have the equal capability to control their belief formation and revision, they need to have a fair share of epistemic resources. Epistemic resources contribute to propositional knowledge (i.e., knowing that things are the case) or knowing how to do something [45].<sup>2</sup> I said earlier that transparency is important in a democratic system because people need to be informed in order to hold representatives accountable or to meaningfully participate in the political sphere. Accordingly, a fair share of epistemic resources requires transparency and public accessibility regarding relevant political information and processes (e.g., recording and publishing the full details of a new federal policy) so that people can exercise their epistemic and political agency accordingly.

Epistemic resources also matter for epistemic and political agency because relations of domination can manifest through the control of epistemic resources. In authoritarian regimes, for instance, access to epistemic resources is often controlled (e.g., certain texts are banned) to deter opposition. By controlling what information is available to the public, authoritarian regimes control the belief formation and revision of citizens. In a democracy, the distribution of epistemic resources should not reflect or enable a relation of domination. Rather, the distribution of epistemic resources should be fair in that it enables each citizen to be equally capable of autonomous belief formation and revision and thus political participation.

## 5 Surveillance Capitalism

So far, I've argued that democratic equality requires a fair distribution of free time – which entails time that is free from surveillance – and epistemic resources. Here I argue that AI used for surveillance capitalism can undermine democratic equality by facilitating the unfair distribution of both of these resources. I borrow the concept of surveillance capitalism from Shoshana Zuboff [55] who defines it as a new economic order that uses human experience as free raw material, resulting in unprecedented disparities in power, particularly with knowledge production and dissemination. The free raw material of human experience that Zuboff [55] refers to is the behavioural data that is collected through machine learning and algorithmic systems. Since the ability to collect and use behavioural data with AI systems requires costly technological infrastructure – like physical storage space and powerful machine learning technologies – economic inequalities in the free market have translated into inequalities in technological power that create mass disparities in epistemic wealth and social power [55].

These disparities arise from what Zuboff [55] calls “the privatization of the *division of learning in society*” (p. 45). The division of learning in society is concerned with three things: 1) who has knowledge and controls who is in the circle that can access it 2) who has the authority to decide what's learned, who gets to learn, and what people can do with what they learn, and 3) the power that “undergirds the authority to share or withhold knowledge” (p. 352) [55]. With surveillance capitalism, the division of learning

<sup>2</sup> Note that while there is debate about whether there is knowledge other than propositional knowledge, this debate does not impact my argument. Even if know-how collapses into know-that, we still need a fair share of epistemic resources. That is, even if there is only one *kind* of epistemic resource, it still needs to be fairly distributed for democratic equality.

becomes privatized because the production and dissemination of epistemic resources is disproportionately controlled by private tech corporations like Google and Facebook. The dissemination of epistemic resources is dominantly controlled by these corporations because the information that users have access to through digital platforms (i.e., major search engines and social media sites) is privately regulated through each company [5, 27, 29, 30, 52]. Moreover, privately owned AI systems are used to generate a plethora of privatized epistemic resources in the form of behavioral data. Thus, information collection and dissemination via AI systems is controlled by major tech companies that get to determine who knows what, who can learn, who decides who knows and learns, and what people can do with what they know [55]. Thus, as Zuboff [55] notes, the division of learning in society has become privatized.

The privatization of the division of learning in society is partly facilitated by AI's unprecedented capacity for systemic surveillance [47]. While surveillance is not a new problem, AI has pushed the problem to new heights. By consistently and ubiquitously tracking people's behaviour on digital devices, AI systems can collect more information about people than their friends or family have – potentially more information than they even possess about themselves [10, 11]. For instance, some algorithmic systems can even track small, seemingly insignificant things like smartphone typing patterns as an indication of one's mental health [31]. Since digital technology has become a somewhat omnipresent feature in many of our lives and vast technological networks composed of various AI systems constitute superhuman capacities for persistent and meticulously detailed surveillance, the power to surveil has reached a new precipice. Accordingly, not only is there a dominant and pervasive exposure to the will of the algorithms – so to speak – in terms of information distribution, but there is also a dominant and pervasive exposure to the surveilling eye of the algorithms that collect information about users.

When thinking about democratic equality and the relevance of time that is free from systemic surveillance, it's important to note that there is a vast disparity of free time between surveillance capitalists and those that use their algorithmic systems. While users of algorithmic systems are extensively surveilled, algorithms are opaque (e.g., it's unclear what information these systems collect or how this information is used [7, 37]) and the market freedom afforded to surveillance capitalists removes them from public surveillance and regulation [55]. Indeed, this disparity of free time is what underwrites the privatization of the division of learning in society. Surveillance capitalism is thus marked by minimal public surveillance and regulation alongside an unprecedented power to systemically surveil users.

I previously noted that accountability and transparency are crucial for a democratic system. Full political participation requires transparency (i.e., around political processes) so one can make informed political decisions, and it also requires being able to hold representative bodies and actors accountable. Even though dominant tech corporations can affect significant political change in the broader social landscape, they are exempt from public surveillance/regulation and thus are not required to be transparent – like about whether their algorithms disproportionately deny jobs to disabled applicants [36] – and they are not beholden to norms of accountability that are important for a democracy. To put it simply, these corporate-political actors can affect major change in the political sphere without having to explain these effects or account for the public's political will

or interests. This creates a disparity of political power that is antithetical to democratic equality because those with substantial power to affect political change via extensive surveillance power are kept above the reach of public input and regulation. I previously noted that a distribution of free time – which entails time that is free from systemic surveillance – is unfair if it creates inequalities in the capability for full political participation. Surveillance capitalism creates an unfair distribution of free time by facilitating inequalities in political power between those who enact political change and those who are affected by it.

Furthermore, the disparity of surveillance capacities and free time facilitates a disparity in epistemic resources. As previously noted, the production and dissemination of information is largely controlled by surveillance capitalists through AI systems. The stats are revealing: over 80% of Americans use digital devices to get news [10]. The problem with receiving news and information through dominant algorithmic systems – like those used by Facebook and Google – is that these systems are programmed to maximally siphon what Zuboff [55] calls “human experience as free raw material” (p. 9). To collect as much behavioural data as possible, algorithmic systems are fundamentally aimed at increasing the amount of time people spend on a digital device. To do this, algorithms use the behavioural data they collect to individually personalize the online experience. That is, information distribution on these platforms is algorithmically determined by personal preference and what will grip a person’s attention, not epistemic standards regarding truth or reliable testimony [5, 27, 29, 30, 52]. This model of information distribution lends to the proliferation of fake news and disinformation, which damages people’s ability to obtain positive epistemic status (e.g., knowing, having justified beliefs, etc.) [1, 6, 11, 17, 20, 33].

For instance, an algorithmic system might direct a person down a rabbit hole of COVID conspiracy theories if it detects a susceptibility to or potential interest in a neurotic or paranoid way of thinking [29]. Since algorithmic systems are designed to tailor the online experience to each user, they are effective at influencing the behaviour [10, 20], epistemic habits [6], and psychological states of users, including their emotions and beliefs [32]. This gives those in control of these algorithmic systems disproportionate power to control the belief formation and revision of others, thereby undermining their epistemic agency [20, 41].

Given the interconnected nature of epistemic and political agency, not having control over one’s belief formation and revision can harm a person’s capability for full political participation. To give a concrete example, consider political micro-targeting: the personalized distribution of political information like campaign ads based on behavioural data [14, 19, 20, 24, 26]. Indeed, it was revealed that Trump’s 2016 presidential campaign algorithmically targeted voters based on behavioural data taken from Facebook [38]. With political micro-targeting, algorithms distribute information that is meant to manipulate their behaviour, not to inform them of the facts needed to make an informed political decision [20, 23, 29]. This means that people are not only vulnerable to being politically manipulated but are vulnerable to being manipulated through purely rhetorical or false information.



It's important to note here that regardless of whether micro-targeting influences an election, the problem is that the guiding epistemic principles behind digitally mediated information (or misinformation) distribution are fundamentally flawed. When the guiding principle controlling the flow of information is based on increasing consumer engagement and catering to subjective preferences, factors like truth and reliability are no longer the guiding principles behind the distribution of information [5]. To maintain epistemic agency and prevent the degradation of political agency, information flow should be dictated by epistemic criteria oriented towards obtaining positive epistemic status, not capitalistic criteria oriented towards profit.

The privatized capacity for algorithmic surveillance results in a disproportionate distribution of epistemic resources, and this disproportionate distribution of epistemic resources gives surveillance capitalists the ability to algorithmically manipulate people's psychological states, including their political beliefs. Thus, using AI for surveillance capitalism creates an unfair distribution of epistemic resources because the vast disparity of epistemic wealth gives surveillance capitalists disproportionate power over the belief formation and revision of others, which threatens their capability for full political participation. Thus, AI that is used for surveillance capitalism threatens democratic equality by creating an unfair distribution of free time and epistemic resources.

## 6 Computational Propaganda

So far, I've argued that AI used for surveillance capitalism can undermine democratic equality by creating unfair distributions of free time and epistemic resources. In this section, I set aside the focus on free time and epistemic resources and argue that AI that's used for computational propaganda can threaten democratic equality by creating relations of domination that undermine the process of collective self-determination. Indeed, democratic equality and collective self-determination requires the equal ability of citizens to express their political interests and views [17]. As noted in Section two, democracy requires that people are positioned in equal relation to each other – that there are no relations of oppression or domination that cause unequal capabilities for full political participation – and that people are obligated to hear each other out respectfully in the process of collective self-determination. But AI used for computational propaganda can directly suppress the ability for political discussion and undermine the obligation people have to hear each other out.

Samuel Woolley and Philip Howard [25, 50] define computational propaganda as politically motivated digital misinformation and manipulation via AI systems and human interference. Computational propaganda includes generating false political support or dissent through AI bots called amplifiers or dampeners [4, 13, 50]. An amplifier bot can share or like posts, flood comments sections, increase visibility of posts, etc., which can inadvertently minimize opposing voices by drowning them out [50]. Dampener bots, on the other hand, actively suppress information, channels, and viewpoints by shutting down websites, flooding people's tweets or posts with negative responses, or supporting negative responses to messages as a way to manufacture disapproval. For instance, dampener bots were found to have been supporting negative responses to the #GoodRiddanceHarper hashtag on twitter and were also used against Black Lives Matter



[50]. Indeed, dampener bots can be used to suppress certain views and movements, sometimes through harassment or bullying tactics.

Of course, propaganda and the way it can undermine democratic processes precedes AI. Nonetheless, it's necessary to recognize the specific ways that AI amplifies this problem and creates new iterations of it. Similar to any propaganda, computational propaganda can undermine people's epistemic agency. The belief formation and revision of some can be controlled by those who artificially manufacture mass consensus or dissent. However, computational propaganda is particularly alarming because it normalizes a political environment of domination and intolerance. Since many people spend a lot of time online, the internet has become a prominent platform for political discussion. But AI used for computational propaganda makes the platform where much political discussion takes place unequal [41]. Amplifiers and dampeners constitute a digitally mediated domination tactic that directly creates inequalities in the capability for open discussion.

The suppression of certain views and voices through computational propaganda also normalizes intolerance and eschews the obligation for collective engagement and action, which can encourage the formation of echo chambers. Suppressing certain views and voices cuts off or limits the ability to foster trusting relationships with those who have different perspectives. An echo chamber can form when people's circle of trust can become narrowly confined to include only those with specific views – even if those views are false or harmful – which makes it hard for them to believe information that is true [6, 36, 44].

The possibility for echo chambers is even further compounded by the fact that algorithmically mediated information distribution is based on personal preference, not epistemic standards, which can narrow a person's epistemic circle to include only those who have the same false and politically harmful beliefs and perspectives [1, 6, 10, 17]. For instance, some experts attribute the attack on Capitol Hill to social media sites like Parler that allow the proliferation of disinformation and encourage the formation of echo chambers [18, 40]. Researchers in political science and communications also found that there was a correlation between getting news off social media and politically polarizing behaviour, like unfriending people or saying harmful things to those with opposing viewpoints [10, 21]. When the environment where people consume belief-altering content and have political discussions encourages intolerance and hostility, political and social relations get further from the ideal of collective self-determination. Yet, collective self-determination is the defining feature of democracy, and it requires that people hear each other out and engage in productive discussion, not dominate and silence each other. In facilitating a hostile, intolerant, and politically polarizing environment, the use of AI for computational propaganda destabilizes the very foundation of democracy.

## 7 Conclusion

I've argued that AI can pose a threat to democratic equality, first by undermining the fair distribution of resources that are important for democratic equality, and second by undermining the process of collective self-determination. When used for surveillance capitalism, AI can facilitate the unfair distribution of free time and epistemic resources, resulting in disparities of political power and agency. When used for computational

propaganda, AI can facilitate relations of domination that disrupt the equal standing of persons. However, AI is not inherently anti-democratic. Rather, it can be used to enhance our social and political world [9, 10, 28, 29, 43, 46]. Some have even argued that AI can serve democratic aims by increasing political participation [43]. If used appropriately, AI has the potential for positive social and political effects. Identifying the problems AI currently poses for democracy, then, should be seen as a starting point for pinpointing practical and relevant solutions.

One possible way to mitigate the effects AI can have on democratic equality is by increasing access of epistemic resources. One way to do this is to increase transparency around algorithmic systems (e.g., increase information about what information these systems collect and how this info is used) [29, 37]. Another possible solution that researchers have explored involves institutionalizing tech literacy training [10, 29]. Since part of the unfair distribution of epistemic resources stems from the unequal ability to understand and apply these technologies, publicly funded and distributed tech literacy training is one promising way to increase the fair distribution of epistemic resources. Though, it's not enough just to be better informed about how these algorithmic systems work or what they know. It's also vital to maintain human epistemic autonomy and authority over AI systems, especially those used in decision-making procedures [54]. Current AI should be viewed as a tool that can aid the decision-making procedure rather than an agent capable of making properly informed and well-considered decisions.

Another way to make current and emerging AI more compatible with the principles and foundation of democracy is to democratize the design and implementation of AI systems [34, 49]. Indeed, the public should have a say in the algorithmic systems that are increasingly shaping the social and political world. This would require government policies that de-privatize what Zuboff calls the division of learning in society. De-privatization of the division of learning in society would disrupt the unfair distribution of free time by alleviating the radical disparity of surveillance power (e.g., privatized data collection).

It's also worth considering that social systems and institutions themselves may need to change and adapt appropriately to mitigate political problems that arise with big data and AI. As Benn and Lazar [5] suggest, moving towards a systematic procedure of collective action may require institutionalizing epistemic authorities that can control the dissemination and flow of digitally mediated information. In other words, it might be critical to direct power away from surveillance capitalists who determine information flow based on personal preference and instead direct it towards institutionally regulated and instated epistemic authorities who can control the flow of information based on relevant epistemic markers (e.g., what's true or false rather than what will grab someone's attention). Though, I make these suggestions only as a starting point for further investigation, as exploring solutions with due consideration is beyond the scope of this paper. My primary aim here is to explore the ways that AI can threaten democratic equality so that there can be further research on targeted solutions.

**Acknowledgments.** Thanks to my mentor Regina Rini for giving me feedback on this paper and for supporting my research aims and academic growth. I am also grateful to the organizing committee of the AISoLA conference October 2023 for giving me the opportunity to present this paper, and I appreciate those who provided helpful feedback during the Q&A.

**Disclosure of Interests.** The author has no competing interests.

## References

1. Aïmeur, E., Amri, S., Brassard, G.: Fake news, disinformation and misinformation in social media: a review. *Soc. Netw. Anal. Min.* **13**(1), 30 (2023). <https://doi.org/10.1007/s13278-023-01028-5>
2. Anderson, E.: What is the point of equality? *Ethics* **109**(2), 287–337 (1999). <https://doi.org/10.1086/233897>
3. Anderson, E.: *Private Government: How Employers Rule our Lives (and why we don't talk about it)*. Princeton University Press, Princeton (2017)
4. Bastos, M.T., Mercea, D.: The Brexit Botnet and user-generated hyperpartisan news. *Soc. Sci. Comput. Rev.* **37**(1), 38–54 (2019). <https://doi.org/10.1177/0894439317734157>
5. Benn, C., Lazar, S.: What's wrong with automated influence. *Can. J. Philos.* **52**(1), 125–148 (2022)
6. Blake-Turner, C.: Fake news, relevant alternatives, and the degradation of our epistemic environment. *Inquiry* 1–21 (2020). <https://doi.org/10.1080/0020174X.2020.1725623>
7. Buhmann, A., Fieseler, C.: Deep learning meets deep democracy: deliberative governance and responsible innovation in artificial intelligence. *Bus. Ethics Quart.* **33**(1), 146–179 (2023). <https://doi.org/10.1017/beq.2021.42>
8. Carlson, M.: The robotic reporter: automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digit. J.* **3**(3), 416–431 (2015). <https://doi.org/10.1080/21670811.2014.976412>
9. Cath, C.: Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos. Trans. Royal Soc. A: Math. Phys. Eng. Sci.* **376**(2133) (2018). <https://doi.org/10.1098/rsta.2018.0080>
10. Christodoulou, E., Iordanou, K.: Democracy under attack: challenges of addressing ethical issues of AI and big data for more democratic digital media and societies. *Front. Polit. Sci.* **3** (2021) <https://doi.org/10.3389/fpos.2021.682945>
11. Coeckelbergh, M.: Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI Ethics* (2022). <https://doi.org/10.1007/s43681-022-00239-4>
12. Coglianese, C., Lehr, D.: Transparency and algorithmic governance. *Admin. Law Rev.* **71** (2019). <https://ssrn.com/abstract=3293008>
13. Confessore, N.: Cambridge Analytica and Facebook: the Scandal and the Fallout so Far. *The New York Times*, New York (2018). <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>
14. Endres, K., Kelly, K.J.: Does microtargeting matter? Campaign contact strategies and young voters. *J. Elect. Public Opin. Parties* **28**(1), 1–18 (2018). <https://doi.org/10.1080/17457289.2017.1378222>
15. Eubanks, V.: *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, Inc. (2018)
16. Fallis, D.: The epistemic threat of deepfakes. *Philos. Technol.* **34**(4), 623–643 (2020)
17. Farkas, J., Schou, J.: *Post-truth, Fake News and Democracy: Mapping the Politics of Falsehood*. Routledge, Milton Park (2019)
18. Frenkel, S.: The Storming of Capitol Hill was Organized on Social Media. *The New York Times*, New York City (2021). <https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html>
19. Franz, M.M.: Targeting Campaign Messages: Good for Campaigns but Bad for America? In T. N. Ridout (Ed.), *New directions in media and politics*. 2nd edn. Routledge (2018)

20. Gorton, W.A.: Manipulating citizens: how political campaigns' use of behavioral social science harms democracy. *New Polit. Sci.* **38**(1), 61–80 (2016). <https://doi.org/10.1080/07393148.2015.1125119>
21. Goyanes, M., Borah, P., Gil De Zúñiga, H.: Social media filtering and democracy: Effects of social media news use and uncivil political discussions on social media unfriending. *Comput. Human Behav.* **120**, 106759 (2021). <https://doi.org/10.1016/j.chb.2021.106759>
22. Helbing, D., et al.: Will democracy survive big data and artificial intelligence? In: Helbing, D. (ed.), *Towards Digital Enlightenment*, pp. 73–98. Springer, Cham (2019). [https://doi.org/10.1007/978-3-319-90869-4\\_7](https://doi.org/10.1007/978-3-319-90869-4_7)
23. Hersh, E.D.: *Hacking the Electorate: How Campaigns Perceive Voters*. 1st edn. Cambridge University Press, Cambridge (2015). <https://doi.org/10.1017/CBO9781316212783>
24. Howard, P.N., Woolley, S., Calo, R.: Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *J. Inform. Tech. Polit.* **15**(2), 81–93 (2018). <https://doi.org/10.1080/19331681.2018.1448735>
25. Jamieson, K.H.: Messages, Micro-targeting, and new media technologies. *The Forum* **11**(3) (2013). <https://doi.org/10.1515/for-2013-0052>
26. Just, N., Latzer, M.: Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media Cult. Soc.* **39**(2), 238–258 (2017). <https://doi.org/10.1177/0163443716643157>
27. Kim, J., Tabibian, B., Oh, A., Schölkopf, B., Gomez-Rodriguez, M.: Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 324–332 Association for Computing Machinery, New York (2018). <https://doi.org/10.1145/3159652.3159734>
28. König, P. D., Wenzelburger, G.: Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Govern. Inf. Quart.* **37**(3), (2020). <https://doi.org/10.1016/j.giq.2020.101489>
29. Lanzing, M.: “Strongly recommended” revisiting decisional privacy to judge hypernudging in self-tracking technologies. *Philos. Technol.* **32**(3), 549–568 (2019). <https://doi.org/10.1007/s13347-018-0316-4>
30. Mastoras, R.-E., et al.: Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Sci. Rep.* **9**(1) (2019). <https://doi.org/10.1038/s41598-019-50002-9>
31. Meyer, R.: Everything we know about Facebook's secret mood-manipulation experiment. *The Atlantic* (2021) <https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/>
32. McKay, S., Tenove, C.: Disinformation as a threat to deliberative democracy. *Polit. Res. Q.* **74**(3), 703–717 (2021). <https://doi.org/10.1177/1065912920938143>
33. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. *philosophical transactions of the royal society a: mathematical. Phys. Eng. Sci.* **376**(2133) (2018). <https://doi.org/10.1098/rsta.2018.0089>
34. Noble, S.U.: *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York (2018)
35. Nguyen, T.C.: Echo chambers and epistemic bubbles. *Episteme* **17**(2), 141–161 (2020). <https://doi.org/10.1017/epi.2018.32>
36. O'Neil, C.: *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. 1st edn. Crown (2016)
37. Persily, N.: Can democracy survive the internet? *J. Democr.* **28**(2), 63–76 (2017). <https://doi.org/10.1353/jod.2017.0025>
38. Rini, R.: Deepfakes and the epistemic backstop. *Philos. Imprint* **20**(24), 1–16 (2020)

39. Romero, L.: Experts say echo chambers from apps like Parler and Gab contributed to attack on Capitol. ABC News (2021). <https://abcnews.go.com/US/experts-echo-chambers-apps-parler-gab-contributed-attack/story?id=75141014>
40. Rose, J.: Free Time. Princeton University Press, Princeton (2016)
41. Bradshaw, S., Philip N.H.: Social media and democracy in crisis. In: Bradshaw, S.S., Howard, P. (ed.) *Society and the Internet*, pp. 212–227. Oxford University Press (2019) <https://doi.org/10.1093/oso/9780198843498.003.0013>
42. Savaget, P., Chiarini, T., Evans, S.: Empowering political participation through artificial intelligence. *Sci. Public Policy* **46**(3), 369–380 (2019). <https://doi.org/10.1093/scipol/scy064>
43. Scheufele, D.A., Krause, N.M.: Science audiences, misinformation, and fake news. *Proc. Natl. Acad. Sci.* **116**(16), 7662–7669 (2019). <https://doi.org/10.1073/pnas.1805871115>
44. Shotwell, A.: Forms of knowing and epistemic resources. In: Kidd, I.J., Medina, J., Pohlhaus, G. (eds.) *The Routledge Handbook of Epistemic Injustice*. Routledge Handbooks Online (2017). <https://doi.org/10.4324/9781315212043.ch7>
45. Stahl, B.C., et al.: Artificial intelligence for human flourishing – beyond principles for machine learning. *J. Bus. Res.* **124**, 374–388 (2021). <https://doi.org/10.1016/j.jbusres.2020.11.030>
46. Stoller, B.: From Instagram to insta-fired: 86% of Canadian companies would fire employees for inappropriate social media posts. *The Financial Post*. (2023). <https://financialpost.com/globe-newswire/from-instagram-to-insta-fired-86-of-canadian-companies-would-fire-employees-for-inappropriate-social-media-posts>
47. Tene, O. Polonetsky, J.: Big data for all: privacy and user control in the age of analytics. *Northwestern J. Technol. Intell. Prop.* **11**(5) (2013). <https://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1>
48. Wachter, S., Mittelstadt, B., Floridi, L.: Transparent, explainable, and accountable AI for robotics. *Sci. Robot.* **2**(6) (2017). <https://doi.org/10.1126/scirobotics.aan6080>
49. Westerstrand, S. Ethics in the intersection of AI and democracy: The AIDEM Framework. *ECIS 2023 Research Papers* **321** (2023). [https://aisel.aisnet.org/ecis2023\\_rp/321](https://aisel.aisnet.org/ecis2023_rp/321)
50. Woolley, S., Howard, P.N.: *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, Oxford (2019)
51. Xu, X.: To repress or to co-opt? Authoritarian control in the age of digital surveillance. *Am. J. Polit. Sci.* **65**(2), 309–325 (2021). <https://doi.org/10.1111/ajps.12514>
52. Yeung, K.: ‘Hypernudge’: big data as a mode of regulation by design. *Inf. Commun. Soc.* **20**(1), 118–136 (2017). <https://doi.org/10.1080/1369118X.2016.1186713>
53. Zarsky, T.: The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. *Sci. Technol. Human Values* **41**(1), 118–132 (2016). <https://doi.org/10.1177/0162243915605575>
54. Zimmermann, A., Lee-Stronach, C.: Proceed with caution. *Can. J. Philos.* **1**, 6–25 (2021)
55. Zuboff, S.: *The age of surveillance capitalism: The fight for a human future at the New Frontier of Power* (Kindle edn). Public Affairs (2020)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# **AI Assisted Programming**



# Safeguarding AI-Based Software Development and Verification using Witnesses (Position Paper)

Dirk Beyer 

LMU Munich, Munich, Germany

**Abstract.** This position paper accompanies a presentation and discussion at AISoLA 2023 on the topic of how (generative) AI influences software engineering, namely, the development and verification of software programs. We outline a few opportunities and challenges that are posed by the application of AI. AI-based techniques can be an efficient method to produce software code. Not only this, AI can also be efficient in producing invariants that help proving correctness of software programs. But unfortunately, the results generated by AI-based approaches are often still imprecise or wrong: Code produced with the help of AI often does not satisfy the specification, and AI-generated invariants are often not helpful to prove the correctness of the program. The solution is to safeguard the process by independently checking the results using verification witnesses and witness validation. The goal of this paper is to convince readers that software witnesses are important and that independent result validation is of utmost importance to ensure correctness.

**Keywords:** Software Verification · Software Development · Generative AI · Machine Learning · Formal Methods · Software Correctness · Invariants · Verification Witnesses · Error Reports · Exchange Formats · Explainability

## 1 Motivation

Generative AI has applications in almost all areas of software engineering, from requirements engineering over code generation to quality assurance. This discussion at AISoLA focused on three applications of AI (1–3 below) that are all important since they affect the correctness of the software, have a large potential for improvement through AI, and, on the negative side, currently still suffer from imprecision of the AI techniques. The three approaches share another commonality: Their negative impact on the correctness can be addressed by the same solution: witness-based validation of result.

(1) Generative AI can be used to suggest software code. It seems that the more integrated this technique is in the software-development process, the more productive it becomes. Recent empirical studies show that approaches like GitHub Copilot can significantly improve the productivity, especially if developers are trained on how to use them. There is a large body of literature on this topic



describing the state of the art [1, 2]. The problem is that generated code often contains bugs that are sometimes obvious but sometimes quite subtle and therefore missed by developers.

(2) Whenever software has to solve an intractable problem, that is, a problem for which no efficient algorithm is known, heuristics are applied that work well in certain circumstances [3]. Often, there are many different heuristics from which the developer or user can choose. For example, in software verification, the tools PESCO [4, 5] and GRAVES-CPA [6, 7] use an ML-based algorithm selection to choose the most promising configuration from the software-verification framework CPACHECKER [8, 9]. The problem here is that the best configuration selected by the AI might turn out to be the wrong choice and produce a wrong result and another configuration producing the correct result remains unused. For example, the AI may select a configuration without pointer analysis (because it might have been fast on other programs without pointers during training) for a program that has pointers.

(3) AI can be used to generate invariants that accelerate the construction of a proof of correctness. Machine learning has been used to infer loop invariants for programs [10]. Also, termination arguments [11] have been derived by using neural networks to represent ranking functions. Generative AI can be used as code pilots for interactive theorem proving [12]. A recent study showed that generative AI (CHATGPT) can be used to generate loop invariants that help FRAMA-C prove the correctness of a program [13]. The problem is that the suggested predicates might not be valid invariants, might not be inductive, or might not aid in proving the safety property.

Imprecise AI-based techniques should be safeguarded by techniques that verify the result, like witness-based result validation.

## 2 Solution

We would like to advocate verification witnesses as a means towards solving the above-mentioned problems, that is, by safeguarding software development and verification with the help of verification witnesses.

### 2.1 Related Work

Traditionally, verifiers returned claims of the form *true* (system satisfies the specification) or *false* (system violates the specification) and the user is left alone with the result. Any kind of useful information that *explains* the result would improve the situation. The formal-methods community has established the practice of witness validation, in the area of software verification [14, 15], termination checking [16], SAT solving [17, 18], SMT solving [19, 20], and hardware model checking [21, 22]. Witnesses were also used for graph algorithms [23]. The technique of execution reports [24] was investigated to provide analysis results

in a more structured way, and the format SARIF [25] is supported by some program-analysis tools. Information exchange also enables the integration of tools, for example, the Evidential Tool Bus (ETB) [26, 27, 28] supported tool integration by producing claims that are supported by evidence, and the Electronic Tools Integration (ETI) [29, 30, 31] offered a platform and web service for integrating model-checking tools. Also cooperative verification [32, 33, 34] requires information in standard exchange formats and tools should be collected using a standard format with information about the tools [35].

## 2.2 Verification Witnesses

Nowadays, software verifiers produce witnesses to certify their result [14, 15, 36]. The competition on software verification (SV-COMP) uses verification witnesses since 2015 [37]. The most recent advancement is that the competition on software verification [38] has for its 2023 edition introduced a new track for the evaluation of tools for witness validation [39]. Furthermore, to make verification witnesses more human-readable and more concise, and semantically well-defined, the community developed a new version 2.0 of the widely supported exchange format for verification witnesses [40]. This new standard format was immediately adopted in the competition in its 2024 edition [41].

*Software verification* is the process of producing, for a given program  $P$  and specification  $\varphi$ , a verdict (from *true*, *false*, and *unknown*) and a verification witness  $w$  (a correctness witness for verdict *true* and a violation witness for verdict *false*). The verdict *true* means that  $\varphi$  holds for  $P$  and the verification tool has constructed a proof of correctness  $\pi$  (denoted as  $\pi : P \models \varphi$ ). In this case, the *correctness* witness contains program invariants aiding the construction of the correctness proof. The verdict *false* means that  $\varphi$  is violated by  $P$  and the verification tool has constructed a counterexample to the proof of correctness  $\pi$  (denoted as  $\pi : P \not\models \varphi$ ). In this case, the *violation* witness describes at least one error path through  $P$  that violates  $\varphi$ , and for the proof of violation it suffices to analyze the semantics of the program along the described error paths.

*Software validation* in this context is the process of reestablishing a verdict, for a given program  $p$ , specification  $\varphi$ , and verification witness  $w$ . To reconstruct a proof of correctness  $\pi' : P \models \varphi$ , the validator takes the invariants stored in  $w$  and checks if they hold. If this is the case, then the validator can use the invariants as lemmata in its own proof of correctness  $\pi'$ . If the invariants in the witness hold and the program satisfies the specification, then the witness is *valid for verdict true*. To reestablish a verdict *false*, the validator explores the paths described by the witness and checks if they are feasible. If a feasible path is found, then the validator can use the path to check whether the strongest post-condition of the operations along the path leads to a specification violation. If a feasible path described by the witness violates the specification, then the witness is *valid for verdict false*. We refer to the literature [15, 40] for more details.

**Program:**

```

1 #include <assert.h>
2 extern unsigned char
↪ __nondet_uchar(void);
3
4 int main() {
5     unsigned char n =
↪     __nondet_uchar();
6     if (n == 0) {
7         return 0;
8     }
9     unsigned char v = 0;
10    unsigned int s = 0;
11    unsigned int i = 0;
12    while (i < n) {
13        v = __nondet_uchar();
14        s += v;
15        ++i;
16    }
17    assert(s >= v);
18    assert(s <= 65025);
19    return 0;
20 }
```

**Specification:**

All assertions in the program must hold.

**Witness:**

```

1 - entry_type: invariant_set
2 metadata:
3     format_version: "2.0"
4     producer:
5         name: "CPAchecker"
6 content:
7     - invariant:
8         type: loop_invariant
9         location:
10            file_name: "inv-a.c"
11            line: 12
12            column: 3
13            function: main
14            value: "s <= i*255 &&
↪ 0 <= i && i <= 255 &&
↪ n <= 255"
15            format: c_expression
```

Fig. 1: Example C program similar to `inv-a.c` (left, adopted from [15]), satisfying the given specification (top right), and correctness witnesses in format 2.0 (right, metadata shortened, with a single nontrivial invariant, adopted from [40])

## 2.3 Example

For illustration, we show an example from the literature [40] in Fig. 1 for a verification witness that contains a loop invariant: The figure shows a C program (left) and a specification (top right), together with a correctness witness in format 2.0 (right). The program mainly consists of a loop in which `n` values are read into variable `v` and summed up in variable `s`. The specification is to ensure that no assertion is violated. The first assertion requires that the sum `s` is at least as large as the last read value `v`. The second assertion requires that the sum `s` is less than or equal to 65025.

The creative task for program verification is now to come up with a loop invariant. For example, let us consider  $s \leq i * 255 \wedge 0 \leq i \wedge i \leq 255 \wedge n \leq 255$ . This predicate captures the knowledge that the upper bound of variable `s` is `i` times the largest possible value for `v`. Since variable `v` has type `unsigned char`, its largest possible value is 255. It further tells us that variable `i` has only values from 0 to 255, because it starts with 0, is counted up by 1, and its largest value is the largest value of `n`, which is 255. The invariant also tells us that the value of `n` is bound by 255, because it is of type `unsigned char`.

Table 1: Witness Validators in SV-COMP 2024, with literature references, the language they support, since when the tools exist, and which formats they support (‘viol’ short for violation witnesses, ‘corr’ short for correctness witnesses)

Validator	Reference	Lang.	Since	Supported Formats			
				1.0-viol	1.0-corr	2.0-viol	2.0-corr
CPACHECKER	[14, 36, 42]	C	2015	✓	✓	✓	✓
UAUTOMIZER	[14, 36]	C	2016	✓	✓		✓
CPA-WITNESS2TEST	[43]	C	2018	✓			
CPROVER-WITNESS2TEST	[43]	C	2018	✓			
METAVAL	[44]	C	2020	✓	✓		
NITWIT	[45]	C	2020	✓			
WITNESSLINT	[40]	C	2021	✓	✓	✓	✓
DARTAGNAN	[46]	C	2022	✓			
GWIT	[47]	Java	2022	✓			
SYMBIOTIC-WITCH	[48]	C	2022	✓			
WIT4JAVA	[49]	Java	2022	✓			
CONCURWITNESS2TEST	[50]	C	2024	✓			
GOBLINT	[51]	C	2024				✓
JCWIT		Java	2024		✓		
LIV	[52]	C	2024		✓		
MOPSA	[53]	C	2024				✓
WITCH	[40, 54]	C	2024			✓	

This predicate has three interesting properties: (a) It is a *loop invariant*, because it holds at every iteration before the evaluation of the loop head (beginning of the loop). (b) It is *inductive*, because if assumed at the loop head, it holds again at the next visit of the loop head. (c) It is *safe*, because it implies that the safety specification holds. The latter is true because both assertions hold. The first assertion holds because after  $v$  (with a value in the interval  $[0, 255]$ ) is added to  $s$ , the value of  $s$  is at least as large as the value of  $v$ . The variable  $s$  cannot overflow because it is of type `unsigned int`, which is sufficiently large and  $s$  does not grow larger than 65025 (see next assertion). The second assertion holds because  $s$  is always less than or equal to the product of the largest possible values for variables  $n$  and  $v$ , which are both bound by 255.

No matter how a verification tool came up with such an invariant, witness validation is available to safeguard the verification result: A witness-based result validator takes as input the program, the specification, and the witness, and checks whether the claimed invariant really holds and the program fulfills the specification.

Figure 1 shows a correctness witness on the right. Besides the invariant, the witness format captures all necessary and useful information, such as the precise location at which the invariant holds, the format in which the invariant is specified, that the invariant is a loop invariant, and metadata about the producer of the witness and the verification task.

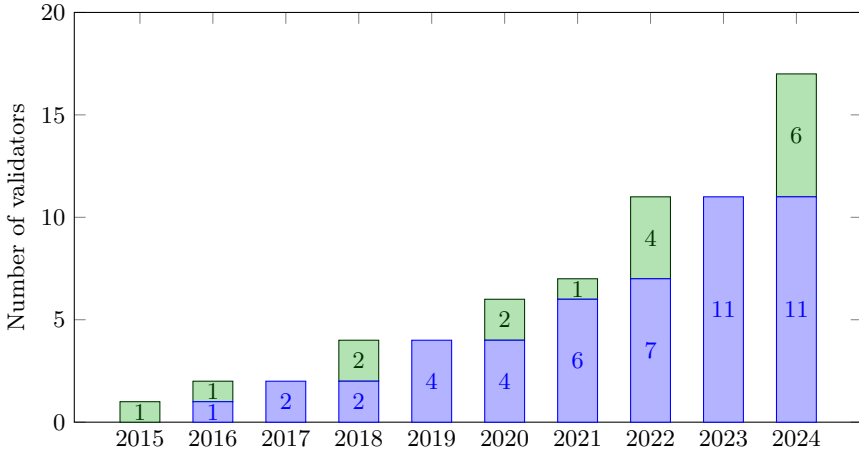


Fig. 2: Number of witness validators evaluated in SV-COMP for each year (first-time participants on top), taken from 2024 report [41, page 317]

## 2.4 Available Witness Validators

Validation of verification results becomes more and more important, as the research on witnesses and their validation matures [39]. The most recent competition on software verification has compared 17 validators [41] (including a syntax checker *WITNESSLINT*). There are currently two kinds of verification witnesses considered: correctness witnesses and violation witnesses. Also, there are two versions of the format for verification witnesses: The first format, version 1.0, is from 2015 [14] and based on GraphML (XML) [55]. It describes a witness automaton. The second format, version 2.0, is from 2023 [40] and based on YAML format. In sum, there are four combinations of witness kind and witness format, version 1.0 for violation (‘1.0-viol’), version 1.0 for correctness (‘1.0-corr’), version 2.0 for violation (‘2.0-viol’), and version 2.0 for correctness (‘2.0-corr’). Table 1 lists all validators, their references, supported languages, and supported formats. Figure 2 shows how the number of available witness validators developed over the last ten years. There is strong interest in developing tools for validation of verification results, which is a valuable enabling factor for the use of imprecise techniques in verification tools that compute invariants or error paths.

## 2.5 Safeguarding Software Development

Now we are equipped with witnesses and can address the three problems outlined in the motivation Sect. 1. First, programs should be accompanied by behavioral specifications and verification should be applied to ensure that the program fulfills the specification. Ideally, the programs are annotated with assertions [56], which are an in-code form of specification. If code is AI-generated, the assertions (and other annotations [57]) should also be generated, or manually added as the contract

between the generated code and its context. As the volume of AI-generated code will grow in the near future, but the precision of the code will still be suboptimal, the correctness of the software must be ensured by verification.

Second, if a particular verification tool is selected or automatically configured (for example by an AI-based selection), then the result of the verifier—that is, the verification witness—must be validated by a witness validator. This way, the verification process is less vulnerable to the risk that a verification tool was wrongly configured and produced a wrong result.

Third, and finally, there is no risk in using AI-generated invariants *if* the invariants are validated by a witness validator. That is, the generated invariants are put into a correctness witness and then given together with the program to the validator to be checked for validity. Note that it is not important for the validation process whether the invariants are annotated in the program or given as a correctness witness: both are interchangeable [58].

### 3 Conclusion

Imprecise approaches (such as AI-based code generation and AI-based invariant generation) can safely be used in the development and verification of software systems if the results are independently checked for correctness. That is, generated code should be analyzed to make sure it adheres to the specification, and generated invariants should be checked by witness-based result validators. The competition on software verification (SV-COMP 2024) has evaluated 16 validators for software witnesses (correctness witnesses and violation witnesses) and shown that their quality is very good. In conclusion, it seems that imprecise AI-based techniques can be empowered by techniques that safeguard the result, like witness-based result validation.

**Funding Statement.** My work on this topic was supported in part by the Deutsche Forschungsgemeinschaft (DFG) – 418257054 (COOP).

**Acknowledgements.** I would like to thank M. Dangl, D. Dietsch, M. Heizmann, T. Lemberger, A. Stahlbauer, and M. Tautschnig for the collaboration on developing the concepts, format, and tools for the verification witnesses in format 1.0 [14, 15], and P. Ayaziová, M. Lingsch-Rosenfeld, M. Spiessl, J. Strejček, and N. Weise for the collaboration on developing the new witness format 2.0 [40]. I am very grateful that the idea of verification witnesses became widely adopted in the community of software verification [59] (more than 60 verifiers generate witnesses) and the developers of the validators (16 validators were created, see Table 1) to support witness validation.

### References

1. Wang, S., Geng, M., Lin, B., Sun, Z., Wen, M., Liu, Y., Li, L., Bissyandé, T.F., Mao, X.: Natural language to code: How far are we? In: Proc. FSE. pp. 375–387. ACM (2023). <https://doi.org/10.1145/3611643.3616323>

2. Shin, J., Nam, J.: A survey of automatic code generation from natural language. *J. Inform. Processing Systems* **17**(3), 537–555 (June 2021). <https://doi.org/10.3745/JIPS.04.0216>
3. Rice, J.R.: The algorithm selection problem. *Adv. Comput.* **15**, 65–118 (1976). [https://doi.org/10.1016/S0065-2458\(08\)60520-3](https://doi.org/10.1016/S0065-2458(08)60520-3)
4. Richter, C., Hüllermeier, E., Jakobs, M.C., Wehrheim, H.: Algorithm selection for software validation based on graph kernels. *Autom. Softw. Eng.* **27**(1), 153–186 (2020). <https://doi.org/10.1007/s10515-020-00270-x>
5. Richter, C., Wehrheim, H.: PESCO: Predicting sequential combinations of verifiers (competition contribution). In: *Proc. TACAS* (3). pp. 229–233. LNCS 11429, Springer (2019). [https://doi.org/10.1007/978-3-030-17502-3\\_19](https://doi.org/10.1007/978-3-030-17502-3_19)
6. Leeson, W., Dwyer, M.B.: Algorithm selection for software verification using graph neural networks. *arXiv/CoRR* **2201**(11711) (January 2022). <https://doi.org/10.48550/arXiv.2201.11711>
7. Leeson, W., Dwyer, M.: GRAVES-CPA: A graph-attention verifier selector (competition contribution). In: *Proc. TACAS* (2). pp. 440–445. LNCS 13244, Springer (2022). [https://doi.org/10.1007/978-3-030-99527-0\\_28](https://doi.org/10.1007/978-3-030-99527-0_28)
8. Beyer, D., Keremoglu, M.E.: CPACHECKER: A tool for configurable software verification. In: *Proc. CAV*. pp. 184–190. LNCS 6806, Springer (2011). [https://doi.org/10.1007/978-3-642-22110-1\\_16](https://doi.org/10.1007/978-3-642-22110-1_16)
9. Baier, D., Beyer, D., Chien, P.C., Jakobs, M.C., Jankola, M., Kettl, M., Lee, N.Z., Lemberger, T., Lingsch-Rosenfeld, M., Wachowitz, H., Wendler, P.: Software verification with CPACHECKER 3.0: Tutorial and user guide. In: *Proc. FM*. LNCS, Springer (2024)
10. Si, X., Dai, H., Raghothaman, M., Naik, M., Song, L.: Learning loop invariants for program verification. In: *Proc. NeurIPS*. pp. 7762–7773. Curran Associates (2018), <https://dl.acm.org/doi/pdf/10.5555/3327757.3327873>
11. Giacobbe, M., Kröning, D., Parsert, J.: Neural termination analysis. In: *Proc. ES-EC/FSE*. pp. 633–645. ACM (2022). <https://doi.org/10.1145/3540250.3549120>
12. Song, P., Yang, K., Anandkumar, A.: Towards large language models as copilots for theorem proving in LEAN. In: *Proc. MATH-AI* (2023), <https://mathai2023.github.io/papers/4.pdf>
13. Janßen, C., Richter, C., Wehrheim, H.: Can ChatGPT support software verification? In: *Proc. FASE*. pp. 266–279. LNCS 14573, Springer (2024). [https://doi.org/10.1007/978-3-031-57259-3\\_13](https://doi.org/10.1007/978-3-031-57259-3_13)
14. Beyer, D., Dangl, M., Dietsch, D., Heizmann, M., Stahlbauer, A.: Witness validation and stepwise testification across software verifiers. In: *Proc. FSE*. pp. 721–733. ACM (2015). <https://doi.org/10.1145/2786805.2786867>
15. Beyer, D., Dangl, M., Dietsch, D., Heizmann, M., Lemberger, T., Tautschnig, M.: Verification witnesses. *ACM Trans. Softw. Eng. Methodol.* **31**(4), 57:1–57:69 (2022). <https://doi.org/10.1145/3477579>
16. Sternagel, C., Thiemann, R.: The certification problem format. In: *Proc. UITP*. pp. 61–72. EPTCS 167, EPTCS (2014). <https://doi.org/10.4204/EPTCS.167.8>
17. Heule, M.J.H.: The DRAT format and drat-trim checker. *CoRR* **1610**(06229) (October 2016), <https://arxiv.org/abs/1610.06229>
18. Wetzler, N., Heule, M.J.H., Jr., W.A.H.: DRAT-TRIM: Efficient checking and trimming using expressive clausal proofs. In: *Proc. SAT*. pp. 422–429. LNCS 8561, Springer (2014). [https://doi.org/10.1007/978-3-319-09284-3\\_31](https://doi.org/10.1007/978-3-319-09284-3_31)
19. Bury, G.: DOLMEN: A validator for SMT-LIB and much more. In: *Proc. SMT Workshop*. CEUR Workshop Proceedings, CEUR (2021), <https://ceur-ws.org/Vol-2908/short4.pdf>

20. Bury, G., Bobot, F.: Verifying models with DOLMEN. In: Proc. SMT Workshop. CEUR Workshop Proceedings, CEUR (2023), <https://ceur-ws.org/Vol-3429/short9.pdf>
21. Yu, E., Biere, A., Heljanko, K.: Progress in certifying hardware model checking results. In: Proc. CAV. pp. 363–386. LNCS 12760, Springer (2021). [https://doi.org/10.1007/978-3-030-81688-9\\_17](https://doi.org/10.1007/978-3-030-81688-9_17)
22. Ádám, Z., Beyer, D., Chien, P.C., Lee, N.Z., Sirrenberg, N.: BTOR2-CERT: A certifying hardware-verification framework using software analyzers. In: Proc. TACAS (3). pp. 129–149. LNCS 14572, Springer (2024). [https://doi.org/10.1007/978-3-031-57256-2\\_7](https://doi.org/10.1007/978-3-031-57256-2_7)
23. McConnell, R.M., Mehlhorn, K., Näher, S., Schweitzer, P.: Certifying algorithms. *Computer Science Review* 5(2), 119–161 (2011). <https://doi.org/10.1016/j.cosrev.2010.09.009>
24. Castaño, R., Braberman, V.A., Garbervetsky, D., Uchitel, S.: Model checker execution reports. In: Proc. ASE. pp. 200–205. IEEE (2017). <https://doi.org/10.1109/ASE.2017.8115633>
25. OASIS: Static analysis results interchange format (sarif) version 2.0 (2019), <https://docs.oasis-open.org/sarif/sarif/v2.0/csprd02/sarif-v2.0-csprd02.html>
26. Rushby, J.M.: An Evidential Tool Bus. In: Proc. ICFEM. pp. 36–36. LNCS 3785, Springer (2005). [https://doi.org/10.1007/11576280\\_3](https://doi.org/10.1007/11576280_3)
27. Cruanes, S., Hamon, G., Owre, S., Shankar, N.: Tool integration with the Evidential Tool Bus. In: Proc. VMCAL. pp. 275–294. LNCS 7737, Springer (2013). [https://doi.org/10.1007/978-3-642-35873-9\\_18](https://doi.org/10.1007/978-3-642-35873-9_18)
28. Cruanes, S., Heymans, S., Mason, I., Owre, S., Shankar, N.: The semantics of Datalog for the Evidential Tool Bus. In: Specification, Algebra, and Software. pp. 256–275. Springer (2014). [https://doi.org/10.1007/978-3-642-54624-2\\_13](https://doi.org/10.1007/978-3-642-54624-2_13)
29. Margaria, T., Nagel, R., Steffen, B.: Remote integration and coordination of verification tools in JETI. In: Proc. ECBS. pp. 431–436 (2005). <https://doi.org/10.1109/ECBS.2005.59>
30. Steffen, B., Margaria, T., Braun, V.: The Electronic Tool Integration platform: Concepts and design. *STTT* 1(1-2), 9–30 (1997). <https://doi.org/10.1007/s100090050003>
31. Margaria, T.: Web services-based tool-integration in the ETI platform. *Software and Systems Modeling* 4(2), 141–156 (2005). <https://doi.org/10.1007/s10270-004-0072-z>
32. Beyer, D., Wehrheim, H.: Verification artifacts in cooperative verification: Survey and unifying component framework. In: Proc. ISoLA (1). pp. 143–167. LNCS 12476, Springer (2020). [https://doi.org/10.1007/978-3-030-61362-4\\_8](https://doi.org/10.1007/978-3-030-61362-4_8)
33. Beyer, D., Kanav, S.: CoVERITEAM: On-demand composition of cooperative verification systems. In: Proc. TACAS. pp. 561–579. LNCS 13243, Springer (2022). [https://doi.org/10.1007/978-3-030-99524-9\\_31](https://doi.org/10.1007/978-3-030-99524-9_31)
34. Beyer, D., Kanav, S., Wachowitz, H.: CoVERITEAM SERVICE: Verification as a service. In: Proc. ICSE, companion. pp. 21–25. IEEE (2023). <https://doi.org/10.1109/ICSE-Companion58688.2023.00017>
35. Beyer, D.: Conservation and accessibility of tools for formal methods. In: Proc. Festschrift Podelski 65th Birthday. Springer (2024), [https://www.sosy-lab.org/research/pub/2024-Podelski65.Conservation\\_and\\_Accessibility\\_of\\_Tools\\_for\\_Formal\\_Methods.pdf](https://www.sosy-lab.org/research/pub/2024-Podelski65.Conservation_and_Accessibility_of_Tools_for_Formal_Methods.pdf)
36. Beyer, D., Dangl, M., Dietsch, D., Heizmann, M.: Correctness witnesses: Exchanging verification results between verifiers. In: Proc. FSE. pp. 326–337. ACM (2016). <https://doi.org/10.1145/2950290.2950351>



37. Beyer, D.: Software verification and verifiable witnesses (Report on SV-COMP 2015). In: Proc. TACAS. pp. 401–416. LNCS 9035, Springer (2015). [https://doi.org/10.1007/978-3-662-46681-0\\_31](https://doi.org/10.1007/978-3-662-46681-0_31)
38. Beyer, D.: Competition on software verification and witness validation: SV-COMP 2023. In: Proc. TACAS (2). pp. 495–522. LNCS 13994, Springer (2023). [https://doi.org/10.1007/978-3-031-30820-8\\_29](https://doi.org/10.1007/978-3-031-30820-8_29)
39. Beyer, D., Strejček, J.: Case study on verification-witness validators: Where we are and where we go. In: Proc. SAS. pp. 160–174. LNCS 13790, Springer (2022). [https://doi.org/10.1007/978-3-031-22308-2\\_8](https://doi.org/10.1007/978-3-031-22308-2_8)
40. Ayaziová, P., Beyer, D., Lingsch-Rosenfeld, M., Spiessl, M., Strejček, J.: Software verification witnesses 2.0. In: Proc. SPIN. Springer (2024)
41. Beyer, D.: State of the art in software verification and witness validation: SV-COMP 2024. In: Proc. TACAS (3). pp. 299–329. LNCS 14572, Springer (2024). [https://doi.org/10.1007/978-3-031-57256-2\\_15](https://doi.org/10.1007/978-3-031-57256-2_15)
42. Beyer, D., Friedberger, K.: Violation witnesses and result validation for multi-threaded programs. In: Proc. ISOLA (1). pp. 449–470. LNCS 12476, Springer (2020). [https://doi.org/10.1007/978-3-030-61362-4\\_26](https://doi.org/10.1007/978-3-030-61362-4_26)
43. Beyer, D., Dangl, M., Lemberger, T., Tautschnig, M.: Tests from witnesses: Execution-based validation of verification results. In: Proc. TAP. pp. 3–23. LNCS 10889, Springer (2018). [https://doi.org/10.1007/978-3-319-92994-1\\_1](https://doi.org/10.1007/978-3-319-92994-1_1)
44. Beyer, D., Spiessl, M.: METAVAL: Witness validation via verification. In: Proc. CAV. pp. 165–177. LNCS 12225, Springer (2020). [https://doi.org/10.1007/978-3-030-53291-8\\_10](https://doi.org/10.1007/978-3-030-53291-8_10)
45. J. Švejda, Berger, P., Katoen, J.P.: Interpretation-based violation witness validation for C: NITWIT. In: Proc. TACAS. pp. 40–57. LNCS 12078, Springer (2020). [https://doi.org/10.1007/978-3-030-45190-5\\_3](https://doi.org/10.1007/978-3-030-45190-5_3)
46. Ponce-De-Leon, H., Haas, T., Meyer, R.: DARTAGNAN: Smt-based violation witness validation (competition contribution). In: Proc. TACAS (2). pp. 418–423. LNCS 13244, Springer (2022). [https://doi.org/10.1007/978-3-030-99527-0\\_24](https://doi.org/10.1007/978-3-030-99527-0_24)
47. Howar, F., Mues, M.: GWIT (competition contribution). In: Proc. TACAS (2). pp. 446–450. LNCS 13244, Springer (2022). [https://doi.org/10.1007/978-3-030-99527-0\\_29](https://doi.org/10.1007/978-3-030-99527-0_29)
48. Ayaziová, P., Strejček, J.: SYMBIOTIC-WITCH 2: More efficient algorithm and witness refutation (competition contribution). In: Proc. TACAS (2). pp. 523–528. LNCS 13994, Springer (2023). [https://doi.org/10.1007/978-3-031-30820-8\\_30](https://doi.org/10.1007/978-3-031-30820-8_30)
49. Wu, T., Schrammel, P., Cordeiro, L.: WIT4JAVA: A violation-witness validator for Java verifiers (competition contribution). In: Proc. TACAS (2). pp. 484–489. LNCS 13244, Springer (2022). [https://doi.org/10.1007/978-3-030-99527-0\\_36](https://doi.org/10.1007/978-3-030-99527-0_36)
50. Bajczi, L., Ádám, Z., Micskei, Z.: CONCURRENTWITNESS2TEST: Test-harnessing the power of concurrency (competition contribution). In: Proc. TACAS (3). pp. 330–334. LNCS 14572, Springer (2024). [https://doi.org/10.1007/978-3-031-57256-2\\_16](https://doi.org/10.1007/978-3-031-57256-2_16)
51. Saan, S., Erhard, J., Schwarz, M., Bozhilov, S., Holter, K., Tilscher, S., Vojdani, V., Seidl, H.: GOBLINT VALIDATOR: Correctness witness validation by abstract interpretation (competition contribution). In: Proc. TACAS (3). pp. 335–340. LNCS 14572, Springer (2024). [https://doi.org/10.1007/978-3-031-57256-2\\_17](https://doi.org/10.1007/978-3-031-57256-2_17)
52. Beyer, D., Spiessl, M.: LIV: A loop-invariant validation using straight-line programs. In: Proc. ASE. pp. 2074–2077. IEEE (2023). <https://doi.org/10.1109/ASE56229.2023.00214>
53. Monat, R., Milanese, M., Parolini, F., Boillot, J., Ouadjaout, A., Miné, A.: MOPSA-C: Improved verification for C programs, simple validation of correctness witnesses

- (competition contribution). In: Proc. TACAS (3). pp. 387–392. LNCS 14572, Springer (2024). [https://doi.org/10.1007/978-3-031-57256-2\\_26](https://doi.org/10.1007/978-3-031-57256-2_26)
54. Ayaziová, P., Strejček, J.: WITCH 3: Validation of violation witnesses in the witness format 2.0 (competition contribution). In: Proc. TACAS (3). pp. 341–346. LNCS 14572, Springer (2024). [https://doi.org/10.1007/978-3-031-57256-2\\_18](https://doi.org/10.1007/978-3-031-57256-2_18)
  55. Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M.S.: GraphML progress report. In: Graph Drawing. pp. 501–512. LNCS 2265, Springer (2001). [https://doi.org/10.1007/3-540-45848-4\\_59](https://doi.org/10.1007/3-540-45848-4_59)
  56. Turing, A.: Checking a large routine. In: Report on a Conference on High Speed Automatic Calculating Machines. pp. 67–69. Cambridge Univ. Math. Lab. (1949), <https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-8>
  57. Baudin, P., Cuoq, P., Filliâtre, J.C., Marché, C., Monate, B., Moy, Y., Prevosto, V.: ACSL: ANSI/ISO C specification language version 1.17 (2021), available at <https://frama-c.com/download/acsl-1.17.pdf>
  58. Beyer, D., Spiessl, M., Umbricht, S.: Cooperation between automatic and interactive software verifiers. In: Proc. SEFM. p. 111–128. LNCS 13550, Springer (2022). [https://doi.org/10.1007/978-3-031-17108-6\\_7](https://doi.org/10.1007/978-3-031-17108-6_7)
  59. Beyer, D., Podelski, A.: Software model checking: 20 years and beyond. In: Principles of Systems Design. pp. 554–582. LNCS 13660, Springer (2022). [https://doi.org/10.1007/978-3-031-22337-2\\_27](https://doi.org/10.1007/978-3-031-22337-2_27)

**Open Access.** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# End-to-End AI Generated Runtime Verification from Natural Language Specification

Itay Cohen and Doron Peled<sup>(✉)</sup>

Bar Ilan University, 52900 Ramat Gan, Israel  
doron.peled@gmail.com

**Abstract.** We demonstrate how LLMs can be harnessed to synthesize runtime verification monitors from natural language specifications. We instruct the LLM through prompts to assemble the monitors through a sequence of stages. To start with, we employ the learned insights LLMs possess regarding linear temporal logic for translating natural language specifications into propositional past time LTL formulas. Next, we utilize a sequence of prompts to synthesize from the LTL specification a runtime verification monitor for the given formula. Part of these prompts enable the LLM to validate its own outputs, thereby significantly improving the probability of obtaining a correct monitor.

## 1 Introduction

Runtime verification (RV) [3–5, 10, 13, 16] allows monitoring the executions of a system either online or offline against a formal specification. The fact that each time a single execution is separately checked allows RV to be simpler than the more comprehensive model checking [8]. On the other hand, performing RV does not provide the level of correctness guaranteed by model checking. Still, RV plays an important role in the production of software systems; in particular, it allows shielding the system against malfunctions, and the collection of statistics about the executions of the system.

A recent trend uses LLM tools, such as ChatGPT and Gemini, for performing simple programming tasks. In fact, LLM-based tools, e.g., Copilot [19], aim specifically at lightweight synthesizing of code. While this kind of use for LLMs already appears on a small scale in the software development industry, the abilities of LLMs to provide useful code is considered to be somewhat limited. In essence, LLMs are trained to construct well-formed and informative sentences, responding to the requests that are prompted to them. This is achieved based on training appropriate neural networks using a huge number of text corpora. Among these texts, some programs and their corresponding descriptions are included, e.g., in repositories such as GitHub. However, LLMs typically do not include a serious deduction mechanism (as opposed to some shallow

---

This research was supported by the Israel Science Foundation grant “Validating and controlling software and hardware systems assisted by machine learning” No. 2454/23.

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 362–384, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_23](https://doi.org/10.1007/978-3-031-73741-1_23)

deduction that is inherent in the language mechanism). Such a mechanism is important for the construction of complex code that includes intricate dependencies between objects and calculations based on nontrivial decision making. Another phenomenon that limits the usability of LLMs for synthesizing code is that they tend to be unreliable, sometimes providing an incorrect answer to a given prompt, rather than admitting the limitation in their mechanism to deliver the desired answer. This phenomenon is often called *hallucination*. This can, in our context, produce incorrect code. Thus, LLMs are particularly useful for the automatic programming of *simple* code, e.g., the interface between the internal representation of two systems, or simulating a finite state system based on its transition system.

We undertake here the task of teaching LLMs (specifically, GPT and Gemini) how to synthesize a runtime verification mechanism, given a natural language description of the requirement that needs to be checked or enforced. Thus, a natural language description is translated into code that performs runtime verification for the described property. This exploits the abilities of LLMs to capture and apply descriptions given in natural languages. It is important for our task, that the essential processing involved in the RV algorithm is simple enough, i.e., does not require the ability to apply deep deduction. This allows us to better trust the outcome of our construction. Furthermore, we apply various checks, also embedded within the LLMs and described via prompts, which enhance the reliability of the generated code.

A central principle that we adopt is to partition the task into smaller steps that sequentially follow each other. This includes (a) the construction of a translation mechanism from natural language to temporal logic, based on training examples, (b) the construction of an appropriate syntax tree that corresponds to the obtained temporal specification, (c) the construction of code that synthesizes the runtime verification engine and (d) the application of various testing procedures. All of these steps are performed within the LLM tools. This decomposition of tasks ensures that each task is simple enough to be enforced within the LLM. This helps us to overcome the limitations of current LLM technologies and harness their advantages. It allows exploiting the interfacing with natural languages, already embedded within the mechanism. Further, it employs simple steps, where the reliability of each one of them is independent of the ability of deep deduction. Our approach also exploits some checks that can be simply called for by prompting the LLM.

It is important to note that part of our process is based on instructing the LLMs, through prompts, to exploit, as part of the generated code, a classical RV algorithm [14]. Although the code for this algorithm is not *dictated* per se, our prompts clearly follow it; our process does not include a stage where this algorithm is somehow discovered by the LLM. It would be interesting to see if an alternative, more autonomous description, can be alternatively used, where the LLM itself discovers some connections, e.g., between the provided subformulas of the specification and the progress of information along the occurrence of the

monitored events; essentially, this is the process of updating a *summary* of the execution that is needed with respect to monitoring the specification property.

The rest of the paper is organized as follows: Sect. 2 reviews LLMs and past time linear temporal logic, which is the temporal logic used for specifications in this paper. Each of the next sections describes a different stage in the process of synthesizing a runtime verification engine from natural language specification with the assistance of the LLM. Section 3 describes the various methods examined for translating a given natural language specification into past time LTL formulas. Section 4 delves into the process of converting these formulas into their abstract syntax tree representations through the use of LLMs. In this stage, we also employ validation prompts designed to increase the probability of obtaining a reliable output. Section 5 describes the stage where the output from the previous stage is used to construct a runtime verification monitor, using a single prompt that leverages the syntax tree based representation of the formula. In Sect. 6, we discuss the last stage of our pipeline, which involves using a version of the LLM that includes an interpreter component, to validate the monitor code synthesized in the previous stage.

## 2 Preliminaries

### 2.1 Propositional Past Time Linear Temporal Logic

Propositional past time linear temporal logic (PLTL) is a specification formalism that allows expressing safety properties [1]. The restriction to past time allows interpreting the formulas on finite traces.

**Syntax.** The formulas of propositional past time linear temporal logic are defined using the following grammar:

$$\varphi ::= \text{true} \mid q \mid \neg\varphi \mid (\varphi \wedge \psi) \mid (\varphi \vee \psi) \mid (\varphi \rightarrow \psi) \mid (\varphi \leftrightarrow \psi) \mid (\varphi \mathcal{S} \psi) \mid \ominus\varphi$$

The symbol  $q$  denotes a Boolean proposition over some finite set  $A$  of propositions. The operators  $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$  have the same meaning as in propositional calculus.

The temporal operators have the following informal meaning: the formula  $(\varphi \mathcal{S} \psi)$ , which reads as  $\varphi$  *since*  $\psi$ , means that  $\psi$  holds in some prefix of the current trace, and for all prefixes between that one and the current trace,  $\varphi$  holds. The *since* operator is the past dual of the future time *until* modality. The property  $\ominus\varphi$  (previous-time  $\varphi$ ) means that  $\varphi$  is true in the current trace that is obtained from the current one by omitting the last event. This is the past dual of the future time *next* modality. We can also define the following additional derived temporal operators:  $\diamond\varphi = (\text{true} \mathcal{S} \varphi)$  (“past” or “once”), and  $\boxplus\varphi = \neg \diamond \neg\varphi$  (“always in the past” or “historically”).

**Semantics.** A past time LTL formula is interpreted over a trace (or an observation) of events of the form  $e_1e_2e_3\dots$ . Each event  $e$  is interpreted (labeled) with a finite set of propositions  $L(e) \subseteq A$ . This labeling is obtained when the event is

observed. Let  $t_i$  denote the suffix trace  $e_i e_{i+1} e_{i+2} \dots$ . The semantics of the logic is as follows:

- $t_i \models \text{true}$       is always true,
- $t_i \models q$       iff  $q \in L(e_i)$ ,
- $t_i \models \neg\varphi$       iff it is not the case that  $t_i \models \varphi$ ,
- $t_i \models (\varphi \text{ op } \psi)$       iff  $t_i \models \varphi$  op  $t_i \models \psi$ , when  $\text{op} \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$ ,
- $t_i \models \ominus\varphi$       iff  $i > 1$  and  $t_{i-1} \models \varphi$ ,
- $t_i \models (\varphi \mathcal{S} \psi)$       iff  $t_j \models \psi$  for some  $1 \leq j \leq i$  and  $t_k \models \varphi$  for all  $j < k \leq i$ .

## 2.2 Large Language Models and Linear Temporal Logic

Translating natural language to linear temporal logic has been studied before the deep learning era. These works mostly focused on handling structured sentences that were generated based on a certain grammar [7]. Recently, neural networks based methods have gained popularity over traditional approaches for translating into temporal logics.

Large language models (LLMs) have become a fundamental part of natural language processing (NLP), significantly enhancing the ability of computers to understand and mimic human language. These models, built on architectures such as *transformers* [21], have shown remarkable capabilities in various applications, from machine translation and question-answering systems to content generation and sentiment analysis. The essence of LLMs lies in their ability to process and generate text by learning patterns and relationships within vast datasets of human language. Their development involves training on extensive corpora, often encompassing billions of words sourced from books, articles, and the internet. This training enables the models to grasp the nuances of language, including grammar, idioms, and context. Notable LLM examples include OpenAI's GPT [18], Google's Gemini [2], and the more recent open-source models like Meta's LLaMA [20] and Mistral [15].

In [12], an LLM is fine-tuned based on a large dataset of LTL formulas. Other works focus on translating natural language specifications of a specific domain, such as grounded robotics [22]. Few-shot prompting techniques [6], in which we provide the model with input and output examples, also became prevalent in this domain. In [17], a sequence of prompts was used to translate natural language specifications for robots into an LTL specification, and [9] presented a framework that uses few-shot prompting to derive LTL formulas from unstructured natural language in an interactive way that involves human feedback.

## 3 Translating Natural Language Specifications into Past Time LTL Formulas

We present an approach for synthesizing a runtime verification mechanism given a natural language description of the requirement that needs to be checked. This general task is divided into four distinct sequential stages, each of which can be

entirely executed using LLM tools. We conduct empirical experiments for each stage, to demonstrate its significance. This section outlines the first stage, with subsequent stages detailed in the following sections.

In this work, we consider natural language specifications that correspond to propositional past time LTL formulas. Such specification is appropriate for defining an agent that operates in an environment and can transition through different states or conditions while running. The majority of the examples tested involve agents like robots, which possess various internal states or conditions. An internal state may suggest, for example, that the agent is at a certain location (landmark) in the environment. An example of a natural language specification of this domain can be, for example, “the robot had to go to the repair station before entering sleep mode”. In this example, the agent is a robot, and there is one landmark associated with it (the repair station), and one additional condition that may hold (entering sleep mode). Note that the specification domain is not restricted to the agent-environment setup described earlier. It can cover different areas, like specifications that describe properties software systems need to follow during an execution. However, our approach is versatile and can easily be modified to accommodate these variations.

At this stage, we specifically aim at translating natural language specifications into propositional past time LTL formulas. To do that, we decompose this stage into two subtasks, where in each subtask we harness the LLM’s strengths via specific prompting techniques to increase the translation success rate:

- *Identification of landmarks and conditions.* An agent can be at a certain internal state (location), or condition, and this fact can be considered as a Boolean proposition. In this subtask, the LLM identifies these elements and replaces them with Boolean variables, similarly to the translation approach described in [17]. This results in a specification in an intermediate form, where there are variable placeholders instead of real propositions. An instance of an intermediate form specification could be: “whenever  $q_1$  held and  $q_2$  happened in the previous step, then one step before being at  $q_2$  it was at  $q_3$ ”.
- *Intermediate form translation to past time LTL.* We once again employ the LLM to directly convert the specification in its intermediate form into a past time LTL formula, incorporating placeholders as Boolean variables.

### 3.1 Landmarks and Conditions Identification

We ask the LLM to identify locations or conditions of an agent that operates in an environment. Following the findings presented in [17], we also demonstrate high performance with GPT-4 on this subtask by providing a prompt with a clear description and a few examples (see Listing 1). The prompt, initially aligned with the scenario of an agent in an environment, can be revised to suit other natural language specification variants.

---

Your task is to repeat exact strings that refer to landmarks or conditions of an object from a given utterance, and then replace each landmark or condition with a different variable. the variables are q1, q2, q3 etc. After identifying the landmarks or conditions, write the modified sentence inside hashtags. You are provided with the following three examples:

1. utterance: the robot had to go to the repair station, and then to the charging station.

Landmarks/conditions: the repair station; the charging station.

new sentence: # the robot had to go to q1, and then to q2. #

2. utterance: the vehicle should start at the garage, then go to the grocery store and finally head back to the garage.

Landmarks/conditions: the garage; the grocery store.

new sentence: # the vehicle should start at q1, then go to q2 and finally head back to q1. #

3. utterance: The phone must always be on silent mode, and eventually should receive an MMS message.

Landmarks/conditions: be on silent mode; receive an MMS message.

new sentence: # The phone must always q1, and eventually should q2. #

Now, your task is to consider the following sentence:

"{insert\_your\_natural\_language\_spec\_here}".

---

**Listing 1.** The prompt that was used for landmarks and conditions extraction. The input specification should be provided inside the curly brackets, and by using hashtags, the LLM is guided to deliver outputs that can be easily parsed.

### 3.2 Translating from Intermediate Form Specification Into Past Time LTL

Many research efforts on temporal logic translations from natural language are focused on translations to LTL, with LLMs playing a key role in recent techniques. LLMs provide better results for subjects that were more common on their training dataset. Consequently, as LLMs have evolved, there has been a significant improvement in translation performance for formalisms such as LTL, which have extensive literature coverage. On the other hand, for formalisms that were less covered in the literature, like past time LTL, which is our selected intermediate specification formalism, the quality of these translations often falls short.

To demonstrate the limited knowledge of LLMs regarding past time LTL, we tasked GPT-4 to list the temporal operators of past time LTL and concisely define their semantics. In its response, the semantics proposed for some operators did not align with the standard semantics associated with past time LTL operators. More details on this are provided in Appendix A. To mitigate this, we experimented with two main approaches:

- *Fine-tuning based approach.* We created a dataset that consisted of 570 different pairs of sentences in natural language and their corresponding past time LTL formulas. Then, we fine-tuned GPT-3.5 on these pairs using a service provided by OpenAI<sup>1</sup>, without introducing the model with past time LTL.

<sup>1</sup> <https://platform.openai.com/docs/guides/fine-tuning>.



We developed our own dataset due to the absence of such resources, as past time LTL is a less common variant of LTL. We aimed to test the model’s ability to generalize beyond the provided examples to those that it has never seen before. Initially, we started with manually creating a smaller dataset of 190 different examples. After some preliminary experiments, we noticed that our fine-tuned model that is based on this dataset had a poor generalization ability; therefore we chose to augment it with the assistance of GPT-3.5. By asking it for two rephrasings of each natural language specification in the dataset, we were able to increase the dataset size threefold and introduce a greater linguistic diversity.

- *Prompting based approach.* Our goal here was to leverage the existing translation abilities of different LLMs, while introducing them to the concept of past time LTL using some prompting techniques. Similar approaches were considered in [9,11,17]. Initially, we devised a prompt detailing the various operators in past time LTL and their meanings. For certain operators, we highlighted their resemblance to future-time LTL operators, such as the connection between “always” and “historically”. We then expanded the prompt by adding a second part, providing the LLM with three simple examples of translations (few-shot prompting). The full prompt is provided in Listing 2.

---

```
Your task is to translate the below natural language sentence into a
past-time LTL formula and explain your translation step by step. Remember
that P means "in the last step"/"in the previous timestep", S means "since",
H means "historically" and O means "in the past"/"once". The formula should
only contain atomic propositions or operators &&, ||, !, -, <->, P, S, H,
O. The atomic propositions would be q1, q2, q3 etc. Important note: do not
use the "globally" (G) operator from LTL, instead, use the H operator
("always in the past", "historically").
```

```
The "S" operator should be used in the following way:
natural language: "the robot was previously at q2, but since then it was
always at q1."
past-time LTL formula: " q1 S q2 ".
```

```
You are provided with three simple examples:
1. natural language: "The robot was at q1 at all times."
past-time LTL formula: # H(q1) #.
2. natural language: "Whenever the robot is was at q1, it is not at q2."
past-time LTL formula: # H(q1 -> !(q2)) #.
3. natural language: "In the past, the robot was at q3."
past-time LTL formula: # O(q3) #.
```

```
Write your final answer inside hashtags.
```

```
Please translate this sentence:
"{insert_your_intermediate_form_sentence_here}"
```

---

**Listing 2.** The few-shot prompt that was used for translating the intermediate form sentence to past time LTL. The sentence should be provided inside the curly brackets, and by using hashtags, the LLM is guided to deliver outputs that can be easily parsed.

**Evaluation.** To evaluate the two approaches, we obtained a benchmark dataset with challenging examples. We based our dataset on an existing dataset that was used in [9]. In that work, five experts in the field were asked to provide

pairs of natural language specifications and their formalizations into LTL. We observed that the majority of the examples in this dataset could be adjusted to reflect a past temporal perspective rather than a future one. Consequently, we manually altered 30 out of 36 natural language sentences to adopt a past form and paired them with their matching past time LTL translations. Note that the translation to a past form does not have to preserve the original semantics; it simply involves deriving one sentence from another. We provide below two examples of these modifications of temporal perspective in Table 1.

**Table 1.** Two examples of the change in the temporal perspective.

Original Sentence	Sentence in Past Form
$q_1$ never holds from some point in time on.	From the beginning of the execution until some point in the past, $q_1$ did not hold.
Whenever $q_1$ holds and $q_2$ holds in the next step, then $q_3$ holds one step after $q_2$ .	Whenever $q_1$ held and $q_2$ happened in the previous step, then one step before being at $q_2$ it was at $q_3$ .

We started by evaluating our fine-tuning based approach for translating intermediate form specifications to past time LTL. Before evaluating it on the benchmark dataset, we explored its generalization abilities. We noted that the fine-tuned model often struggled to generalize to past time LTL patterns that were slightly different variants of the 570 examples in its dataset. For instance, the model encountered in its dataset the following sentences and their correct translations: “the robot was at  $q_5$  three steps ago”, “the robot visited  $q_4$  two steps ago”. However, it could not properly translate the sentence: “if the robot was at  $q_7$  three steps ago, then it was at  $q_3$  two steps ago”. When we assessed its performance on our benchmark dataset, it reached a 30.0% accuracy rate.

We tested the prompting based approach on the benchmark dataset using four different configurations:

- *GPT-3.5-no-examples* - we equipped GPT-3.5 with the first part of the prompt that discusses the different temporal operators of past time LTL and their correspondence to regular LTL, without providing the few-shot translation examples at this point.
- *GPT-3.5-few-shot* - we used GPT-3.5 with the full prompt shown in Listing 2, including the three few-shot translation examples.
- *Gemini-few-shot* - we equipped Google’s language model Gemini [2] with the full prompt.
- *GPT-4-few-shot* - we used GPT-4 with the full prompt.

During the experiments, we noticed that the configuration *GPT-3.5-no-examples* tends to mix the regular LTL operators with the past time LTL operators. It seems that it does not recognize the concept of “always in the past” and

tends to confuse between “previous-time” and “past” operators. Moreover, the word “whenever” was not associated with the concept of the “historically” operator. Overall, this setup achieved an accuracy rate of 10.0% on the benchmark dataset.

After adding the three examples to the prompt of GPT-3.5 (*GPT-3.5-few-shot*), there was a noticeable improvement in its performance. However, the model still struggled to generalize in areas where it previously had difficulties. For instance, consider the natural language specification: “the robot is always at  $q_1$ , and whenever it is at  $q_3$ , it cannot be at  $q_2$ ”. The correct translation of this sentence is  $(\Box q_1 \wedge \Box(q_3 \rightarrow \neg q_2))$ , while the output translation was  $(\Box q_1 \wedge (q_3 \rightarrow \neg q_2))$  (i.e., the *whenever* clause only applies to the first event). Overall, this setup achieved an accuracy rate of 43.33% on the benchmark dataset. The results were similar when we evaluated Gemini with the full prompt (*Gemini-few-shot*). In this case, the setup achieved an accuracy rate of 46.67%. Gemini seemed to misinterpret the notion of the “since” operator, and occasionally failed to differentiate between “previous-time” and “past”.

The evaluation of the GPT-4 setup (*GPT-4-few-shot*) showed remarkable progress, outperforming all other configurations we examined. It successfully translated 28 of 30 sentences, achieving a 93.33% accuracy rate. It demonstrated a better grasp of the “since” operator and a more accurate use of the “previous-time” operator. The evaluation results of the two approaches and all the configurations are listed in Table 2. Due to its superior performance, we selected the *GPT-4-few-shot* configuration as the initial stage (translation stage) in the full monitor synthesis process.

**Table 2.** Evaluation accuracies (%) of all the translation approaches on the benchmark dataset.

Setup	GPT-3.5-ft	GPT-3.5-no-ex	GPT-3.5-fs	Gemini-fs	GPT-4-fs
<b>Accuracy</b>	30.00	10.00	43.33	46.67	<b>93.33</b>

## 4 Translating Past Time LTL Formula into Abstract Syntax Tree

The next stage towards synthesizing a propositional past time LTL monitor is teaching the LLM to analyze the formula that was generated in the previous stage. This analysis is conducted exclusively through prompts given to the LLM, specifically GPT-4 in this instance. We split this task into two subtasks, where each subtask is self-validated by the LLM, to increase the probability for a reliable output.

In the first subtask, we ask the LLM to create an abstract syntax tree (AST) out of the propositional past time LTL formula. In the prompt, as presented in

Listing 3, we included all the possible operators and specified their arity. Aiming at a more standardized tree structure, we specified two further requirements: (1) leaves in the AST must only contain Boolean variables, without any operators, and (2) each AST node may have two immediate descendants at most. An example of a partial response for this prompt is shown in Listing 4. This subtask is followed by an additional prompt, presented in Listing 5, in which the LLM self-validates its previous output. In this validation, the LLM counts the number of leaves in the AST and checks if it is consistent with the number of Boolean variable occurrences. If this validation fails, we repeat this stage.

We experimented with two different approaches to repeat a subtask after failure. The first approach involved informing the model of its incorrect output and requesting an alternative solution. This method tended to yield solutions that were originally less likely to be produced and frequently incorrect. As a second approach, we removed the latest incorrect solution from the model message history, giving the model a fresh start as if we started a new conversation. This approach yielded better alternative solutions, so we applied it for all instances where repeating a task was required.

---

Past-time linear temporal logic is a language with  $q_1, q_2, \dots, q_n$  as boolean variables. "!", "P", "H", "O" are unary operators, while "&&", "||", "->", "<->", "S" are binary operators in this language.

Given a valid expression in this language, your task is to generate the abstract syntax tree of it.

Do it for the following expression inside the backticks:

{insert\_past\_time\_ltl\_formula\_here}'.

Generate the abstract syntax tree while meeting the following requirements:

-do not attach unary nodes to the leaves.

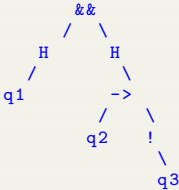
-every node in the tree must have two children at most.

---

**Listing 3.** A prompt that creates a standardized AST out of a past time LTL formula. The formula should be provided inside the curly brackets.

---

### Abstract Syntax Tree Representation ###



This tree represents the structure of the expression 'H(q1) && H(q2 -> !q3)' according to the requirements:

- Unary operators ('H' and '!') are not attached directly to leaves ('q1', 'q2', 'q3').

- Every node has at most two children, adhering to the binary structure of the tree.

---

**Listing 4.** A partial response from GPT-4 to a prompt that asked for a standardized AST for the formula  $H(q_1) \wedge H(q_2 \rightarrow \neg q_3)$ .

---

```
Your next task is to count the number of leaves in the abstract syntax tree,
and the number of boolean variable occurrences in the formula provided
earlier.
Please write your final answer in the following format:

leaves count: <number>
variable occurrences count: <number>
```

---

**Listing 5.** A self-validation prompt that verifies that the number of leaves in the AST is consistent with the number of Boolean variable occurrences in the formula.

In the second subtask, we take every non-leaf node in the AST and formulate an equation that outlines its relation with its immediate descendants, which are past time LTL subformulas. In addition, every leaf in the AST would be associated with its corresponding Boolean variable. Here, the LLM is required to assign an index to every node and generate the corresponding equations. The prompt for this subtask is described in Listing 6. For instance, the following equations should be generated for the formula  $\diamond (q_1 \wedge q_2)$ :

$$\begin{aligned} node_1 &= q_1 \\ node_2 &= q_2 \\ node_3 &= node_1 \wedge node_2 \\ node_4 &= \diamond node_3 \end{aligned}$$

This subtask is also followed by a self-validation prompt that can be found in Listing 7. Here, the LLM is required to check if the number of operator appearances in the original formula is consistent with the number of equations that contain an operator. A failure in this test also leads us to initiate this stage again from the first subtask.

---

```
Your next task has two steps, given the same expression in this language:
1. Name every leaf and non-leaf node in the tree. The name format should be
"node{index}", where an ascending index would be in the placeholder. Start
from bottom to top. The first nodes should be the leaves. If two leaves
represent the same boolean variable, give them the same name.
2. Describe every node with the operator it represents and with one or two
of its immediate descendants. For example: "node2 = node3 && node6". If the
node is a leaf node, describe it as "node{index}=q1" if its corresponding
boolean variable is q1.

The order of the statements is important: you can use a certain node as an
operand in a statement only if its associated statement was already written.
You should finish with the root of the tree.
Please use the following format:
#####
<node description>
<node description>
...
#####
```

---

**Listing 6.** A prompt that creates a list of equations out of the AST of the past time LTL formula.

---

```

Your next task has two steps:
1. Count how many statements have an operator in their description. To do that, thoroughly go over every statement and think if it has any operator inside. Please explain your solution.

2. Count how many operators appearances are there in the original formula. Remember, the different operators are: "!", "P", "H", "O", "&&", "||", "->", "<->", "S". Please explain your solution.

Finally, make sure your answer is in the following format:
- operator statements: <number>
- operator appearances: <number>

```

---

**Listing 7.** A self-validation prompt that checks consistency between the number of operators appearances in the original formula and the number of equations that contain an operator.

**Experiments.** We tested the approach of this stage on the benchmark dataset that we created, while focusing on the effectiveness of the self-validation checks of the two subtasks. For each past time LTL formula in the dataset, we generated its corresponding set of subformulas equations twice: initially omitting the self-validation checks and then including them. With the self-validation checks activated, the model was allowed up to ten attempts to correctly generate the equations. For these two variants, we report the quantity of formulas that were successfully handled, alongside those that failed during the syntax tree generation and equation generation steps. This experiment was conducted for both GPT-3.5 and GPT-4 models, with the findings detailed in Table 3.

**Table 3.** Experimental results for the syntax tree generation stage on the benchmark dataset.

Model	Approach	Success	Syntax Tree Fail.	Eq. Gen. Fail.
GPT-3.5	No Validation	11/30	11/30	8/30
	Self-Validation	23/30	3/30	4/30
GPT-4	No Validation	25/30	3/30	2/30
	Self-Validation	30/30	0/30	0/30

The table shows that self-validation of the LLM significantly improves the performance in this stage. The models produce probabilistic outputs, leading to potential inaccuracies concerning the subtasks under test. However, the two self-validation checks are designed to be simple, minimizing the chance of failure. Thus, passing these checks increases the probability that the stage’s final output is accurate. Clearly, GPT-3’s performance without self-validation is relatively poor, achieving a success rate of only 36.67%. Introducing self-validation boosts its accuracy to 76.67%. Given GPT-4’s superior capabilities, the improvement is less pronounced, yet self-validation still enhances its performance.

## 5 AST Nodes to Runtime Verification Monitor

The third stage in our prompts pipeline is responsible for generating the monitoring algorithm for the desired specification. We employed a single GPT-4 prompt to produce a monitor coded in Python. In this work we consider propositional formulas, hence we assume that the input trace to the monitor consists of Boolean variable assignments.

Our prompt in this stage is a verbal outline of the classic monitoring algorithm presented in [14]. The algorithm is based on calculating a *summary* for the current monitored trace. The summary is used, instead of storing and consulting the entire trace, for providing verdicts, and is updated when new monitored events are reported. It consists of two vectors of bits. One vector, *pre*, keeps the Boolean (truth) value for each subformula, based on the trace observed so far *except* the last observed event. The other vector, *now*, keeps the Boolean value for each subformula based on that trace *including* the last event. Given a new event  $e$  consisting of a set of propositions, which extends the monitored trace, the vector *now* is calculated based on the vector *pre* and the event  $e$ . The symbol  $q$  denotes a single Boolean proposition. This is summarized below:

- $\text{now}(\text{true}) = \text{True}$
- $\text{now}(q) = (q \in e)$
- $\text{now}(\neg\varphi) = \neg\text{now}(\varphi)$
- $\text{now}((\varphi \text{ op } \psi)) = (\text{now}(\varphi) \text{ op } \text{now}(\psi))$
- $\text{now}((\varphi \mathcal{S} \psi)) = (\text{now}(\psi) \vee (\text{now}(\varphi) \wedge \text{pre}((\varphi \mathcal{S} \psi))))$ .
- $\text{now}(\ominus \varphi) = \text{pre}(\varphi)$
- $\text{now}(\Leftrightarrow \varphi) = (\text{now}(\varphi) \vee \text{pre}(\Leftrightarrow \varphi))$
- $\text{now}(\boxplus \varphi) = (\text{now}(\varphi) \wedge \text{pre}(\boxplus \varphi))$

When a new event appears, *now* becomes *pre*, and the *now* values are calculated according to the above cases.

The monitor is generated based on the prompt described in Listing 8 (the description of the classic monitoring algorithm), and on the equations that were generated in the previous stage, where each equation corresponds to a different subformula. Each node's value from the preceding stage corresponds to a single bit in the summary of the monitored trace.

Longer prompts that involve code generation with many requirements are known to be less reliable - as the prompt becomes longer, the likelihood of the code meeting all requirements decreases. Hence, we employed here an *adaptive prompting* technique. Most of the prompt is dedicated to the update rules for the *now* bit vector, although not every rule is needed in every case (as some operators might not be used in every formula). To reduce prompt length, we selectively include only the update rules that are relevant to the formula being addressed.

In this stage, we explicitly outlined the different update rules of each one of the relevant operators. We also began exploring alternative methods to apply these rules without directly mentioning them in the prompt, instead deriving

them directly using the LLM. This early research direction is further described in Appendix B.

---

```

Your final task is to generate a Python program that receives at each step
an assignment for all the Boolean variables in the above Past-time linear
temporal logic expression, and then prints an output. This assignment of the
Boolean variables is referred to as an event. At each step, the event would
be inserted by the user. The output will be based on a certain analysis of
the event. The program would then be ready to receive another event from the
user and analyze it in a similar manner, and finally print another output,
and so forth. The program would terminate when the user inserted the word
"abort".

To perform the analysis, we would keep in memory two versions of the values
of each node from the previous task. The first version would be the current
version while the other one would be the previous version. Before the first
step, the current version of all nodes can be assumed to be a "False" value,
except of the current version of an "H" node that should be initialized to
"True".

The current version of each node will be updated based on its type and on
its previous version. The update order of the nodes should be the same as in
the previous task - the current version of the node may be updated only
after its operands are updated.
- A leaf node would be the truth value of the corresponding variable
according to the new event.
- The current version of node of type "!" would be the negation of the
current version of its single operand.
Similarly, we specify all the update rules that are relevant for the input
formula being addressed.
.....
.....
The previous version of all nodes will be updated only after the current
version of all nodes is updated. Finally, the current version of the root
node should be printed.
- When asking the user for an event, you must not print anything to the
user, just expect an input of the format "q1=<boolean>,q2=<boolean>,....",
where "boolean" is either "True" or "False".
- When printing the current version of the root node, print just a single
word that describes its boolean value (True or False).

```

---

**Listing 8.** A prompt that generates propositional past time LTL monitor, based on the subformulas equations and the monitoring algorithm description.

## 6 Self-validation of the Monitor Code

### 6.1 Method

Code generated by LLMs based on prompts may be inaccurate on both syntactic and semantic levels. As part of our efforts to synthesize a runtime verification monitor using solely the capabilities of LLMs, we explored methods to validate the generated monitor. While we understand that a validation process started by LLMs might not be entirely accurate, our goal was to increase the likelihood of producing a reliable monitor. We aimed for two validation types: (1) syntactic validation of the monitor code, and (2) logical/semantic sanity check of the code.

We used GPT-4 to perform the self-validation stage. First, we queried it to produce two traces of events with respect to the original natural language specification (serving as a semantic sanity check), where one trace would satisfy



the specification and the other would fail to do so. Then, we employed the code interpreter variant of GPT-4 to evaluate the monitor on the synthesized traces, and validate the output verdicts of the monitor (syntactic validation). We will now provide further details on these two steps.

**Semantic Sanity Check with Trace Generation.** We noticed that GPT-4 is able to generate short traces that are consistent with past time LTL formulas and their natural language interpretation. We did some preliminary experiments in which we prompted GPT-4 with natural language specifications that correspond to past time LTL formulas, and asked it to output a short trace (up to ten events) that satisfies the specification. Each event was defined as an assignment to all the Boolean variables in the intermediate form sentence that was given as input. We then repeated this process but this time we provided the model with past time LTL formulas instead of their intermediate natural language representation. We observed that GPT-4 performs better when its input is a natural language sentence rather than a structured formula. We conclude that GPT-4 better captures temporal dependencies when the specification is provided in natural language.

Following this observation, we tested the trace generation ability on the benchmark dataset from Sect. 3.2. For each natural language specification in its intermediate form, we instructed GPT-4 to produce three short traces (up to ten events) that satisfy it, and three additional traces that do not. The prompt we provided for this task listed all the possible temporal operators of past time LTL, followed by their meaning. In addition, we provided a single-shot example triplet of a specification, a short satisfying trace and an unsatisfying trace, where each event was defined as an assignment to the different Boolean variables in the specification.

GPT-4 managed to output different traces for each specification, except two cases where there was a single satisfying trace. Among 90 traces designed to satisfy the specifications, 78 indeed satisfied their corresponding intermediate form sentences, marking an accuracy rate of 86.67%. Remarkably, it succeeded in producing accurate traces for the two samples that were not translated correctly in the first stage. For these samples, 5 out of 6 generated traces were accurate. Of the 90 traces that were not supposed to satisfy their specifications, 85 indeed did not, yielding an accuracy rate of 94.44%. The higher accuracy here may be attributed to the fact that, within our benchmark dataset, violating a few specifications tends to be slightly simpler than fulfilling them.

**Syntactic Validation with Code-Interpreter.** Recently, OpenAI has unveiled an improved version of GPT-4, namely the GPT-4 Code Interpreter<sup>2</sup>. This version is capable of offering reasoning in natural language while producing Python code in a step-by-step manner. Notably, it can generate and execute code incrementally, then relay the output of the executed code back to the LLM.

<sup>2</sup> <https://platform.openai.com/docs/assistants/tools/code-interpreter>.

This variant of GPT-4 was utilized not only for syntactical verification of the generated code but also for executing it on two traces, one that satisfied the specification and one that did not, which were also produced with the help of the LLM. The prompt we provided had three requirements:

- Check that the monitor executes correctly, and that it produces a Boolean output for each event.
- Verify that the last verdict of the monitor for the satisfying trace is “True”.
- Verify that the last verdict of the monitor for the unsatisfying trace is “False”.

A failure in any of these checks requires us to go back to the code generation step and generate a new monitor. The prompt for this step is presented in Listing 9, and an example of a response is shown in Listing 10.

---

```
Given a first trace of events: {insert_satisfying_generated_trace}
and a second trace of events: {insert_unsatisfying_generated_trace}

Read the contents of the file "monitor.py", and modify it so instead of
getting inputs from the user, each time an event will be loaded from the
provided traces. When the first trace ends, rerun the code with the second
one. Make sure that each event is given as input in the format: "q1=True,
q2=False, ..".

Finally, let me know what were the outputs of the monitor for the two
traces. Summarize the process with three outputs:
1. Whether or not the monitor executed properly and produced a boolean
output for each event. (just write success/failure)
2. Write the last output that the monitor produced for the first trace.
3. Write the last output that the monitor produced for the second trace.

Write the summary in the following format:
execution status: (Success/Failure)
first trace last output: (True/False)
second trace last output: (True/False)
```

---

**Listing 9.** A prompt that uses the Code Interpreter variant of GPT to test the generated monitor on two traces. The traces to be tested should be provided inside the curly brackets, and the monitor itself is passed to the model as a separate file.

---

```
To address your request, I'll follow these steps:

1. Modify the code from "monitor.py" to process the provided traces instead
of reading from user input.
2. Load and run the first trace through the modified code.
3. Reload the code to start fresh and then run the second trace.
4. Summarize the results according to the format you provided.
First, I'll modify the code from "monitor.py" based on the description.

Let's start by writing the modified code to properly handle and iterate
through the two given traces.

.....
.....
<modified code displayed>
.....
.....
```

---

```
The modification of the code from "monitor.py" to process the provided
traces instead of getting inputs from a user was successful. Here's the
summary based on the results obtained:
```

```
execution status: Success
first trace last output: True
second trace last output: False
These outcomes indicate that the monitor executed properly for each event in
both traces and the specified final Boolean outputs for the last events were
True for the first trace and False for the second trace.
```

**Listing 10.** The response of the Code Interpreter variant after testing a generated monitor of two given traces.

## 6.2 Experiments

We conducted end-to-end experiments, applying all four stages of our proposed approach to synthesize runtime verification monitors from natural language specifications. The main goal was to evaluate the accuracy of the monitors produced, with an additional objective to test the effectiveness of the monitor self-validation technique described earlier.

Our experiments involved fifteen natural language specifications randomly selected from the benchmark dataset. In the first stage of the process, we used the *GPT-4-few-shot* configuration of the prompting based approach to translate from intermediate form specification into past time LTL. In the second stage, where the formula was translated into an abstract syntax tree representation, we allowed up to three repetitions in case of subtask failures. Subsequently, we permitted up to three attempts at generating monitor code (in the third stage) if there was a failure in the monitor self-validation stage (fourth stage). Once the process terminated, each monitor created was evaluated against a real runtime verification tool, regardless of the self-validation outcome. We based our comparison on 1000 distinct traces, randomly generated with 30 events each, where the comparative tool was supplied with the past time LTL formula derived from the respective natural language specification in the benchmark dataset.

**Table 4.** Accuracies (%) of the end-to-end experiments.

Experiment	Correctness of Generated Monitor	Self-Validation Test Consistency
Accuracy	94.34	73.34

Table 4 indicates that our four-stage process succeeded in generating accurate monitors for 94.34% of the specifications tested. Focusing on the self-validation test results, it was observed that these results were consistent with the actual correctness of the monitor in 73.34% of the cases (11 out of 15). Moreover, all discrepancies involved false negatives, where the self-validation stage incorrectly reported a correct monitor as faulty.

It is important to note that our work is still in progress, and the limited sample size in these experiments makes it challenging to draw definitive conclusions from the performance presented above. Nevertheless, these preliminary results are encouraging and represent a positive progression in the effort to teach LLMs the algorithmic foundations of runtime verification approaches.

## 7 Conclusion

In this work, we attempted to teach LLMs to synthesize code for runtime verification from natural language specification. While the LLM abilities were vital for the translation stage from natural language, we identified further opportunities to leverage natural language interfacing across additional stages of the synthesis process. Specifically, we employed a number of self-validation prompts that improved the creation of the syntax tree for a specific formula. Additionally, we observed that trace generation achieved greater accuracy when it is derived from the natural language specification, as opposed to the formula itself. This suggests that depending on the task domain, integrating language comprehension abilities can be beneficial at various points in the process.

Furthermore, at the current stage of LLM evolution, their inherent knowledge appears adequate for converting relatively complex natural language sentences into temporal logics, including less common variants such as past time LTL. For our purposes, drawing parallels between LTL and past time LTL, coupled with a few illustrative examples, sufficed to handle a variety of unstructured natural specifications. However, it is important to acknowledge that LLMs understanding of temporal dependencies is not flawless, and their translation efficiency declines as sentences increase in length and complexity.

For a future research, there is still work to be done with respect to teaching LLMs the algorithmic foundations of classic runtime verification approaches. In our work, the directives provided to the model in the stage of the runtime verification monitor creation were relatively straightforward. We anticipate that as LLMs continue to evolve, they will feature more advanced deduction capabilities, leading to a more profound comprehension of these algorithmic principles. Such advancements would, ideally, pave the way for instructing LLMs on comprehensive runtime verification concepts through more autonomous and concise explanations. We have already made a preliminary exploration in this direction by effectively extracting a few specific update rules for the RV summary using GPT-4. Further details on this are provided in Appendix B.

## A GPT’s Zero-Shot Understanding of Past Time LTL

When GPT-4 was instructed to list the temporal operators of past time LTL and concisely define their semantics, it proposed some operators that do not semantically align with the standard associated past time LTL operators. Specifically, the semantics of the proposed operator “until in the past” was not consistent with the conventional semantics of “since”, which is the past dual of the future

time “until” modality. In addition, it appeared not to recognize the “previous-time” operator. A prompt and a response that describe this phenomenon are shown in Listing 11.

---

**User:**  
List all the temporal operators of past time linear temporal logic, and shortly describe their semantics.

**ChatGPT:**  
In linear temporal logic (LTL), temporal operators are used to reason about sequences of events over time. Here are the primary temporal operators for past time in LTL:

1. Globally in the past ( $P$ ): Denoted as  $P\varphi$ , this operator means that the formula  $\varphi$  holds true for all past time points.  
- Semantics:  $\varphi$  holds true at every point in the past.
2. Eventually in the past ( $P\Diamond$ ): Denoted as  $P\Diamond\varphi$ , this operator means that  $\varphi$  holds true at some point in the past.  
- Semantics: There exists a past time point at which  $\varphi$  holds true.
3. Until in the past ( $PU$ ): Denoted as  $P\varphi U\psi$ , this operator means that  $\varphi$  holds true until  $\psi$  becomes true in the past.  
- Semantics: For every point in the past where  $\psi$  holds true,  $\varphi$  holds true at that point and all preceding points until  $\psi$  becomes true.
4. Release in the past ( $PR$ ): Denoted as  $P\varphi R\psi$ , this operator means that  $\varphi$  holds true until  $\psi$  becomes true, and once  $\psi$  becomes true,  $\varphi$  remains true for all subsequent points in the past.  
- Semantics: If  $\psi$  holds true at any point in the past, then  $\varphi$  has to hold true at that point and all preceding points until  $\psi$  becomes true, and  $\varphi$  continues to hold true for all past points thereafter.

These operators are used to reason about the past-time behavior of systems and specifications in temporal logic.

---

Listing 11. GPT’s zero-shot understanding of past time LTL.

## B Extracting RV Summary Update Rules from GPT

We explored the possibility of extracting from GPT-4 the necessary update rules that help the classic RV algorithm update its summary, with only minimal information provided. We created a prompt that mentions the two Boolean vectors *now* and *pre*, the fact that each bit in the vectors is responsible for a different subformula, and the temporal operators “historically” and “since”. Then, we requested it to formulate the update rules for these operators. GPT-4 was able to retrieve the correct update rules for these operators, as shown in Listing 12. Note that the interaction detailed here represents the complete interaction with the model, with no preliminary knowledge provided.

---

**User:**  
Your task is to build a runtime verification monitor that monitors specifications written in propositional past time LTL. To track the boolean values of each subformula along the execution, you have two Boolean vectors

- "now" and "pre" vectors. Assume that one bit of the "now" vector is responsible for the formula "H(q)", where "H" is the "always in the past" operator in past time LTL, and q is a subformula. how would you calculate this bit, based on the other bits of the "now" vector and the bits of the "pre" vector?

ChatGPT:

To build a runtime verification monitor that monitors specifications written in propositional past time Linear Temporal Logic (LTL), you have to manage the logical evaluation of LTL formulas over the traces of system execution. The "H(q)" formula, where "H" denotes "always in the past" (historically), requires that for the property "q" to be true at the current point in time, it must have been true at all previous points in time during the execution.

Given two Boolean vectors, "now" and "pre", which track the truth values of LTL subformulas at the current and previous points in time, respectively, calculating the bit for "H(q)" involves the following steps:

1. Identify the Bit for q: First, identify the bit in both the "now" and "pre" vectors that corresponds to the subformula "q". Let's denote these bits as "now[q]" and "pre[q]", respectively.

2. Calculate H(q) for the Current Time: The truth value of "H(q)" at the current time step depends on:

- The truth value of "q" at the current time step (i.e., "now[q]").
- The historical truth value of "H(q)" up to the previous time step (i.e., "pre[H(q)]").

3. Update Rule for H(q): The bit for "H(q)" in the "now" vector (denoted as "now[H(q)]") can be calculated as follows:

- "now[H(q)]" = "now[q]" AND "pre[H(q)]"

This calculation is based on the semantic rule of the "H" operator, which states that for "H(q)" to be true at the current moment, "q" must be true now (i.e., "now[q]" must be true), and "H(q)" must have been true at the previous moment (i.e., "pre[H(q)]" must be true). This reflects the property that "q" must have always been true in the past for "H(q)" to hold.

User:

Now, how would you calculate the bit that corresponds to the formula "q1 Since q2"? Since q2"?

ChatGPT:

The "Since" operator in propositional past time Linear Temporal Logic (LTL) captures a form of historical dependency between two conditions. Specifically, the formula "q1 S q2" means "q1 has been true since q2 was last true." In other words, for "q1 S q2" to be true at a certain point, "q2" must have been true at some previous time, and "q1" must have been true continuously from that point up to the current time.

Given two Boolean vectors, "now" and "pre", which represent the truth values of subformulas at the current and previous time steps, respectively, calculating the bit for "q1 S q2" involves considering both the current and historical truth values of "q1" and "q2". Let's break down how to calculate this bit:

1. Identify the Bits for q1 and q2: First, identify the bits in the "now" and "pre" vectors corresponding to the subformulas "q1" and "q2". Let's denote these bits as "now[q1]", "pre[q1]", "now[q2]", and "pre[q2]", respectively.

2. Calculate "q1 S q2" for the Current Time: The truth value of "q1 S q2" at the current time step depends on the current truth values of "q1" and "q2", as well as the historical truth value of "q1 S q2". Specifically:

- If "q2" is true now ("now[q2]" = true), then "q1 S q2" becomes true regardless of its past value because "q2" occurring now satisfies the "since" condition.

```
- If "q2" is not true now ("now[q2]" = false), then for "q1 S q2" to be true, both "q1" must be true now ("now[q1]" = true) and "q1 S q2" must have been true in the previous step ("pre[q1 S q2]" = true).
```

```
3. Update Rule for "q1 S q2": Based on the above, the bit for "q1 S q2" in the "now" vector (denoted as "now[q1 S q2]") can be calculated as follows:
- "now[q1 S q2]" = "now[q2]" OR ("now[q1]" AND "pre[q1 S q2]")
```

This calculation reflects the semantics of the "Since" operator: "q1 S q2" is true if either "q2" is true now (capturing the idea that "q2" initiates the truth of "q1 S q2") or if "q1" has been true since the last time "q2" was true (reflected by "now[q1]" being true and "q1 S q2" also being true in the previous step).

**Listing 12.** A prompt that extracts update rules of past time LTL temporal operators.

## References

1. Alpern, B., Schneider, F.B.: Recognizing safety and liveness. *Distrib. Comput.* **2**(3), 117–126 (1987)
2. Anil, R., et al.: Gemini: a family of highly capable multimodal models. *CoRR abs/2312.11805* (2023)
3. Barringer, H., Rydeheard, D.E., Havelund, K.: Rule systems for run-time monitoring: from eagle to ruler. *J. Log. Comput.* **20**(3), 675–706 (2010)
4. Bartocci, E., Falcone, Y., Francalanza, A., Reger, G.: Introduction to runtime verification. In: Bartocci, E., Falcone, Y. (eds.) *Lectures on Runtime Verification*. LNCS, vol. 10457, pp. 1–33. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-75632-5\\_1](https://doi.org/10.1007/978-3-319-75632-5_1)
5. Basin, D.A., Klaedtke, F., Müller, S., Zalinescu, E.: Monitoring metric first-order temporal properties. *J. ACM* **62**(2), 15:1–15:45 (2015). <https://doi.org/10.1145/2699444>
6. Brown, T.B., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 6–12 December 2020, virtual* (2020)
7. Brunello, A., Montanari, A., Reynolds, M.: Synthesis of LTL formulas from natural language texts: State of the art and research directions. In: Gamper, J., Pinchinat, S., Sciavicco, G. (eds.) *26th International Symposium on Temporal Representation and Reasoning, TIME 2019, 16–19 October 2019, Málaga, Spain*. LIPICs, vol. 147, pp. 17:1–17:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2019)
8. Clarke, E.M., Grumberg, O., Peled, D.A.: *Model Checking*, 1st Edition. MIT Press, Cambridge (2001). <http://books.google.de/books?id=Nmc4wEaLXFEC>
9. Cosler, M., Hahn, C., Mendoza, D., Schmitt, F., Trippel, C.: nl2spec: interactively translating unstructured natural language to temporal logics with large language models. In: Enea, C., Lal, A. (eds.) *Computer Aided Verification - 35th International Conference, CAV 2023, Paris, France, July 17–22, 2023, Proceedings, Part II*, LNCS, vol. 13965, pp. 383–396. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-37703-7\\_18](https://doi.org/10.1007/978-3-031-37703-7_18)
10. D'Angelo, B., et al.: LOLA: runtime monitoring of synchronous systems. In: *12th International Symposium on Temporal Representation and Reasoning (TIME 2005), 23–25 June 2005, Burlington, Vermont, USA*, pp. 166–174. IEEE Computer Society (2005)

11. Fuggitti, F., Chakraborti, T.: NL2LTL - a python package for converting natural language (NL) instructions to linear temporal logic (LTL) formulas. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, 7–14 February 2023, pp. 16428–16430. AAAI Press (2023)
12. Hahn, C., Schmitt, F., Tillman, J.J., Metzger, N., Siber, J., Finkbeiner, B.: Formal specifications from natural language. CoRR **abs/2206.01962** (2022)
13. Havelund, K., Peled, D., Ulus, D.: First-order temporal logic monitoring with BDDs. Formal Methods Syst. Des. **56**(1), 1–21 (2020). <https://doi.org/10.1007/S10703-018-00327-4>
14. Havelund, K., Roşu, G.: Synthesizing monitors for safety properties. In: Katoen, J.-P., Stevens, P. (eds.) TACAS 2002. LNCS, vol. 2280, pp. 342–356. Springer, Heidelberg (2002). [https://doi.org/10.1007/3-540-46002-0\\_24](https://doi.org/10.1007/3-540-46002-0_24)
15. Jiang, A.Q., et al.: Mistral 7b. CoRR **abs/2310.06825** (2023)
16. Leucker, M., Schallhart, C.: A brief account of runtime verification. J. Log. Algebraic Methods Program. **78**(5), 293–303 (2009)
17. Liu, J.X., Yang, Z., Idrees, I., Liang, S., Schornstein, B., Tellex, S., Shah, A.: Grounding complex natural language commands for temporal tasks in unseen environments. In: Tan, J., Toussaint, M., Darvish, K. (eds.) Conference on Robot Learning, CoRL 2023, 6–9 November 2023, Atlanta, GA, USA. Proceedings of Machine Learning Research, vol. 229, pp. 1084–1110. PMLR (2023)
18. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. OpenAI (2018)
19. Tan, C.W., Guo, S., Wong, M., Hang, C.N.: Copilot for xcode: exploring AI-assisted programming by prompting cloud-based large language models. CoRR **abs/2307.14349** (2023)
20. Touvron, H., et al.: Llama: open and efficient foundation language models. CoRR **abs/2302.13971** (2023)
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, pp. 5998–6008 (2017)
22. Wang, C., Ross, C., Kuo, Y., Katz, B., Barbu, A.: Learning a natural-language to LTL executable semantic parser for grounded robotics. In: Kober, J., Ramos, F., Tomlin, C.J. (eds.) 4th Conference on Robot Learning, CoRL 2020, 16–18 November 2020, Virtual Event/Cambridge, MA, USA. Proceedings of Machine Learning Research, vol. 155, pp. 1706–1718. PMLR (2020)



**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





# AI-Assisted Programming with Test-Based Refinement

Bernhard K. Aichernig<sup>1</sup> and Klaus Havelund<sup>2</sup>(✉)

<sup>1</sup> Institute of Software Technology, Graz University of Technology, Graz, Austria

<sup>2</sup> Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA  
havelund@gmail.com

**Abstract.** This work explores the utilization of a Large Language Model (LLM), specifically OpenAI’s ChatGPT, to develop a program as a sequence of refinements. Traditionally in formal methods literature such refinements are proven correct, which can be time consuming. In this work the refinements are tested using property-based testing. This approach addresses the problem of ensuring that the code generated by an LLM is correct, which is one of the main challenges of code generation with LLMs. Programs are developed in Scala and testing is performed with ScalaCheck. This approach is demonstrated through the development and testing of a classical bridge controller, originally presented in documentation for the refinement-based Event-B theorem prover.

## 1 Introduction

Neural program synthesis, based on Large Language Models (LLMs) which are trained on open source code, are quickly becoming a popular addition to the software developer’s toolbox. Services like, for instance, Open AI’s ChatGPT<sup>1</sup>, Google’s Gemini<sup>2</sup> (the next generation of Bard), and GitHub’s Copilot<sup>3</sup>, can generate code in many different programming languages from natural language requirements entered as “prompts”. A system like ChatGPT is for example particular good at answering questions like “*how do you do X in programming language Y?*”.

Prompt-based programming, however, seems to work best for development of smaller programs. It currently appears infeasible to generate large and complex programs from natural language prompts. It is also not clear how good such a

---

<sup>1</sup> <https://chat.openai.com>.

<sup>2</sup> <https://gemini.google.com>.

<sup>3</sup> <https://github.com/features/copilot>.

© 2024. All rights reserved

The research performed by this author was carried out at Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 385–411, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_24](https://doi.org/10.1007/978-3-031-73741-1_24)

system is for generating code that strays away from common patterns. We refer to this as the *complexity problem*. Furthermore, neural systems do not come with guarantees of producing correct, safe, or secure code, what we shall refer to as the *verification problem*.

We propose *test-based refinement* to address these two problems. Program refinement is an old idea, but usually performed with *proofs* of correctness [1, 2, 6, 13, 16]. We experiment with this concept but using *testing*. Specifically, to approach the *complexity* problem, we propose a method based on program refinement, where a program is developed in a step-wise manner, starting with a very high-level abstract program, and then refining it iteratively, towards a final implementation. At each refinement step, the LLM can provide assistance by generating code suggestions and refining existing code snippets. To approach the program *verification* problem, we suggest to apply automated *testing*, including test-case generation, to test that each refinement implements the previous step. Literature on program refinement usually approaches the verification problem as a deductive proof problem. Proofs are, however, hard to carry out for humans, even with automated proof tools.

This work is in line with the TriCo (Triple Co-Piloting) principle to software development described in [3], which argues that implementation, (formal) specification, and tests should be developed hand-in-hand, assisted by machine learning, such as e.g. an LLM. We do, however, not address the formal specification part of TriCo. In general, there has been published numerous articles on the use of LLMs for software development. The paper [7] provides an overview of some of these, in addition to reflecting on how LLMs can be used for software development and quality assurance. Closest to our work is the contribution in [9]. The authors first instantiate computer adventure games with the help of ChatGPT from natural language requirements. Then, they mine finite-state models from the code and check their correctness with model checkers. A yet different approach is to verify neural networks themselves, as discussed in [8].

Our experiments are carried out by developing a classical bridge controller, originally introduced as part of the tutorial for the Event-B theorem prover [2], where a proof-based refinement is performed. Our programs are written in the Scala<sup>4</sup> programming language. Testing is performed using the ScalaCheck<sup>5</sup> property-based testing library. Property-based testing is a method for defining general properties the code should satisfy, quantifying over the input domains (similar to universal quantification in mathematics), and the ScalaCheck library automatically generates test data to verify those properties across a wide range of inputs. Property-based testing is also mentioned in [3] as a way of ensuring correctness in the TriCo approach. We use property-based testing in two ways. First we use property-based testing as a way to test each single program, following the original intent of ScalaCheck. Second, we use property-based testing to test the refinement relation between two versions of a program.

---

<sup>4</sup> <https://www.scala-lang.org>.

<sup>5</sup> <https://scalacheck.org>.

We demonstrate the approach with two different refinements of the bridge controller. The first refinement follows the Event-B proof of the bridge controller, which describes a system as a set of events, each consisting of a precondition and an action having side effects on a shared state. Conceptually one outer loop repeatedly picks an event where the precondition is true and executes the action, until no event preconditions are true. This concept has its roots in Dijkstra’s language of guarded commands [11], which is Turing complete. The second refinement follows a more classical object-oriented approach, the way many would normally write programs in a language like Scala.

The objective of this work is to (i) explore writing a program iteratively in steps using program refinement, and test the refinements; (ii) augment refinement tests with property-based testing; and (iii) use ChatGPT to generate code, refinement tests and property-based tests. All code should be (and was largely) generated by ChatGPT. Note that we are using ChatGPT’s web-based interface Version 4 (and in some cases Version 3 when the upper limit per day for use of Version 4 was reached). We did not use any APIs to the LLM. We also did not use Copilot or any such IDE.

The paper is structured as follows. Section 2 introduces the background for this work, including the well established theory of refinement, in particular its proof theoretic focus, as well as property-based testing. Section 3 explains how one can test refinements instead of proving them correct. Section 4 introduces the bridge controller example. Section 5 presents the guarded command approach to test-based refinement in Scala, while Sect. 6 presents the object-oriented approach. Finally, Sect. 7 concludes the paper with observations.

## 2 Background

In this section we introduce the background for the work, including the established notions of *refinement* and *property-based testing*.

### 2.1 Refinement

Refinement is the idea of developing a system as a series of specifications (which can be programs), each refining the previous specification, except the first, which constitutes the top-level specification of the system. An implementation at the bottom is then defined as a refinement of the top level abstract specification through transitivity of the refinement relation. A system can be a physical system, a software system, or a mix of the two.

Traditionally in formal methods literature, refinements are proved correct using refinement mappings. In this section we formalize this concept. The theory presented is a minor modification of the theory developed by Abadi and Lamport [1] with the addition of observation functions and invariants as described in [12] (the definitions below are from [12] with a few modifications). We introduce the concepts of *transition systems*, *execution traces*, *invariants*, *refinements*, and *refinement mappings*. Specifications are written as transition systems. A *transition system* is defined as follows.

**Definition 1 (Transition System).** A transition system is a five-tuple  $(\Sigma, I, N, \Sigma_o, \pi)$  where

- $\Sigma$  is a state space.
- $I \subseteq \Sigma$  is the set of initial states.
- $N \subseteq \Sigma \times \Sigma$  is the next-state relation. Elements of  $N$  are denoted by pairs of the form  $(s, t)$ , meaning that there is a transition from the state  $s$  to the state  $t$ .
- $\Sigma_o$  is the state space of observations.
- $\pi : \Sigma \rightarrow \Sigma_o$  is an observation function.

The observation function  $\pi$  in the above definition, which when applied to a state in  $\Sigma$  returns an observation in  $\Sigma_o$ , allows us to compare traces from two transition systems using a shared state space  $\Sigma_o$  in the case where the internal states  $\Sigma$  of the two systems differ.

An *execution trace* is an infinite sequence  $\sigma = \langle s_0, s_1, s_2, \dots \rangle$  of states<sup>6</sup>, where the first state satisfies the  $I$  predicate and every pair of adjacent states is related by the  $N$  relation. We let  $\sigma_i$  denote the  $i$ 'th element  $s_i$  of the sequence. The traces of a transition system can be defined as follows.

**Definition 2 (Traces).** The traces of a transition system are defined as follows:

$$\Theta(\Sigma, I, N, \Sigma_o, \pi) = \{\sigma \in \Sigma^\omega \mid \sigma_0 \in I \wedge \forall i \geq 0 \cdot N(\sigma_i, \sigma_{i+1})\}$$

A projection function  $\pi$  applied to a trace  $\langle s_1, s_2, \dots \rangle$  results in the projected trace  $\langle \pi(s_1), \pi(s_2), \dots \rangle$ .

An *invariant* is a state predicate true on all states reachable from an initial state via the next-state relation.

**Definition 3 (Invariant).** Given a transition system  $S = (\Sigma, I, N, \Sigma_o, \pi)$ , then a predicate  $P : \Sigma \rightarrow \mathcal{B}$  is an  $S$  invariant iff

$$\forall \sigma \in \Theta(S) \cdot \forall i \geq 0 \cdot P(\sigma_i)$$

We can now define the concept of *refinement*.

**Definition 4 (Refinement).** A transition system  $S_2 = (\Sigma_2, I_2, N_2, \Sigma_o, \pi_2)$  refines a transition system  $S_1 = (\Sigma_1, I_1, N_1, \Sigma_o, \pi_1)$  iff for every trace of  $S_2$  there exists a trace of  $S_1$  with the same observed states:

$$\forall \sigma_2 \in \Theta(S_2) \cdot \exists \sigma_1 \in \Theta(S_1) \cdot \pi_1(\sigma_1) = \pi_2(\sigma_2)$$

---

<sup>6</sup> A finite execution can be thought of as being represented by an infinite trace with the last state of the finite execution being repeated infinitely.

Note that it is here assumed that the abstract transition system  $S_1$  can perform *stuttering* (no-op) steps where the state does not change, reflecting that the concrete system  $S_2$  makes an internal move not corresponding to a state changing move in  $S_1$ .

This definition of refinement in terms of infinite traces, however, is not useful for practical proofs or tests. For practical purposes, the notion of a *refinement mapping* from a lower level transition system  $S_2$  to a higher-level one  $S_1$  is introduced, allowing to reason about pairs of states.

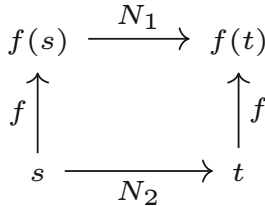
**Definition 5 (Refinement Mapping).** *A refinement mapping from a transition system  $S_2 = (\Sigma_2, I_2, N_2, \Sigma_o, \pi_2)$  to a transition system  $S_1 = (\Sigma_1, I_1, N_1, \Sigma_o, \pi_1)$  is a mapping  $f : \Sigma_2 \rightarrow \Sigma_1$  such that there exists an  $S_2$  invariant  $P$  (representing reachable states in  $S_2$ ), where:*

1.  $\forall s \in \Sigma_2 \cdot \pi_1(f(s)) = \pi_2(s)$
2.  $\forall s \in \Sigma_2 \cdot I_2(s) \Rightarrow I_1(f(s))$
3.  $\forall s, t \in \Sigma_2 \cdot P(s) \wedge N_2(s, t) \Rightarrow N_1(f(s), f(t))$

The main property is 3, expressing that a move on the concrete transition system  $S_2$  simulates a corresponding move in the abstract transition system  $S_1$ , see Fig. 1. We can now state the main theorem (which is stated in [1]):

**Theorem 1 (Existence of Refinement Mappings).** *If there exists a refinement mapping from a transition system  $S_2$  to a transition system  $S_1$ , then  $S_2$  refines  $S_1$ .*

Defining the refinement mapping turns out typically to be easy, whereas proving that it is indeed a refinement mapping (Property 3 in Definition 5) can be hard, in particular discovering a sufficiently strong invariant. As we shall see in Sect. 3 we will test this relationship instead of proving it.



**Fig. 1.** Commuting refinement diagram

## 2.2 Property-Based Testing

Property-based testing (PBT) is a testing technique that tries to falsify a given property by generating random input data and verifying the expected behaviour. The motivation is to replace hand-written unit tests with property specifications that are tested automatically. The first tool implementing this style of testing

was QuickCheck [10]. Properties can range from simple algebraic equations to complex state machine models. Like in all model-based testing techniques the properties serve as a source for test-case generation as well as test oracles. It is a well-known testing practice in functional programming, but nowadays we see a growth of applications outside its traditional domain, including automotive software [5] and web-services [4, 15].

A simple example of an algebraic property is that the reverse of the reverse of a list must equal the original list:

$$\forall xs \in List[T] : reverse(reverse(xs)) = xs$$

A PBT tool will generate a series of random lists  $xs$ , execute the reverse function and evaluate the property. This test-case generation is realized via the composition of test-case generators. For the example above, the test-case generator for lists will use a test-case generator for type  $T$ . In addition to default generators for basic types, a user can define custom generators and mix them arbitrarily.

If a property is violated, a counter-example will be presented to the user. For easier debugging, the counter-example is simplified in a process called *shrinking*.

The original QuickCheck tool was implemented in, and supported testing of, Haskell programs, but has been ported to many programming languages including FsCheck<sup>7</sup> for C#, Hypothesis<sup>8</sup> for Python, and ScalaCheck<sup>9</sup> for Scala. In this work we rely on the latter. In ScalaCheck the above property for integer lists would be represented as follows:

```
1 property("reverse") = forall { (xs: List[Int]) => reverse(reverse(xs)) =
  = xs }
```

ScalaCheck will by default generate 100 random integer lists in order to test this property.

### 3 Refinement and Testing

In this section, we present our AI-assisted method for developing programs with LLMs. We will use step-wise refinement in order to develop a program iteratively from an abstract method down to a concrete implementation. Hence, the LLM shall assist us in generating improved versions of an abstract program that preserve the refinement mapping. In each refinement step, additional requirements are considered while preserving correctness with respect to the more abstract versions. Traditionally, the refinement relation is formally verified via model checking, like in TLA [14], or via theorem proving, like in Event-B [2].

In our method, we propose a more lightweight approach, namely the automated testing of the refinement relation with the help of property-based testing. More concretely, we use a refinement property and data generators to test if two objects respect refinement.

<sup>7</sup> <https://fscheck.github.io/FsCheck/>.

<sup>8</sup> <https://github.com/HypothesisWorks/hypothesis>.

<sup>9</sup> <https://scalacheck.org>.

```

1  property("Class C2 refines C1 for method m()") =
2  forall(validStates) { object2: Class2 =>
3    val object1 = refinement_mapping(object2)
4    if (object1.guards && object2.guards) {
5      object1.m()
6      object2.m()
7      object1 == refinement_mapping(object2)
8    }
9    else true
10 }

```

**Fig. 2.** Sketch of a ScalaCheck property for testing that a method of Class1 correctly refines its abstract version in Class0.

For the testing of refinement, we can simplify the definition of refinement in Definition 5. Assuming executable Scala methods to be tested, we can specialize the general next-relation  $N$  to a next-function resulting in a simpler version of Def. 5.3:

$$\forall s \in \Sigma_2 \cdot P(s) \Rightarrow N_1(f(s)) = f(N_2(s))$$

Note that this property directly corresponds to the commuting diagram in Fig. 1. Hence, when testing if a method  $N_2^m$  refines an abstract method  $N_1^m$ , we simply initialize both objects to a state such that the refinement mapping  $f$  and the invariants hold, then execute both mappings and check if the results are equal under the mapping  $f$ .

Figure 2 shows how such a refinement property can be realized in ScalaCheck. The property checks for all states of the concrete class `Class2` that a method `m` respects refinement. A custom generator `validStates` randomly generates objects of `Class2` that also respect its invariant (Line 2). Hence, the invariant check is part of this custom generator. Then, the refinement mapping is used to construct or initialize the abstract version `Class1` (Line 3). If the guards of the methods are satisfied, both methods are executed and the equivalence of the resulting object states is checked (lines 4–7).

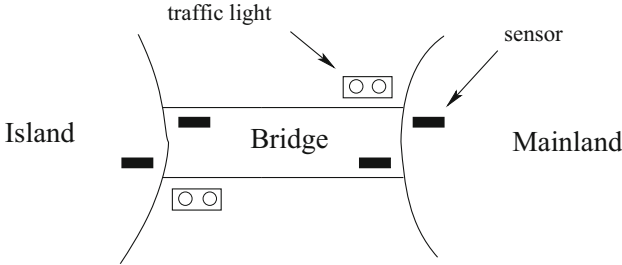
Note that we may introduce new methods under refinement where no abstract counterpart method exists. In such cases, we refine an abstract skip-method, i.e. we simply need to test that the new method  $N_2^m$  (`object2.m`) does not change the abstract state. This latter case is also known as stuttering. In addition, we need test-properties for checking that all methods preserve their (class) invariants.

The bold idea, we are following here, is that we will prompt ChatGPT in order to generate the test properties as well. One may argue that this generation of test-properties introduces another verification problem. Therefore, we need to review the test-properties. Our hypothesis is that these declarative test-properties are easier to review than the generated class implementations (code review vs. specification review).

## 4 The Bridge Controller Example

In this section we outline the requirements for the system to be developed, a bridge controller originally described in [2]. The objective is to develop a soft-





**Fig. 3.** The bridge environment from [2].

ware program, the *controller*, which controls the movement of cars on a bridge connecting a mainland and an island, the *environment*, shown in Fig. 3.

Cars can move from the mainland onto the bridge, and from there to the island, and similarly in the other direction. The following requirements are defined in [2], divided into requirements concerned with the functionality of the controller, labeled  $\text{FUN}_i$ , and requirements concerned with the environment, labeled  $\text{ENV}_i$ .

$\text{FUN}_1$ : *The system is controlling cars on a bridge connecting the mainland to an island.*

$\text{FUN}_2$ : *The number of cars on bridge and island is limited.*

$\text{FUN}_3$ : *The bridge is one-way or the other, not both at the same time.*

$\text{ENV}_1$ : *The system is equipped with two traffic lights with two colors: green and red.*

$\text{ENV}_2$ : *The traffic lights control the entrance to the bridge at both ends of it.*

$\text{ENV}_3$ : *Cars are not supposed to pass on a red traffic light, only on a green one.*

$\text{ENV}_4$ : *The system is equipped with four sensors with two states: on or off.*

$\text{ENV}_5$ : *The sensors are used to detect the presence of a car entering or leaving the bridge: “on” means that a car is willing to enter the bridge or to leave it.*

Note that  $\text{FUN}_3$  describes what is commonly referred to as a *one-lane* bridge. In the following text it will be referred to as a one-way bridge, like in [2]. In [2] the bridge controller is developed in four steps (three refinements), first a mainland and an island (without a bridge), then the bridge is introduced, then the traffic lights, and finally the sensors. In this paper we shall just develop the two first of these, including their refinement relation.

## 5 Guarded-Commands Development

In this first experiment we aim for developing a Scala program that is as close as possible to Abrial’s original Event-B model of the bridge controller [2]. Event-B follows a guarded-command style, i.e. a set of variables is iteratively updated by guarded commands. A guarded command is a conditional assignment that is enabled when the guard-condition is satisfied in a given state. In Event-B these guarded commands are called events. If more than one event is enabled, one is chosen non-deterministically.

In Scala, we can mimic this style with a class of methods with preconditions. Each method resembles an event. Semantically, we assume that these methods are non-deterministically called from the environment. If a precondition is not satisfied, the state does not change and another method is tried. In this way, we can write Event-B-like models in the Scala programming language. The main difference is that we require executable events. In the following, we report about our experience of letting ChatGPT write, refine and test the model.

## 5.1 Initial Model

Abrial starts with a very abstract model, consisting of two events that increase and decrease the number of cars on the island. In a first prompt, we tell ChatGPT what we want using the variable and method names of Abrial's model.

**Prompt:** Give me a Scala class `Bridge0` that abstractly models an island where at most `d` cars are allowed. A method `ml.in` shall model a car going from the island to the mainland. A method `ml.out` shall model the entering of a car from the mainland. At this level, the model state shall only be a variable `n`, the number of cars on the island. The following class invariant shall be respected: `n` shall be positive and not exceeding `d` and `d` shall be greater than 0.

The initial result was correct, but the style was not what we aimed for: ChatGPT produced a version with if-conditions, but we want preconditions. Furthermore the counter variable `n` was initialized to zero, but we prefer a more general parameterized initialisation. This is what we tell ChatGPT.

**Prompt:** The code compiles. Conditions are used, it would be better to use preconditions. Furthermore, it would be better to make `n` a variable parameter such that it can be set when the class is constructed.

This led to the following solution satisfying our requirements, which we name GC.V1 (Guarded Commands Version 1).

```

1  class Bridge0(d: Int, private var n: Int = 0) {
2      require(d > 0, "d must be greater than 0")
3
4      // Ensure n starts off with a valid value
5      require(n >= 0 && n <= d, "Initial number of cars on the island is
        invalid")
6
7      // method to model the entering of a car from the mainland to the
        island
8      def ml_out(): Unit = {
9          require(n < d, "Number of cars on the island exceeds limit!")
10         n += 1
11     }
12
13     // method to model a car going from the island to the mainland
14     def ml_in(): Unit = {
15         require(n > 0, "No cars on the island to remove!")
16         n -= 1
17     }
18
19     def carsOnIsland: Int = n
20 }

```

GC.V1.bridge

The above code is a parameterized Scala class with the maximum number of cars  $d$  and the state variable  $n$  being parameters. Note that  $d$  is constant while  $n$  can be changed after construction. The two methods resemble the two abstract events with guards ensuring that the invariant holds. The require-expressions have an error message to be included in the exception when the condition is violated. Here, also the invariant is encoded via two require-clauses. These clauses are checked upon construction of the object. Note that ChatGPT did a good job in commenting the program.

Next, we want to generate test-cases with ScalaCheck. Therefore, we let ChatGPT generate ScalaCheck properties that can be tested automatically. Obvious properties to be tested are the invariants.

**Prompt:** Given the class Bridge0, give me a ScalaCheck property that checks if the constructor and methods respect the following invariants: (1)  $n$  shall be positive, (2)  $n$  shall not exceed  $d$ , (3)  $d$  shall be greater than 0.

Two Iterations were needed, because ChatGPT generated a property that had several flaws. The code did not compile. Furthermore, the tests failed when a precondition was violated, but such tests should be ignored. In addition, one property did not catch exceptions and another property was useless, in fact, testing nothing. Hence, in the following prompt we communicated these observations in order to correct the flawed property.

**Prompt:** The last property can be skipped. It also does not compile. There seems to be a misunderstanding: we wanted a property that checks that the maximum allowed number of cars on the island is greater than zero. Furthermore, I think that the property “constructor” shall also test if an exception in case of invalid arguments is raised. Similar to the other two properties.

The result was the following ScalaCheck property.

```

1  object Bridge0Properties2b extends Properties("Bridge0") {
2    // Generator for valid d values (greater than 0)
3    val genD: Gen[Int] = Gen.posNum[Int]
4    ...
5    property("ml_out respects invariants") = forAll(genD) { d: Int =>
6      val bridge = new Bridge0(d, d) // initialize with maximum cars
7      val result = try {
8        bridge.ml_out()
9        bridge.carsOnIsland <= d
10     } catch {
11       case e: IllegalArgumentException => bridge.carsOnIsland == d
12     }
13     result :| s"ml_out with d = $d and n = ${bridge.carsOnIsland}"
14   }
15   ...

```

GC.V1.property-basted-test

The above listing shows the correct ScalaCheck property for method `ml_out()`. The code for `ml_in` is similar. First in Line 3, a generator for natural numbers is defined. It is used in the following property (Line 5) to generate values of the maximum number of cars  $d$ , i.e. the property should hold for all natural numbers  $d$ . Next, in Line 6 a bridge with the maximum number of allowed cars is created. Then, the method is called at Line 8. This call shall raise an exception leaving the number of cars on the island unchanged. Note that this property constitutes a boundary test at the upper limit, initializing the initial number of cars  $n$  to the maximum  $d$ . When running ScalaCheck it produces 100 random tests and all pass. The same holds for the property of method `ml_in`.

## 5.2 Refinement

In the following, we will refine the initial Scala model by introducing the one-way bridge going to the island.

**Prompt:**

- Given the class Bridge0, give me a refined class Bridge1 with three variables:
  - a: the number of cars on the bridge to the island,
  - b: the number of cars on the island,
  - c: the number of cars on the bridge from the island to the mainland.
  - The refinement relation is  $n = a + b + c$ .
- We need now two additional methods: `il_in` for cars going from the bridge to the island and `il_out` for cars leaving the island and going on the bridge.
- There is a new invariant: the bridge shall be a one-way bridge.

Again, the first attempt was not successful. The preconditions of `m1_out` and `m1_in` violated the one-way rule of the bridge, because ChatGPT mixed up variable names. In addition, the first version did not allow the free initialization of `a`, `b`, and `c`. Formulating these observations in a new prompt created a refined version of the bridge controller.

```

1  class Bridge1(d: Int, initialA: Int = 0, initialB: Int = 0, initialC:
    Int = 0) {
2  require(d > 0, "d must be greater than 0")
3  require(initialA + initialB + initialC <= d, "Initial number of cars
    exceeds
4                                     limit")
5
6  private var a: Int = initialA // cars on the bridge to the island
7  private var b: Int = initialB // cars on the island
8  private var c: Int = initialC // cars on the bridge from the island
    to the
9                                     // mainland
10
11 // One-way bridge invariants
12 private def oneWayInvariant(): Boolean = {
13   !(a > 0 && c > 0)
14 }
15 ...
16 // method to model the entering of a car from the mainland to the
    bridge
17 // towards the island
18 def m1_out(): Unit = {
19   require(c == 0, "Bridge is already in use by cars from the island
    to mainland")
20   require(b + c + a < d, "Number of cars exceeds limit!")
21   a += 1
22 }
23 ...
24 require(oneWayInvariant())
25 }

```

GC.V2.bridge

The above code snippet shows the part of the refined model with the new variables, its invariants and the refined version of `m1_out()`. This class can initialize

a, b, c individually, it correctly checks that their sum does not exceed the maximum d, and a new one-way invariant has been introduced and is checked. The preconditions are correct including nicely generated error messages.

The above code seems to be correct. In the following, we attempted to generate a property to test that refinement holds. The following prompt guided ChatGPT.

**Prompt:** Given the two classes Bridge0 and Bridge1. Generate a ScalaCheck property that checks if ml\_out of Bridge1 is a refinement of the same method in Bridge0 using the following refinement relation:  $n = a + b + c$ . The generated classes shall respect their invariants, hence all invalid classes shall be filtered out.

In addition, we included the code of Bridge0 and Bridge1 into the above prompt. This time, the results of ChatGPT were less convincing as the generated code included many logical flaws. In total, we needed ten iterations: four iterations in order to get a correct refinement property for method ml\_out and six iterations for the two new methods for entering and leaving the island from the bridge. The refined version of method ml\_in having the simplest precondition was immediately correct. The two main issues were (1) sophisticated but slightly wrong class generators that raised exceptions due to invariant violations and (2) precondition violations of methods were wrongly classified as test failures. The latter is an error that also appeared in the first test property for testing the initial model. Finally, after ten attempts, ChatGPT produced a test property for testing that the refinement property holds.

```

1  ...
2  def genValidBridgeStates(d: Int): Gen[(Bridge0, Bridge1)] = {
3    for {
4      a <- Gen.choose(0, d)
5      b <- Gen.choose(0, Math.max(0, d - a))
6      c <- Gen.choose(0, Math.max(0, d - a - b))
7      if (a > 0 && c == 0) || (c > 0 && a == 0) || (a == 0 && c == 0)
8      bridge0 = new Bridge0(d, a + b + c)
9      bridge1 = new Bridge1(d, a, b, c)
10   } yield (bridge0, bridge1)
11 }
12 ...
13 property("Bridge refinement for ml_out, ml_in, il_in, and il_out") =
14   forAll(genD) { d: Int =>
15     val validBridgeStates = genValidBridgeStates(d)
16     Prop.forAllNoShrink(validBridgeStates) { case (bridge0, bridge1) =>
17
18       // Properties for ml_out
19       val mlOutResult = if (bridge0.carsOnIsland < d && bridge1.
20         carsFromIsland == 0) {
21         bridge0.ml_out()
22         bridge1.ml_out()
23         bridge0.carsOnIsland == bridge1.carsToIsland + bridge1.carsOnIsland
24       }

```

```

23         bridge1.carsFromIsland
24     } else true
25     ...
26     mlOutResult && mlInResult && ilInResult && ilOutResult
27         GC.refinement.property

```

The generated property tests that `ml_out` and `ml_in` are correctly refined and that the new methods `il_in`, `il_out` do not change the abstract state, i.e. they refine *skip* (stuttering). Interestingly, ChatGPT came up with a property that combines the test results for all four methods (Line 26). The property initializes the classes such that  $a+b+c$  stay below maximum  $d$ . It also filters states that violate the one-way invariant. Note that both initialized bridge objects respect the refinement mapping  $n = a + b + c$ .

## 6 Object-Oriented Development

In this approach we gave ChatGPT more freedom by not asking for a solution based on guarded commands. Perhaps not surprising, it pursued an object-oriented solution, which we shall describe in this section. There were in fact two attempts, the first of which was abandoned after several prompts. We illustrate both attempts below, the first one only briefly and the second one in more detail. First, we provided ChatGPT with the following prompt.

**Prompt:** Write a Scala program, which models cars moving between two zones: a mainland and an island connected by a bridge. There are some requirements defined. These are as follows: [FUN1 ... FUN3, ENV1 ... ENV5 provided]. However, we want to develop this program using step-wise refinement where we start with an abstract program that ignores certain details, and then we want to refine the program in steps, adding more and more details. In the first step we ignore the bridge and just consider there to be a mainland and an island. There can only be maximally  $d$  cars on the island at any point in time. Initially there are 0 cars on the island. We need an operation for moving a car from the mainland to the island and for moving a car from the island back to the main land. Can you write the Scala 3 program for the first step?

### 6.1 A First Attempt

A first generated solution was cleverly suggesting a class representing the notion of a *place*, which could then either be the island, the mainland, or the bridge:

```

1     class Place(var name: String, var carLimit: Int, var currentCars: Int)
2     {
3         def addCar(): Boolean = {...}
4         def removeCar(): Boolean = {...}
5     }

```

Although the idea was appealing, the solutions were flawed, needing numerous prompts. Moving a car from a place to another place was placed in the main program and not as part of the `Place` class. If the source place had cars but the target place had no more space, a car will be moved from the source anyway, but not transferred to the target. It generally had problems with the bridge, how many cars to allow and in what directions they could drive, going from allowing only one car on the bridge to allowing cars in both directions, each of which is wrong. Also it moved the wrong number of cars. There were also compiler errors. It was decided to abandon this thread and start over. The following sections cover the second attempt.

## 6.2 Initial Model

We started over again with the exact same prompt. This led interestingly to a different solution introducing an `Island` class and a `Mainland` class, instead of the generic `Place` class. It, however, took two additional prompts to solve some issues, including some style issues concerning the design of classes (there was initially only an `Island` class and no `Mainland` class), and our wish that moving a car should just be one operation, not require two method calls. An error was also detected, that a car is moved to the island even when there are no cars on the mainland. This led to the following self explanatory solution, which we name OO.V1 (Object-Oriented Version 1).

```

1  class Island(maxCars: Int) {
2      private var currentCars: Int = 0
3
4      def canAcceptCar: Boolean = currentCars < maxCars
5
6      def addCarFromMainland(): Boolean = {
7          if (canAcceptCar) {
8              currentCars += 1
9              true
10         } else {
11             false
12         }
13     }
14
15     def moveToMainland(): Boolean = {
16         if (currentCars > 0) {
17             currentCars -= 1
18             true
19         } else {
20             false
21         }
22     }
23
24     def carsOnIsland: Int = currentCars
25 }

```

OO.V1.island



```

1  class Mainland {
2      private var cars: Int = 0
3
4      def addCar(): Unit = {
5          cars += 1
6      }
7
8      def moveToIsland(island: Island): Boolean = {
9          if (cars > 0 && island.canAcceptCar) {
10             cars -= 1
11             island.addCarFromMainland()
12             true
13         } else {
14             false
15         }
16     }
17
18     def carsOnMainland: Int = cars
19 }

```

OO.V1.mainland

We also asked ChatGPT to rename `Island.addCarFromMainland()` as `Island.addCar()` since it really just adds a car to the island. We then proceeded to ask ChatGPT to generate a property-based test with the following prompt.

**Prompt:** Can you write tests in ScalaCheck, using property-based testing, which test these classes?

This first resulted in a solution containing the following code.

```

1  property("moveToMainland transfers car from island to mainland") =
2  forall { (initialMainlandCars: Int, initialIslandCars: Int) =>
3      ...
4      if (beforeIsland > 0) {
5          (afterMainland == beforeMainland) && (afterIsland ==
6              beforeIsland - 1)
7      } else {
8          ...
9      }

```

As can be seen, this is not quite correct in that it only checks that a car is removed from the island, and not that it is added to the mainland. In fact, this corresponds to ChatGPT's generated explanation of the solution.

**Response:** ... If you move a car from the mainland to the island, the number of cars on the mainland should decrease by one, and the number on the island should increase by one, unless the island is full or the mainland has no cars. **If you move a car from the island to the mainland, the number of cars on the island should decrease by one, unless the island has no cars.**

The generated English language requirement is wrong with respect to moving a car from the island to the mainland by not requiring the car to be added to the mainland, and the code generation “correctly” follows that. A second issue was that it generated negative numbers and very large positive numbers as test inputs, which caused the test to run for a long time. Finally for the amusement, we prompted it with *“Is the ‘foreach’ construct the preferred Scala3 way to write a for loop?”* (finding it slightly unappealing), to which it responded: *“foreach is a method on collections and is still idiomatic for iterating over collections and performing side-effects”*, sticking to its guns. The final result after four prompts was as follows, including here only the test of moving a car from the island to the mainland.

```

1  object CarSystemSpec extends Properties("CarSystem") {
2      val maxCarsOnIsland = 5
3      val limitedGen = choose(0, 10)
4      ...
5      property("moveToMainland transfers car from island to mainland") =
6          forAll(limitedGen, limitedGen) {
7              (initialMainlandCars: Int, initialIslandCars: Int) =>
8                  val mainland = new Mainland()
9                  val island = new Island(maxCarsOnIsland)
10                 (0 until initialMainlandCars).foreach(_ => mainland.addCar())
11                 (0 until initialIslandCars).foreach(_ => island.addCarFromMainland
12                     ())
13
14                 val beforeMainland = mainland.carsOnMainland
15                 val beforeIsland = island.carsOnIsland
16                 island.moveToMainland()
17                 val afterMainland = mainland.carsOnMainland
18                 val afterIsland = island.carsOnIsland
19
20                 if (beforeIsland > 0) {
21                     (afterMainland == beforeMainland + 1) && (afterIsland ==
22                         beforeIsland - 1)
23                 } else {
24                     (afterMainland == beforeMainland) && (afterIsland ==
25                         beforeIsland)
26                 }
27             }
28         }
29     }

```

OO.V1.property-basted-test

As can be seen it has switched to use a “user defined” generator (Line 3) as a response to the prompt on negative and too large positive numbers. It first initializes the island and the mainland (lines 8-11). It then stores the pre-values of cars on the mainland and island (lines 13-14). Then it moves a car (Line 15).

Then it obtains the post-values of cars on the mainland and island (lines 16-17). The condition itself checks that a car is moved in case there were cars on the island, or else nothing is changed (lines 19-23).

When running the test, however, it failed with a message from ScalaCheck, which we directly fed back to ChatGPT as a prompt without any further explanations.

### Prompt:

```
! CarSystem.moveToMainland transfers car from island to mainland:
  Falsified after 0 passed tests.
> ARG_0: 0
> ARG_1: 1
> ARG_0_ORIGINAL: 10
> ARG_1_ORIGINAL: 8
Found 1 failing properties.
```

The error is caused by the fact that `island.moveToMainland()` does not actually add a car to the mainland. What is interesting is that ChatGPT itself corrects the error based on ScalaCheck's error message as a prompt, suggesting that one can imagine an automated feedback loop between a testing tool and an LLM. The corrected part of the code is shown below where a reference to the mainland has been added as parameter to the `moveToMainland` method (Line 3), and a `mainland.addCar()` statement has been added (Line 6).

```
1  class Island(maxCars: Int) {
2    ...
3    def moveToMainland(mainland: Mainland): Boolean = {
4      if (currentCars > 0) {
5        currentCars -= 1
6        mainland.addCar()
7      }
8      true
9    } else {
10     false
11   }
12   ...
13 }
```

OO.V1.island-corrected

## 6.3 Refinement

At this point we move forward and ask ChatGPT to perform a refinement with the following prompt.

**Prompt:** Now let's do the first refinement. We introduce a bridge between the mainland and the island. The bridge is a one way bridge in the sense that cars can only go in one direction at a time. We now need four operations: (1) for moving a car from the mainland to the bridge, (2) for moving a car from the bridge to the island, (3) for moving a car from the island to the bridge, and (4) for moving a car from the bridge back to the mainland. We do not yet introduce traffic lights.

There were several obstacles on the way to a working solution. There were some style issues, including again two method calls to make a move, inconsistent ways of checking whether a move is possible (in a condition: asking for forgiveness after a failed attempt versus asking for permission), naming of methods, and code that could be simplified. There was a syntax error in a private name not being accessible. There were also semantic errors, including that a car did not get added to mainland. There was wrongly also no limit to how many cars can be on the bridge. The  $\text{FUN}_2$  requirement stated: *The number of cars on bridge and island is limited*. This requirement was admittedly not quite precise. To make it precise, we required that in order to move a car from the mainland to the island, the sum of cars on the island and on the bridge going in the direction of the island should be less than the maximal allowed number of cars on the island. Further errors included that cars could go in both directions at the same time, or that it only allowed one car on the bridge at a time, and that the maximal number of cars on bridge was defined as constant rather than a function of cars on island.

Below is the final result after seven prompts, showing the implementation of the bridge, the island, and the mainland. First the `Bridge` class together with an enumerated type of directions.

```

1  enum Direction:
2      case ToIsland, ToMainland
3
4  class Bridge(island: Island, mainland: Mainland) {
5      private var _carsOnBridge: Int = 0
6      private var direction: Direction = Direction.ToIsland
7
8      def canAcceptCar(d: Direction): Boolean = {
9          if (d == Direction.ToIsland) {
10             return _carsOnBridge + island.carsOnIsland < island.maxCars &&
11                 (_carsOnBridge == 0 || direction == d)
12         }
13         _carsOnBridge == 0 || direction == d
14     }
15
16     def moveCarOntoBridge(d: Direction): Unit = {
17         _carsOnBridge += 1; direction = d
18     }
19
20     def moveToIsland(dest: Island): Boolean = {
21         if (_carsOnBridge > 0 && direction == Direction.ToIsland && dest.
22             canAcceptCar)

```

```

23     _carsOnBridge -= 1; dest.acceptCar(); true
24   } else { false }
25   }
26
27   def moveToMainland(): Boolean = {...}
28 }

```

OO.V2.bridge

The `Bridge` class above is relatively straightforward, except perhaps for the condition determining whether a car can be moved onto the bridge in some direction  $d$ , defined by the method `canAcceptCar` (lines 8-14). If the direction is towards the island, the number of cars on the island plus the current number of cars on the bridge should be less than the allowed maximum number of cars on the island. In any case, either there should be no cars on the bridge or the direction  $d$  must match the current direction of the other cars. The classes `Island` and `Mainland` below are relatively straightforward.

```

1  class Island(val maxCars: Int) {
2    private var currentCars: Int = 0
3
4    def canAcceptCar: Boolean = currentCars < maxCars
5
6    def carsOnIsland: Int = currentCars
7
8    def moveToBridge(bridge: Bridge): Boolean = {
9      if (currentCars > 0 && bridge.canAcceptCar(Direction.ToMainland))
10     {
11       bridge.moveCarOntoBridge(Direction.ToMainland)
12       currentCars -= 1
13     } else {
14       false
15     }
16   }
17
18   def acceptCar(): Unit = {
19     currentCars += 1
20   }
21 }

```

OO.V2.island

```

1  class Mainland {
2    private var cars: Int = 0
3
4    def addCar(): Unit = cars += 1
5
6    def carsOnMainland: Int = cars
7
8    def moveToBridge(bridge: Bridge): Boolean = {
9      if (cars > 0 && bridge.canAcceptCar(Direction.ToIsland)) {
10     bridge.moveCarOntoBridge(Direction.ToIsland)
11     cars -= 1
12     true
13   } else {
14     false
15   }
16 }
17 }

```

OO.V2.mainland

It was interesting to observe that the object-oriented design distributes the logic in a manner that makes it more difficult to ensure oneself that the code is correct compared to the guarded command approach. E.g. to understand how a car is moved from the island to the bridge one has to study several methods distributed over the two classes. In the guarded command solution the entire action is defined in one place.

We have now developed two versions of our bridge control system, an abstract initial version OO.V1 and a refinement OO.V2. We first performed a property-based test of OO.V2, which we will not report upon in detail here due to lack of space. It did not reveal any errors in V2. However, it is worth mentioning that it took 10 prompts to get the property-based test itself right. The main issue was that it repeatedly did not initialize all of mainland, bridge, and island correctly with an initial random (in ScalaCheck style) number of cars, it only initialized the place moved from. What is more interesting is that we fed the failed test outputs as prompts (as shown earlier) repeatedly without additional explanations. However, this process did not seem to converge, and ChatGPT needed an additional small hint at the end before success. After this property-based testing exercise we then proceeded to the actual refinement test with the following prompt.

**Prompt:** The task is now to show that the second version V2 indeed is a refinement of the first version V1. This is done by first defining a mapping  $M$  from the state of V2 to the state of V1, and then show that given a reachable state  $S2$  of V2, if we apply an action  $A2$  of V2 and reach a state  $S2'$ , then if we apply the corresponding action  $A1$  in V1 to  $M(S2)$  we get a state  $S1'$  such that  $S1' = M(S2')$ . Can you write a test using ScalaCheck that performs this test? I will give you the two versions V1 and V2 of the code [... the code of V1 and V2 ...].

Generating a refinement test was non-trivial and required 56 prompts, involving the following issues, amongst others. The initial test moved several cars from mainland to island over the bridge before testing the intermediate states. It should check one move at a time. It also initially tested only one direction. Also, at some point the condition verified was wrong, verifying that all cars on the bridge got moved to the island in one move. In another attempt, the test for moving a car from the bridge in V2 to the island compared this to moving a car from the island in V1 to the mainland (wrong direction). It also made wrong subtractions and additions. Some iterations were spent ensuring that the concrete version only made moves from the bridge to the mainland or island if the direction was correct. At some point it wanted to compare island (and mainland) objects with equality (`==`). This was an interesting suggestion, but it had not defined equality (the `equals` method), and this would also cause a refactoring of the code, which we at this point tried to avoid.

At some point it produced the following code for moving a car from the mainland to the bridge in the concrete version V2.

```

1  val movedToBridge = mainlandV2.moveToBridge(bridge)
2
3  val (mainlandV1, islandV1) = mapStateV2toV1(mainlandV2, islandV2, bridge
4      )
5
6  movedToBridge == mainlandV1.moveToIsland(islandV1)

```

First, a move is performed at the concrete V2 level (Line 1), storing the Boolean result (success or failure) in the variable `movedToBridge`. The *resulting* concrete V2 state is then mapped to the corresponding abstract state (Line 3). Finally, a move is performed at the abstract V1 level, starting from that abstracted state, and the result is asserted equal to the result of the concrete move, stored in the variable `movedToBridge` (Line 5).

There are a few errors here. First of all, it maps the post-state of the V2 move to V1, whereas it should have been the pre-state such that V1 can make a move from that. Second, it compares the move from mainland to the bridge in V2 to a move from the mainland to the island in V1, and these yield different results. Note that in V2, a car on the bridge going in the direction of the island is considered as still belonging to the mainland in V1, so in V1 no car should be moved to the island. In general, it had a hard time learning that cars on the bridge moving to the island in V2 belong to the mainland in V1, and similarly the other way, cars on the bridge moving to the mainland in V2 belong to the island in V1.

This is related to the problem of *stuttering* in V1. This is the situation where an action in the concrete version V2 does not correspond to an action in V1. This is normally in refinement proofs handled by allowing the abstract version, here V1, to stutter, to make a move that does not change its state. Stuttering was initially not modeled in the generated test code, which, however, interestingly was caught by the refinement test. Specifically, in the property above a move from mainland to the bridge in V2 is attempted compared to a move between mainland and island in V1, whereas it should have been a stuttering step in V1. We informed ChatGPT about stuttering with the following prompt.

**Prompt:** The [problem] is “stuttering”: stuttering is the situation where an action in the concrete version (V2 here) does NOT have a corresponding action in the abstract version. In this case, the starting state and the result state of the concrete action map to the same abstract state. We say that the action in the concrete version corresponds to a stuttering action in the abstract version (that does not change the state). So in the first property, `mainlandV2.moveToBridge` causes a car to be moved from the mainland to the bridge but since cars on the bridge going to the island are counted as belonging to the mainland in the abstract version, this corresponds to no action in the abstract state.

A different problem was related to the generator of V2 states, caught by a failing test. The error was caused by the refinement mapping from V2 to V1, which only adds to the island if there is space. The mapping calls `islandV1.addCar()`, which contains the check: `if (canAcceptCar)...`. This means that not all cars that are on the island and on the bridge in Mainland direction in V2 are necessarily added to the abstract island by the mapping if there are “too many” cars. The concrete state should be initialized such that the sum of cars on the island and on the bridge going towards the mainland is less than or equal the maximum number of cars allowed on the island. In the end we manually had to fix the generator of V2 states.

At several stages it wanted to change the V2 implementation to have the direction be part of the implementation. However, considering how fragile ChatGPT is it was decided not to follow that advice, trying to avoid redoing too much work. It was also necessary to remind it of what versions V1 and V2 were. It also changed the properties to test even when they were ok, and it had to be asked to go back to previous definitions, reminding it what they were. These are issues that likely would be avoided with a system such as Copilot.

The following refinement test using ScalaCheck was finally generated. First it generated the following generator of inputs.

```

1  object RefinementProperties extends Properties("Refinement") {
2    val genCars = for {
3      m <- Gen.choose(0, 10)
4      maxIslandCars <- Gen.choose(0, 10)
5      i <- Gen.choose(0, maxIslandCars)
6      d <- Gen.oneOf(Direction.ToMainland, Direction.ToIsland)
7      b <- Gen.choose(0, maxIslandCars - i)
8    } yield (m, b, i, maxIslandCars, d)
9    ...
10 }                                     OO.refinement.generator

```

What is of interest here is the number of cars generated for the bridge (Line 7), taking into account that the number of cars on the island and bridge should reflect (not surpass) the maximum number of cars allowed on the island. This required numerous prompts.

Next, the refinement mapping is defined by the `mapStateV2toV1` method shown below.

```

1  def mapStateV2toV1(mainlandV2: MainlandV2, islandV2: IslandV2,
2                      bridge: Bridge): (MainlandV1, IslandV1) =
3  {
4    val mainlandV1 = new MainlandV1()
5    val islandV1 = new IslandV1(islandV2.maxCars)
6
7    (1 to mainlandV2.carsOnMainland +
8     (if (bridge.getDirection == Direction.ToIsland) bridge.
9        carsOnBridge else 0))
10   .foreach(_ => mainlandV1.addCar())
11
12   (1 to islandV2.carsOnIsland +
13    (if (bridge.getDirection == Direction.ToMainland) bridge.
14       carsOnBridge else 0))

```



```

13     .foreach(_ => islandV1.addCar())
14
15     (mainlandV1, islandV1)
16   }
OO.refinement.mapping

```

It takes as argument a mainland, an island, and a bridge of version OO.V2 and returns a tuple consisting of the corresponding mainland and island of version OO.V1. It first creates new instances of OO.V1 mainland and island (lines 4-5). It then adds cars to the mainland (lines 7-9). That is, for each car in the OO.V2 mainland plus the cars on the bridge going in direction of the island, it adds a car to the OO.V1 mainland. Similarly, it adds cars to the OO.V1 island that are on the bridge going in the direction of the mainland (lines 11-13), in addition to the cars on the OO.V2 island. At the end, it returns the updated OO.V1 mainland and island (Line 15).

Finally, the refinement property, corresponding to Definition 5.3, for moving a car from the bridge to the island is shown below.

```

1   property("Bridge to Island move is consistent") = forAll(genCars) {
2     case (m, b, i, maxIslandCars, d) =>
3       val mainlandV2 = new MainlandV2()
4       val islandV2 = new IslandV2(maxIslandCars)
5       val bridge = new Bridge(islandV2, mainlandV2)
6       (1 to m).foreach(_ => mainlandV2.addCar())
7       (1 to b).foreach(_ => bridge.moveCarOntoBridge(d))
8       (1 to i).foreach(_ => islandV2.acceptCar())
9
10      val (mainlandV1, islandV1) =
11        mapStateV2toV1(mainlandV2, islandV2, bridge)
12
13      val movedToIslandV2 = bridge.moveToIsland(islandV2)
14      val movedToIslandV1 =
15        if (movedToIslandV2) mainlandV1.moveToIsland(islandV1) else
16          false
17
18      val (mainlandV1Post, islandV1Post) =
19        mapStateV2toV1(mainlandV2, islandV2, bridge)
20
21      (movedToIslandV2 == movedToIslandV1) &&
22        mainlandV1Post.carsOnMainland == mainlandV1.carsOnMainland &&
23        islandV1Post.carsOnIsland == islandV1.carsOnIsland
OO.refinement.property

```

It first initializes a OO.V2 state consisting of a mainland, an island, and a bridge (lines 3-8). It then maps this concrete OO.V2 state to an OO.V1 state using the refinement mapping (Line 10). It then performs the moves, a concrete on OO.V2 and an abstract on OO.V1 if there was a concrete move (lines 13-15). Then it maps the resulting OO.V2 state to a OO.V1 state using the refinement mapping (Line 17). At the end, it checks the commuting diagram condition, that the OO.V1 state obtained by the abstract move is the same as the state obtained by mapping the resulting OO.V2 state to a OO.V1 state using the refinement mapping (lines 20-22).

## 7 Conclusion

The developments using an LLM like ChatGPT turned out not to be effortless. This is likely due to the fact that the problem being solved is not a common pattern occurring in open source software. We let ChatGPT generate all the code (except for a couple of exceptions), including test properties, since we tried to see how far this approach could be taken. In some cases ChatGPT went off in a wrong direction and had to be steered back on the right track. The experience was different than the “*how do you do X in programming language Y*” prompts that ChatGPT is so good at.

It turned out that the object-oriented approach required almost 5 times as many prompts (80) than the guarded command approach (17). There can be several reasons for that but we did observe that it was easier to convince oneself that a guarded command program was correct than an object-oriented program. We attribute this to the fact that in an object-oriented solution, a move (e.g. from bridge to mainland) was distributed over several methods in different classes. This aspect could potentially impact the performance of ChatGPT as well.

One learns to write better requirements through an effort like this, what is also referred to as *prompt engineering*. One can imagine requirements being translated to a formalism, as in our case, which can then be checked for soundness and completeness. Once the generated artifact is “good enough”, the requirements are possibly also “good enough”.

However, what required the most effort was to read and comprehend the generated code, since many prompts lead to many results that have to be investigated, which was quite tiresome. ChatGPT does not point out (in the code) what has been changed, so we had to read all the generated code after each prompt. A system integrated in an IDE, such as Copilot, should do a better job here. As it turned out, it was easier to read and understand the generated tests, and in particular the property-based tests. This suggests that tests, and in particular succinct and readable specifications, become important in evaluating generated code.

Our process consisted, for each result, of writing down observations, and then re-formulating them as a prompt. In case of failed test cases we first tried to feed them back directly as prompts without modification. As we showed, this worked in some cases in the sense that ChatGPT “understood” what the problem was and fixed it. In other cases this approach did not converge. Since the process involved prompt engineering in natural language (English), the results shall be interpreted with a grain of salt. Prompt engineering is an informal process. Nonetheless, we believe that we can conclude that testing and specification are important concepts in this context.

Wrt. future work, a challenge for research in this area is the unknown nature of future improvement of LLMs’ capabilities. If these models become significantly better, research done now may quickly become obsolete (depending on the kind of research of course). It will be interesting to carry out the same experiment with Version 5 of ChatGPT, which is expected to be launched within months of the time of writing, as well as with other LLM models. It would also be interesting to

carry out the same experiment using an LLM IDE such as Copilot. One feature that we were missing during the work was access to a tree-like structure of the development branches (perhaps supported by a github), where one easily can “walk” up and down the tree of experiments.

## References

1. Abadi, M., Lamport, L.: The existence of refinement mappings. *Theoret. Comput. Sci.* **82**, 253–284 (1991)
2. Abrial, J.: *Modeling in Event-B - System and Software Engineering*. Cambridge University Press, Cambridge (2010)
3. Ahrendt, W., Gurov, D., Johansson, M., Rümmer, P.: TriCo-triple co-piloting of implementation, specification and tests. In: Margaria, T., Steffen, B. (eds.) *Leveraging Applications of Formal Methods, Verification and Validation. Verification Principles, ISO/FA 2022, LNCS*, vol. 13701, pp. 174–187. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-19849-6\\_11](https://doi.org/10.1007/978-3-031-19849-6_11)
4. Aichernig, B.K., Schumi, R.: Property-based testing of web services by deriving properties from business-rule models. *Softw. Syst. Model.* **18**(2), 889–911 (2019)
5. Arts, T., Hughes, J., Norell, U., Svensson, H.: Testing AUTOSAR software with QuickCheck. In: *Software Testing, Verification and Validation Workshops (ICSTW)*, 2015 IEEE Eighth International Conference on, pp. 1–4, April 2015
6. Back, R.-J., Wright, J.: *Refinement Calculus - A Systematic Introduction*. Texts in Computer Science (TCS) (1998)
7. Belzner, L., Gabor, T., Wirsing, M.: Large language model assisted software engineering: prospects, challenges, and a case study. In: Steffen, B. (eds.) *Bridging the Gap Between AI and Reality, AISoLA 2023, LNCS*, vol. 14380, pp. 355–374. Springer, Cham (2024). [https://doi.org/10.1007/978-3-031-46002-9\\_23](https://doi.org/10.1007/978-3-031-46002-9_23)
8. Bensalem, S., Cheng, CH., Huang, W., Huang, X., Wu, C., Zhao, X.: What, indeed, is an achievable provable guarantee for learning-enabled safety-critical systems. In: Steffen, B. (ed.) *Bridging the Gap Between AI and Reality, AISoLA 2023, LNCS*, vol. 14380, pp 55–76. Springer, Cham (2024). [https://doi.org/10.1007/978-3-031-46002-9\\_4](https://doi.org/10.1007/978-3-031-46002-9_4)
9. Busch, D., Nolte, G., Bainsczyk, A., Steffen, B.: ChatGPT in the loop: a natural language extension for domain-specific modeling languages. In: Steffen, B. (ed.) *Bridging the Gap Between AI and Reality, AISoLA 2023, LNCS*, vol. 14380, pp. 375–390. Springer, Cham (2024). [https://doi.org/10.1007/978-3-031-46002-9\\_24](https://doi.org/10.1007/978-3-031-46002-9_24)
10. Claessen, K., Hughes, J.: QuickCheck: a lightweight tool for random testing of Haskell programs. In: *Proceedings of the Fifth ACM SIGPLAN International Conference on Functional Programming, ICFP 2000*, pp. 268–279. New York, NY, USA, ACM (2000)
11. Dijkstra, E.W.: Guarded commands, nondeterminacy and formal derivation of programs. *Commun. ACM* **18**, 453–457 (1975)
12. Havelund, K., Shankar, N.: A refinement proof for a garbage collector. In: Bartocci, E., Cleaveland, R., Grosu, R., Sokolsky, O. (eds.) *From Reactive Systems to Cyber-Physical Systems*. LNCS, vol. 11500, pp. 73–103. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-31514-6\\_6](https://doi.org/10.1007/978-3-030-31514-6_6)
13. Jones, C.B.: *Systematic Software Development Using VDM*. UK, 2nd, Prentice Hall, Hemel Hempstead (1990)

14. Lamport, L.: The temporal logic of actions. *ACM Trans. Program. Lang. Syst.* **16**(3), 872–923 (1994)
15. Lampropoulos, L., Sagonas, K.F.: Automatic WSDL-guided test case generation for PropEr testing of web services. In: Silva, J., Tiezzi, F. (eds.), *Proceedings 8th International Workshop on Automated Specification and Verification of Web Systems*, vol. 98 of EPTCS, pp. 3–16 (2012)
16. Morgan, C.: *Programming from Specifications*. Prentice Hall, Hoboken (1990)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# **Automotive Driving**



# Safer Than Perception: Increasing Resilience of Automated Vehicles Against Misperception

Martin Fränzle<sup>1</sup> and Andreas Hein<sup>2</sup>

<sup>1</sup> Research Group Foundations and Applications of Systems of Cyber-Physical Systems, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[martin.fraenzle@uo1.de](mailto:martin.fraenzle@uo1.de)

<sup>2</sup> Research Group Assistance Systems and Medical Technology, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[andreas.hein@uo1.de](mailto:andreas.hein@uo1.de)

**Abstract.** Autonomous vehicles (AV) rely on environmental perception to take manoeuvre decisions. Safety assurance for AV thus hinges on achieving confidence in all percepts that are safe-guarding critical manoeuvres. As the safety targets for such critical manoeuvres are considerably higher than the statistical figures for the reliability of at least current learning-enabled classification algorithms within the environmental perception, we need means for assuring that the overall system is “safer than perception” in that the frequency of erratically adopting a critical manoeuvre is considerably lower than the frequency of individual misclassifications. We present a methodology for constructively generating reformulations of guard conditions that are more resilient to misperception than the original condition. The synthesized, rephrased guard conditions reconcile a given safety target, i.e. a given a societally accepted upper bound on erratically activating a critical manoeuvre due to a false positive in guard evaluation, with maximal availability, i.e. maximal true positive rate. By synthesizing a resilient rephrasing of the guard condition, the constructive setting presented herein complements the analytical setting from a previous companion paper [6], which merely analysed a given condition for its safety under uncertain perception.

**Keywords:** Safety-critical perception · Decision making · Robustification

## 1 Introduction

Decision-making based on—inherently uncertain to some extent—environmental perception is a key element of providing cyber-physical systems, like transportation systems in general and road vehicles in particular, with forms of autonomy,

Supported by the Ministry of Science and Culture of the State of Lower Saxony under grant number ZN3493 as well as by Deutsche Forschungsgemeinschaft under grant no. DFG FR 2715/5-1.

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 415–433, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_25](https://doi.org/10.1007/978-3-031-73741-1_25)

as in highly automated or autonomous driving. Such decision-making obviously is safety-critical, as the actions adopted in consequence of a decision have physical impact and can consequently incur risk to life, health, and property. Especially in the field of automated vehicles, societal expectations concerning the risk induced by automated driving functions, and thus ultimately for error rates decision-making, are very high: while manual driving already is amazingly safe at considerably more than a million kilometres driven on average between two accidents incurring some form of injury, the public debate as well as relevant authorities tend to require highly automated vehicles (HAV) to even further reduce the overall rate of injuries and fatalities compared to human-operated vehicles.

How rare accidents of HAV must be is a matter of ongoing societal debate, but the societal acceptance threshold will obviously be orders of magnitude below the misperception rates that can be realized by or guaranteed of<sup>1</sup> any perception system containing machine-learned components, which can only be trained and qualified on examples. Concerning the three types of uncertainties that these systems inevitably are prone to, namely

1. *existential uncertainty*, i.e. not knowing whether all or at least all relevant environmental objects have been detected,
2. *classification uncertainty*, i.e. uncertainty in exactly classifying the type, like “car“, “adult pedestrian“, “playing child“, or “waste bin“, of any detected object, and
3. *state uncertainty*, i.e. inaccuracies in determining relevant physical quanta, like speed or distance, of a classified object,

uncertainties especially concerning the first two remain relatively high. Even if those machine-perception systems could (and currently they cannot) guarantee significantly better performance w.r.t. these two criteria than human vision within complex street scenes and at any level of environmental disturbance, like rain, fog, or blinding sun, their error rates would still remain orders of magnitude higher than the inherently strict safety target expected of automated vehicles. This implies that a significant gap remains to be bridged here, namely the gap between actual performance of technical perception and expected societal acceptance thresholds for unjustified manoeuvres.

This paper sets out to narrow this gap by answering the following three questions affirmatively:

1. Can we provide a mathematical or logical formalization of relevance of a percept such that we understand when a misperception remains redundant to a decision, i.e. either does not propagate into a—then unjustified—decision or does not harm the safety of the decision?
2. Can we demonstrate the positive safety impact of such redundancies in that we rigorously show that actual guard conditions are “safer than perception“

---

<sup>1</sup> Note that for extremely low error rates, realizing them technically and providing evidence for their satisfaction are completely different, both very hard, issues.

in that their evaluation exposes considerably lower error rates than the perception, which is their input?

3. Can we provide a mechanism automatically rewriting a safety-critical guard condition into a more robust variant that retains the logical content of the original condition, yet offers resilient evaluation under uncertainties in that it provides false positive rate below the societal acceptance threshold while maximizing true positive rates?

In answering the first question, we will follow the lines of the precursor paper [6], which also addressed the second question by means of safe, i.e. pessimistic approximation of the quantitative risk. We will herein complement its analysis by a precise closed-form analysis of a special case, which sheds more light on the actual safety level to be expected and confirms that safety gains overarching orders of magnitude are indeed plausible. The last question remained an open issue for future research in [6] and we are pleased of now being able to expose an algorithm that can constructively construct optimized guard formulations.

This paper is organised as follows: Sect. 2 sketches a reference architecture used throughout the subsequent discussions while Sect. 3 provides a simple example showcasing the effect of why and how the evaluation of a complex guard condition safeguarding a safety-critical manoeuvre can be “safer than perception”. The subsequent Sects. 4 and 5 develop the mathematical framework facilitating quantitative analysis of this effect and prove its existence. Section 6 then sketches an automatic rewriting technique maximizing the resilience of a critical guard condition while keeping detection performance at a requested level. Sections 7 and 8, finally, refer to related work and provide a summary and pointers to future work.

## 2 Preliminaries

For solidly basing our analysis, we postulate a certain reference architecture. As discussed in [6], which this section is based on, the exemplary reference architecture uses labelled occupancy grids for collecting the output of machine-learning based algorithms that classify objects in the environment of the ego-vehicle. Class labels are assigned according to a (generally partially ordered, e.g., collecting cars, trucks, motor-cycles, etc. into a super-class of vehicles) ontology. The occupancy grid partitions the geometric vicinity of the ego car into finitely many grid elements. Its grid elements are filled with the corresponding class labels from the ontology whenever they have been perceived as being occupied by an object. Postulating this particular reference architecture is a matter of convenience, as it provides the subsequent analysis with a defined basis, but does by no means imply that the analysis would fail for other architectures, like those representing the world model by an object list—in fact, it carries over, as the models are mostly isomorphic (we discuss this in some more depth in the conclusion). Our analysis ought consequently, *cum grano salis*, carry over to the highly proprietary implementations of original equipment manufacturers and their suppliers.

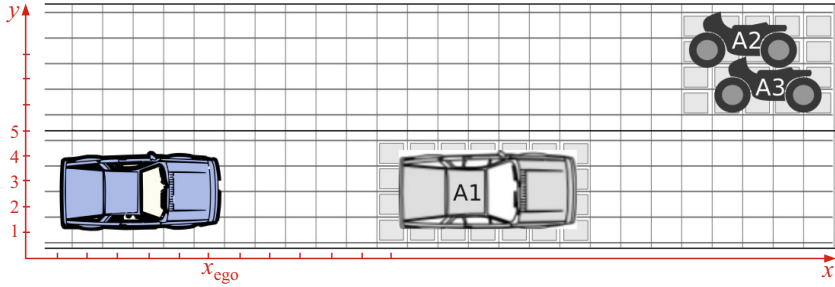


Typical conditions enabling or blocking—and thereby meant to safeguard—critical manoeuvres then take the form of Boolean combinations of statements concerning the occupancy of certain elements of the occupancy grid, with these elements together forming grid areas which correspond to subspaces of the surrounding traffic space. The atoms of such statements query occupancy of a particular grid element by certain object types named in the ontology, plus maybe additional unlocalized environmental conditions, like general visibility conditions. As an example take an evasive manoeuvre of a car across the curb to the footpath in order to make room for an emergency vehicle: Initiation of such a manoeuvre by the ego car would naturally be safeguarded by a Boolean condition requiring (1) presence of an emergency vehicle somewhere in the occupancy grid elements belonging to the traffic space reasonably close behind the ego car, (2) absence of vulnerable road users within some sufficiently large and connected group of occupancy grid elements belonging to the bike lane and footpath just ahead of the ego vehicle, (3) absence of any obstacles, including parked or stopped cars, on the line between the current ego position and the space for evasion identified via the previous condition, and finally (4) general (like absence of dense fog, presence of illumination) and geometric (like absence of occlusions) visibility conditions pertaining to the critical objects mentioned throughout the previous conditions.

While all the sub-conditions of the above guard condition intuitively make sense as being necessary conditions for safe execution of the safety-critical evasive manoeuvre, a safety risk due to misperception of some of the atomic statements occurring in the guard condition prevails, as no technical (nor a biological) perception system is perfect. In complex road scenes, we can neither expect to detect all potentially relevant objects nor are safe from misclassification of harmful objects as harmless and irrelevant. With absolute object detection rates often dropping below  $\frac{2}{3}$  and classification accuracy easily falling below 90% in non-ideal visibility conditions [8], would our reliance into the evaluation of the guard condition drop into similar ranges due to the weakest link principle?

As the condition guarding the manoeuvre decision is a massive Boolean combination of atomic percepts, individual misperceptions might mask each other: not every single pedestrian needs to be detected, as safely crossing a pedestrian lane does not depend on the particular number of pedestrians being present. Likewise, slight misplacements of perceived objects is irrelevant, as e.g. a slight offset in locating a cyclist will not change drivability of the manoeuvre. When the guard condition reflects these properties, this can induce a considerably lower misevaluation rate for the overall condition than for its constituents, i.e. than for the atomic percepts dealing with detecting, locating, and classifying objects. Within this note, we are rendering this intuitive argument rigorous and formal, thus lifting reliability levels of combinatorial critical environmental perception well beyond the figures for atomic percepts achieved by state-of-the-art perception [8] paired with fusion techniques [14].

The main result of this paper is a methodology for, first, formally establishing and, second, constructively optimizing refined bounds on the risks of



**Fig. 1.** A traffic situation requiring a safety-critical manoeuvre decision

misperception for guard conditions concerning safety-critical manoeuvres based on the rates of misperception of atomic environmental artefacts. It complements formal synthesis-based approaches towards achieving safe controllers as well as engineered control architectures, as our reference architecture does not restrict the typically highly proprietary planning and manoeuvre control of HAV, and instead provides a generic interface between any such proprietary solutions and the perception chain. It does so by allowing to tune guard conditions pertaining to critical manoeuvre decisions based on the confidence level of the underlying individual percepts, allowing to optimise the trade-off between availability (induced from enabling the guarded action) and safety (bounding the error of erratically enabling an action due to misconception to societally accepted risk).

### 3 A Simple Example

The following example is taken from [6]. Consider the blue ego car being in the situation depicted in Fig. 1. When the ego car detects the obstacle A1, it will consider an evasive manoeuvre across the dividing line into the oncoming lane. This manoeuvre would, however, only be adopted if (1) it is necessary to avoid collision on the originally planned track and (2) it is considered safe w.r.t. the available information about the environmental state.

The manoeuvre would thus be (safe-)guarded by a guard condition  $g$  defined as  $g \equiv \textit{necessary} \wedge \textit{safe}$ , where

$$\textit{necessary} \equiv \bigvee_{x=6}^7 \bigvee_{y=1}^5 \textit{obstacle}@(\mathit{x}_{\textit{ego}} + x, y)$$

denotes that some type of obstacle is detected as being present on the own lane, i.e. between 1 and 5 in  $y$  position, within relevant  $x$  distance (here, for the sake of being able to depict the example, shown as just 6 to 7 grid elements ahead in the  $x$  direction; the real figure would be considerably larger). Within the ontology, *obstacle* denotes an arbitrary type of road-blocking object and is defined as a disjunction about different basic object classification labels, like *trash container*,

*tire, debris, or parked/slow car* (cf. Fig. 1). Note that this very definition already induces some fault-tolerance w.r.t. to misperception of atomic percepts: identifying the necessity for circumvention neither requires identifying the full back frontier of the obstacle, as the disjunction across  $y$  positions would evaluate to true already if only a fraction of the frontier is detected, nor identifying correctly the exact type of obstacle, as *obstacle* is a disjunction across numerous obstacle types. Even identification of the  $x$  position of the obstacle would permit for tolerances if circumvention manoeuvres are dimensioned with a safety margin: locating A1 further left than it actually is would not cause risk (yet extend the circumvention), while locating it too far right stays collision-free if the misplacement remains within the safety margin. Note the combinatorially vast number of distorted perceptions of A1 that would thus still lead to the same truth value as the ground truth does. The likelihood of failing to detect the necessity of a circumvention consequently remains considerably lower than the unreliability of atomic percepts. This implies that the rate of false negative verdicts in the evaluation of *necessary* remains comparatively low. We will later see that, by just some rewriting to the way *necessary* is expressed, we will also be able to reduce the false-positive rate of the evaluation of *necessary* further to a frequency well below the false-positive rate of the atomic percepts.

We now turn to the safety condition, yet do in this note simplify its exposition slightly by omitting some additional conditions that are structurally perfectly similar to the ones shown. These omissions deal with occluded areas and are perfectly symmetric to the conditions on oncoming traffic explicated in the sequel. With these simplifications, the safety condition reads

$$safe \equiv \neg \bigvee_{y=6}^{10} \left( \begin{array}{cc} \bigvee_{x=1}^{20} pedestrian@ (x_{ego} + x, y) & \vee \\ \bigvee_{x=-4}^{40} car@ (x_{ego} + x, y) & \vee \\ \bigvee_{x=-1}^{60} motorcycle@ (x_{ego} + x, y) & \end{array} \right).$$

Its evaluation determines the presence of critical objects in the oncoming lane within the ego car's vicinity, constituting the safety condition that may block the circumvention manoeuvre when its execution may become hazardous. As this condition *safe* structurally resembles *necessary* with an outermost negation added, its fault-tolerance properties are in principle dual: where *necessary* is massively disjunctive and therefore tolerant against some or even numerous lacking or inaccurate percepts, *safe* as a negation over a disjunction essentially is conjunctive and consequently seems to require completeness of all percepts across the large set of atomic observations it mentions. This would imply that the very safety condition *safe* were not only as, but even orders of magnitude more fragile against misperception than any of the atomic percepts involved! Sufficiently reliable evaluation of the safety condition would consequently seem elusive, given that reliability of atomic percepts already falls considerably short of our actual safety targets. Fortunately, we will see that also here, a rewriting of the condition *safe* can help. Due to the duality, this rewriting now would have to reduce the false-negative rate in the evaluation of *safe*, thus being dual to the one applied to *necessary*.

The argument that we want to either minimize false-positive rate or false-negative rate of a disjunctive state condition, depending on the polarity of its occurrence, does however only apply when formula like *necessary* or *safe* stand in isolation. Boolean combinations of such disjunctive state conditions, where the satisfying violations of subformulae occurring in opposite polarity (like *necessary* and *safe* in the example) can overlap, may require compromises. Furthermore, an—in principle desirable—minimization of false detections (i.e. of false-positive rate or false-negative rate, resp.) may not always be appropriate, as it unfortunately also tends to minimize the true detection (i.e. true-positive rate or true-negative rate, resp.), thus maximizing safety at the price of minimizing availability of a—presumably useful—action.

Luckily, this problem can be alleviated by careful analysis (and modification whenever beneficial) of the Boolean problem structure of the conditions *safe*, *necessary*, and the guard condition  $g$  within the general mathematical framework provided in the next section. Subsequent sections will then exploit the framework to rigorously quantify the reliability gain that the Boolean structure of the guard condition provides over the atomic percepts, as well as show how to constructively rewrite the guard condition by phrasing its true-positive and false-positive rates as a constrained optimization problem.

## 4 Boolean Guard Formulae as Classifiers Under Uncertainty

Let  $\Phi$  be a formula that guards a safety-critical manoeuvre in the sense that the driving function will only adopt the manoeuvre when it has positive evidence of the validity of  $\Phi$  in the current situation, implying that the manoeuvre would be avoided (and a safer substitute adopted) whenever  $\Phi$  is violated *or* evaluation of  $\Phi$  remains inconclusive. The formula  $g \equiv \textit{necessary} \wedge \textit{safe}$  from the previous section is an example of such a guard condition  $\Phi$ .

Generally, such formulae  $\Phi$  comprise massive Boolean combinations of conditions on individual cells of the occupancy grid, where both the particular cells referenced and the individual conditions vary situationally. E.g.  $\Phi$  may safeguard a transit through a shared traffic space by ensuring that there are no vulnerable road users in the street, where the geometric position of the referenced areas of the occupancy grid depend on the own car’s position as well as the particular geometry of the shared space and the planned trajectory through that space. In this particular setting,  $\Phi = s_1 \wedge s_2 \wedge \dots \wedge s_n$  is a conjunction of statements  $s_i = \neg o_i$ , where  $o_i = \textit{pedestrian}_i \vee \textit{cyclist}_i$  is a disjunction of atomic percepts  $a_{i,j}$  expressing the property “there is a vulnerable road user of type  $j$  at the cell  $c_i$  of the occupancy grid”. The truth value of each atom  $a_{i,j}$  therein directly depends on a classifier output, which is a classifier for the object classes “pedestrian” or “cyclist”, resp., in this particular example.

The central problem we are facing obviously is that the percepts  $a_{i,j}$  are not reliable wrt. ground truth  $A_{i,j}$ , as there is a non-trivial risk for lacking perception of an object or for misclassification of a perceived object. We consequently

have to distinguish between the ground truth  $A_{i,j}$  underlying such an atomic percept and the possibly distorted percept  $a_{i,j}$ . E.g., it may be true that there is a pedestrian at cell  $i$  (i.e.  $A_{i,pedestrian}$  holds), but we misperceive her as a waste bin (i.e.  $\neg a_{i,pedestrian}$  as well as  $a_{i,waste\ bin}$  hold).

The first—rather trivial, yet crucial—observation is that there is no direct need to align  $a_{i,j}$  with  $A_{i,j}$ , i.e. to minimize the misperception rates of individual classifiers, but only a need to sufficiently reduce the misevaluation rate of the compound condition  $\Phi$ . Let us denote by  $\mathcal{GT}(\Phi)$  the formula  $\Phi[a_{1,1}, \dots, a_{n,m}/A_{1,1}, \dots, A_{n,m}]$  where all percepts  $a_{i,j}$  have been replaced by their (factually unknown) ground truth  $A_{i,j}$ . Then, in any situation  $\sigma$  assigning truth values to all ground-truth atoms  $A_{i,j}$  as well as to all percepts  $a_{i,j}$ , the truth value  $\mathcal{GT}(\Phi)(\sigma)$  represents the (desired, yet unknown in practice) actual value of the guard condition  $\Phi$ , while  $\Phi(\sigma)$  is the result of evaluating  $\Phi$  on the potentially distorted percepts. We follow the tradition to write  $\sigma \models \psi$  if  $\psi(\sigma)$  evaluates to true and  $\sigma \not\models \psi$  if  $\psi(\sigma)$  evaluates to false, for any formula  $\Phi$ . Thus, we call  $\sigma$  a *false positive* for  $\Phi$  iff  $\sigma \models \mathcal{GT}(\Phi)$  while  $\sigma \not\models \Phi$ . We call  $\sigma$  a *true positive* for  $\Phi$  iff  $\sigma \models \mathcal{GT}(\Phi)$  and  $\sigma \models \Phi$ .

As false positives induce risk (e.g. due to suggesting overtaking when it actually is unsafe) while true positives are constitutional for system performance (e.g. enabling overtaking whenever safely possible), our obligation then is to ensure that the false-positive rate remains below a defined threshold  $\Theta$  pertaining to societally acceptable risk while maximizing true-positive rate. As  $\Phi$  is a complex Boolean combination of atomic statements  $A_{i,j}$ , this is not identical to the problem of maximizing the true-positive rates and minimizing the false-positive rates for the classifiers generating the percepts  $a_{i,j}$ . Simple as the above observation is, it already has profound consequences for the pragmatics of developing safety-critical autonomous systems, as it implies that the currently prevailing isolated optimization of computer vision for high detection rates may not be the most effective approach towards overall system safety and performance. We will not elaborate on this issue within this note, yet leave it to future exploration.

## 5 Equivalence of Logically Distinct Guard Conditions with Respect to Ground Truth

The second—and more profound—observation is that the above problem statement provides us with liberty in phrasing the condition  $\Phi$ . Especially if the real world satisfies some relevant invariants  $\iota$ —which it inevitably does—then we can rephrase  $\Phi$  into  $\Phi'$  such that in all worlds satisfying the invariant  $\iota$ , the two formulae  $\Phi$  and  $\Phi'$  evaluate identically, i.e.

$$\iota \models \mathcal{GT}(\Phi) \iff \mathcal{GT}(\Phi') \quad , \quad (1)$$

where, as usual,  $\Phi \models \psi$  denotes that any model of  $\Phi$  (i.e. every world satisfying  $\Phi$ ) also is a model of the formula  $\psi$  (i.e.  $\psi$  holds too in those worlds). In the

sequel, we call formulae satisfying Eq. (1) world equivalent (W-equivalent for short):

**Definition 1.** *Given a propositional invariant  $\iota$ , two propositional formulae  $\Phi$  and  $\Phi'$  over atoms  $a_{i,j}$  referring to percepts are called W-equivalent modulo  $\iota$  iff property (1) holds.*

The interesting property is that W-equivalent formulae, despite always agreeing over the ground truth, may well feature substantially different positive rates, both for true and for false positives.

For an extremely simplified example, consider that flat obstacles are only relevant if they have a size of at least 6 grid elements—all smaller ones we circumvent while staying in lane or drive over them, letting them pass between our wheels. The invariant  $\iota$  for relevant flat obstacles thus is that they cover six grid elements at least, such that in the ground truth, either none or at least six grid elements feature a relevant flat obstacle. Then the formula

$$\Phi \equiv \bigvee_{x=1}^3 \bigvee_{y=1}^5 \text{flatobstacle}@ (x_{\text{ego}} + x, y) \tag{2}$$

expressing presence of a flat obstacle directly in front of the ego car is, for any  $k \in \{1, \dots, 6\}$ , W-equivalent wrt.  $\iota$  to

$$\Phi_k \equiv \sum_{x=1}^3 \sum_{y=1}^5 \text{flatobstacle}@ (x_{\text{ego}} + x, y) \geq k \ , \tag{3}$$

where we adopt the standard convention to identify false with 0 and true with 1 when taking the sum. Nevertheless, their positive rates vary obviously, as satisfying formula  $\Phi_k$  gets harder for larger  $k$ , with the easiest instance given by  $k = 1$  being logically equivalent to  $\Phi$ . Being harder to satisfy means that the positive rates get smaller. This applies both to the false positive rates—which is beneficial—and—detrimentally—to the true positive rates.

To provide an analytically solvable example of this effect, consider the stronger invariant  $\iota'$  that any relevant flat obstacle covers *exactly* six grid elements. Then all instance  $\Phi_k$  of (3) are still W-equivalent modulo  $\iota'$  and their false-positive rates and true-positive rates can, by straightforward reduction to binomial distributions, be analytically described by the formulae in Table 1, where  $tp$ ,  $fp$ ,  $tn$  and  $fn$  denote the true-positive rates, false-positive rates, true-negative rates, and false-negative rates, resp., of the atomic classifiers. For simplicity of the example, these are assumed to be uniform and stochastically independent across the occupancy grid. In practice, stochastic dependencies are obviously to be expected between adjacent grid elements. Extending the analysis to such spatial and furthermore to spatio-temporal dependencies is subject to future work. Qualitatively, the expected effects are, however, similar, as spatial dependencies affect both the ground truth (where geometric connectedness of objects permits additional inference) and the perception (where mispercepts may be correlated). These true-positive rates etc. are in practice determined empirically by

the usual statistical testing methods and thus known (up to a confidence) for the operational design domain.

**Table 1.** False-positive rates and true-positive rates for the guard formula  $\Phi_k$  from (3) in dependence of  $k \in \{1, \dots, 6\}$  for the cases of none to two obstacles of size 6 grid cells being present in the critical region.

obstacles present	$fp(\phi_k)$	$tp(\phi_k)$
0	$\sum_{i=k}^{15} \binom{15}{i} fp^i tn^{15-i}$	—
1	—	$\sum_{i=k}^{15} \sum_{j=0}^i \binom{6}{j} tp^j fn^{6-j} \binom{15-6}{i-j} fp^{i-j} tn^{15-6-(i-j)}$
2	—	$\sum_{i=k}^{15} \sum_{j=0}^i \binom{12}{j} tp^j fn^{12-j} \binom{15-12}{i-j} fp^{i-j} tn^{15-12-(i-j)}$

**Table 2.** False- and true-positive rates for the guard formula  $\Phi_k$  from (3) in dependence of  $k \in \{1, \dots, 6\}$  when the detection rates for the atomic classifiers are  $tp = 0.85$ ,  $fp = 0.2$ ,  $tn = 0.8$ , and  $fn = 0.15$ .

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	
0 o.	<b>0.965</b>	0.833	0.602	0.352	0.164	<b>0.061</b>	$fp(\phi_k)$
1 o.	$1 - 2 \cdot 10^{-6}$	$1 - 6 \cdot 10^{-4}$	0.999	0.992	0.954	0.834	$tp(\phi_k)$
2 o.	$1 - 6 \cdot 10^{-11}$	$1 - 5 \cdot 10^{-9}$	$1 - 1 \cdot 10^{-7}$	$1 - 3 \cdot 10^{-6}$	$1 - 4 \cdot 10^{-5}$	$1 - 4 \cdot 10^{-4}$	$tp(\phi_k)$

When these classification rates are  $tp = 0.85$ ,  $fp = 0.2$ ,  $tn = 0.8$ ,  $fn = 0.15$  then this results in the false and true positive rates for  $\Phi_k$  reported in Table 2. As expected, the positive rates decrease rapidly when  $k$  grows. For the safety-critical false-positive rate, we observe a  $\approx 16$ -fold reduction when going from the original formula  $\Phi$  or its logical equivalent  $\Phi_1$  to  $\Phi_6$ : In detail,  $\Phi = \Phi_1$  has  $\approx 5$ -fold  $fp$  of the atomic sensor, while  $\Phi_6$  has just  $\approx \frac{3}{10}fp$  of the atomic sensor (cf. entries in Table 2 marked in bold-print). The true-positive rate (i.e., performance) degrades also significantly to 83.4% in worst case (cf. second line of Table 2), but still remains at the level of an individual classifier, which features  $tp = 85\%$ .

Note that these reductions in positive rates come with no change in the ground-truth semantics, as W-equivalent formulae evaluate identically wrt. the ground truth according to Definition 1 and property (1). It also is worth noting that the above reductions do not degrade when going from this toy example to realistically fine and large grids—to the contrary, due to the binomials involved, these results obviously become exponentially better over a finer grid!

## 6 Synthesizing Optimal Representations of Guard Conditions

Having seen that the replacement of a safety-critical guard condition by a W-equivalent variant can significantly alter true-positive and false-positive rates, an obvious follow-up question is whether we can effectively rewrite such a guard condition into a more appropriate form. The exact problem we would like to solve is the following:

*Problem 1.* Given a guard condition  $\Phi$  and a societally accepted maximum risk  $\theta \in ]0, 1[$ , automatically synthesize a formula  $\Phi'$  that is W-equivalent to  $\Phi$  and satisfies the two requirements

**Safety:** the false positive rate of  $\Phi'$  remains below  $\theta$ , i.e.

$$fp(\Phi') \leq \theta \quad , \tag{4}$$

**Performance:** the true positive rate of  $\Phi'$  be maximal among the safe W-equivalent rewritings of  $\Phi$ , i.e.

$$tp(\Phi') = \max\{tp(\Phi'') \mid \Phi'' \text{ W-equivalent to } \Phi, tp(\Phi'') \leq \theta\} \quad . \tag{5}$$

To solve Problem 1, we first observe that it is a non-standard instance of don't-care optimization. For the sake of rendering the solution representable in a conference paper, we adopt an even smaller example than in Sect. 5. Assume three observational atoms  $A_1, A_2, A_3$  and a guard condition  $g(A_1, A_2, A_3)$  as given in Table 3 as a truth table. The ground-truth invariant is that never a single atom  $A_i$  can be true. Entries in the truth table satisfying  $\sum_{i_1}^3 A_i = 1$  are consequently “don't-cares”, as they cannot arise in reality.

**Table 3.** Truth table of a guard condition. Entries satisfying  $\sum_{i_1}^3 A_i = 1$  have an arbitrary “don't care” truth value (marked with \*), as they do not arise in ground truth due to a ground-truth invariant that no single atom  $A_i$  can be true in isolation.

$A_1$	$A_2$	$A_3$	$g(A_1, A_2, A_3)$
0	0	0	0
0	0	1	*
0	1	0	*
0	1	1	1
1	0	0	*
1	0	1	1
1	1	0	1
1	1	1	0

The don't-care entries in the truth table allow for setting them arbitrarily to 0 or 1. It would, however, in general be a bad idea to set all of them to 0, as this



nicely minimizes  $fp(g')$ , but also minimizes  $tp(g')$  unfortunately, thus optimizing safety at the price of minimizing performance. Vice versa, setting all don't-cares to 1 would maximize both  $fp(g')$  and  $tp(g')$ , thus optimizing performance while minimizing safety. In general, we need clever compromises, to be achieved by a differentiated setting of the individual don't-cares to 0 or 1.

To achieve such, we observe that both the true-positive rate and the false-positive rate of  $g'$  can be represented as affine terms over the don't-care assignments as follows: Denote by  $p_{abc}^{xyz}$  the likelihood of perceiving the ground truth  $(A_1, A_2, A_3) = (a, b, c)$  as  $(a_1, a_2, a_3) = (x, y, z)$ . Then

$$fp(g') = p_{000001x}^{1+} p_{000}^{010} x_2 + p_{000}^{011} + p_{000}^{100} x_3 + p_{000}^{101} + p_{000}^{110} + p_{111}^{001} x_1 + p_{111}^{010} x_2 + p_{111}^{011} + p_{111}^{100} x_3 + p_{111}^{101} + p_{111}^{110} \quad (6)$$

$$tp(g') = p_{011}^{001} x_1 + p_{011}^{010} x_2 + p_{011}^{011} + p_{011}^{100} x_3 + p_{011}^{101} + p_{011}^{110} + p_{101}^{001} x_1 + p_{101}^{010} x_2 + p_{101}^{011} + p_{101}^{100} x_3 + p_{101}^{101} + p_{101}^{110} + p_{110}^{001} x_1 + p_{110}^{010} x_2 + p_{110}^{011} + p_{110}^{100} x_3 + p_{110}^{101} + p_{110}^{110} \quad (7)$$

holds, where  $x_1, x_2,$  and  $x_3$  are the truth values assigned to the three don't-cares  $(A_1, A_2, A_3) = (0, 0, 1), (A_1, A_2, A_3) = (0, 1, 0),$  and  $(A_1, A_2, A_3) = (1, 0, 0),$  respectively.

An assignment  $x_1 \in \{0, 1\}, x_2 \in \{0, 1\},$  and  $x_3 \in \{0, 1\}$  to the don't-cares satisfying the above two requirements Safety and Performance can now mechanically be found by solving the following 0-1 integer-linear program:

$$\begin{aligned} & \text{Maximize } tp(g') \\ & \text{subject to } fp(g') \leq \theta \text{ and } x_1 \in \{0, 1\}, x_2 \in \{0, 1\}, x_3 \in \{0, 1\}, \end{aligned}$$

where  $fp(g')$  and  $tp(g')$  are the affine expressions from the right-hand sides of Eqs. (6) and (7). Note that the objective function  $tp(g')$  and the domain constraint  $fp(g') \leq \theta$  do directly encode the two requirements (5) and (4) from Problem 1. The above 0-1 integer-linear program can routinely be solved by any integer-linear programming (ILP) solver. The values reported for  $x_1$  to  $x_3$  in the optimal solution do then directly fill the don't-care entries in Table 3 if a solution exists. If no solution exists, then it is impossible to satisfy the societal acceptance threshold  $\Theta$  on false positives by an W-equivalent rewriting.

The above construction, however, does not scale. Being based on enumerating the entries of the truth table, its size is strictly exponential in the number of atoms  $A_i$  involved in the guard condition. The construction consequently becomes impractical when considerably more than 20 atoms are involved, which still constitutes a clearly minuscule occupancy grid. But luckily there is a lot of symmetry in formulae (6) and (7): one would for example expect that  $p_{000}^{001} = p_{000}^{010} = p_{000}^{100},$  as all of them involve flipping one bit from 0 to 1 in  $(A_1, A_2, A_3) = (0, 0, 0).$  Likewise,  $p_{111}^{001} = p_{111}^{010} = p_{111}^{100}$  and  $p_{100}^{011} = p_{010}^{101} = p_{001}^{110}$  and  $p_{101}^{001} = p_{101}^{100} = p_{110}^{010} = p_{110}^{100} = p_{011}^{010} = p_{011}^{001}$  etc. Grouping together equal factors and exploiting the symmetry in the solution space stemming from the fact that for subexpressions of the form  $ax_1 + ax_2 + ax_3$  only the sum  $x_1 + x_2 + x_3$  is

decisive while it is irrelevant which of  $x_1$ ,  $x_2$ , and  $x_3$  is set to 1, we can replace above ILP by the ILP

$$\begin{aligned} & \text{Maximize } tp(g'') \\ & \text{subject to } fp(g'') \leq \theta \text{ and } x \in \{0, \dots, 3\}, \end{aligned}$$

where  $fp(g')$  and  $tp(g')$  are defined by the affine integer expressions

$$\begin{aligned} fp(g'') &= 3p_{000}^{001}x + 3p_{111}^{001}x + 3p_{000}^{011} + 3p_{111}^{110} \\ &= \binom{3}{0}\binom{3}{1}\binom{0}{0}p_{000}^{001}x + \binom{3}{0}\binom{0}{0}\binom{3}{2}p_{111}^{001}x + \binom{3}{0}\binom{3}{2}\binom{0}{0}p_{000}^{011} + \binom{3}{0}\binom{0}{0}\binom{3}{1}p_{111}^{110} \quad (8) \end{aligned}$$

$$\begin{aligned} tp(g'') &= 6p_{011}^{001}x + 3p_{011}^{011} + 3p_{011}^{100}x + 6p_{011}^{101} \\ &= \binom{3}{2}\binom{1}{0}\binom{2}{1}p_{011}^{001}x + \binom{3}{2}\binom{1}{0}\binom{2}{0}p_{011}^{011} + \binom{3}{2}\binom{1}{1}\binom{2}{2}p_{011}^{100}x + \binom{3}{2}\binom{1}{1}\binom{2}{1}p_{011}^{101} \quad (9) \end{aligned}$$

Note that the binomial factors in front of the probabilities  $p_{abc}^{xyz}$  directly reflect the numbers of bits set in the ground truth  $(a, b, c)$  in the first binomial factor, the number of bits flipped from 0 to 1 among the 0 bits in the ground truth  $(a, b, c)$  to obtain the perception  $(x, y, z)$  in the second binomial factor, and the number of bits flipped from 1 to 0 in the third binomial factor. Therefore, the rather compact—and therefore as computationally inexpensive to formulate and solve—formulae (8) and (9) can be constructed directly by combinatorial reasoning without enumerating truth table entries. This process would in practice start from a Don't-Care-BDD representation of the guard condition  $g$  rather than a truth table like that from Table 3 used here for illustration. Implementation of this procedure is underway such that experimental results cannot yet be reported.

## 7 Related Work

Partially or fully autonomous cyber-physical systems, like highly automated vehicles, operate in an uncertain dynamic environment, which they have to perceive and understand in order to draw often safety-critical decisions. Such systems consequently tend to be learning-enabled—not necessarily end-to-end, but at least in central components relevant to perception and situation assessment. Their perception of the environment, i.e. the detection of properties about the dynamic environment, is enabled through inherently noisy sensors and subsequent machine-learnt classifiers. Especially in environmental perception based on computer vision, the uncertainties and the misperception rates induced by such machine-learnt classification algorithms remain substantial when visibility conditions are non-ideal [10]. The resultant misperception rates are orders of magnitude higher than the safety targets of, e.g., HAV [15].

Characterisation and control of perceptive uncertainty can be achieved at multiple stages of an architecture for automated cyber-physical systems, starting from the individual sensor level over fusion of multiple sensors and sensing modalities to control of uncertainty propagation through the decision and planning layers of a robotics architecture. The measures taken at these different

stages complement each other, with our contribution being located at the last of the aforementioned three stages.

Representations of uncertainty impacting the inferences underlying planning decisions have been investigated within the paradigm of probabilistic robotics [23], among others, particularly as applied to vehicle localization in urban environments [13, 14, 19], with localization being a special and historically more well-understood instance of the general problem of safe-guarding critical manoeuvre decisions. In these and related works such as [1, 17], the environment uncertainty is usually represented as probabilistic beliefs. Our constructive approach in this article, as well as its analytical counterpart in [6], complements such approaches by analysing and optimising the uncertainty propagation through the complex Boolean guard conditions usually employed for enabling and safeguarding safety-critical manoeuvres, be it as enabling preconditions of such manoeuvres or as side-conditions in safety shields [2] for AI components [12] or as fallback conditions in SIMPLEX-type fault-tolerant control architectures [16, 22].

It ought to be noted that approaches confining and controlling error propagation in the decision layer complement optimizations on the previous layers of sensing and sensor fusion, directly benefiting from, but also enhancing the impact of, enhancements at these layers. Various approaches to combining multiple classifiers can be found in the literature, e.g., see [7, 21] for an overview. The goal of such a combination is often to compensate for individual shortcomings in the performance by a better performance of the multitude of classifiers [21]. While in the pre-classification level the combination happens at the sensor or raw data level, the focus of this note is on fusion of classifiers at the post-classification level, as on-the-fly combinations of the decision of multiple atomic classifiers are considered. A major challenge for fusion on the decision level arises from the fact that the least genuine information about the object of observation is available at this level [11].

Our approach inherits the traditional setting of balancing between true positive rates, i.e. performance, and false positive rates, i.e. quantitative risk induced by misclassification, of adjustable classifiers, which often is pursued by analysing the empirical ROC (receiver operating characteristics) curve [3, 20]. Such analysis facilitates the optimisation of individual classifiers as well as their combination to obtain a better performance by a multitude of classifiers [21].

In a sense, our approach can be interpreted as a combination of classifiers too, namely one per grid element, albeit with given Boolean combination logic, which distinguishes it from the aforementioned fusion approaches, where the combination is to be designed based on the mutual performance figures of the multiple classifiers. The only degree of freedom we have thus is to modify the combinatory logic such that it maintains the same logical function on all ground-truth instances (see Definition 1, yet still enhances resilience to misclassification. The key to modification of the combinatory logic is the identification of don't cares in its truth tables due to ground-truth invariants. The exploitation of such ground-truth invariants itself is not original; it has already been investigated in visual classification tasks to enhance the accuracy of scene segmentation tasks,

e.g. in [5], where the natural vertical layering (e.g. that vehicles stand on the tarmac and not vice versa) of visual scenes is used as invariant.

It is also interesting to note that the paradigm of occupancy grids and hence the approach suggested carries well beyond automotive manoeuvre planning and similar real-time path planning problems requiring distance between objects. Similar approaches have been used to cover safety problems of contact robotics, like the safe use of robot manipulators in collaborative scenarios (cobots) as well as physical human-robot interaction. To generate collision-free trajectories, models of human motion ought to be integrated for better estimation, and path planning needs to be optimised for execution speed and safety. Both the environment and the human operator are represented via occupancy grids in [24] and exploited for planning that adapts to different human operators or their hand positions. If direct contact between the manipulator and the human is relevant to the task, the contact forces must also be limited. [18] have integrated biomechanical injury information into the robot controller for this purpose, with different force thresholds applying to different body parts, thus requiring occupancy maps for representing the positions of human body parts. Symmetrically, occupancy maps of the full robot arm, rather than just the tool centre point, are employed in [9] to limit joint torques in whole arm manipulations with their multiple contact points, inducing spatially distributed safety constraints.

## 8 Conclusion

Reliable guarantees for the safety of autonomous systems are a prerequisite for their societal acceptance. The quest for such guarantees cannot easily be served, at least not at the appropriate quantitative safety levels for safety-critical systems like autonomous vehicles at usual speed of traffic, due to the relatively high misperception rates of technical perception chains. When mapping an autonomous vehicle’s vicinity, their error rates tend to—currently as well as for the foreseeable future—be orders of magnitude higher than the pertinent safety targets for autonomous operation. Any compositional analysis inducing a weakest-link principle, i.e. suggesting that the overall system’s reliability would be bounded by perception performance, is thus bound to generate grossly insufficient evidence of system safety.

We consequently are in need of analytical methods or even constructive means to ensure that our systems actually are “safer than perception”. More precisely, this requires a rigorous assessment of the likelihood that a safety-critical manoeuvre is erroneously adopted, and this assessment has to provide much tighter bounds for such erratic manoeuvre adoption than for any misperception. In a companion report [6], we have addressed the analytical problem of rigorously proving a quantitative reliability figure for the evaluation of a complex Boolean guard condition that is safeguarding a safety-critical manoeuvre, in the sense that its evaluation to true is a necessary prerequisite for adoption of the manoeuvre. We have been able to show that for complex guard conditions, the rates of critical misevaluations can be proven to be significantly lower than misperception rates concerning atomic percepts.

In this article, we drove this analysis further and gave it a constructive tweak in that we asked for solving an optimization problem that deals with finding that rewriting of a given guard condition that is most resilient to misperceptions while retaining the semantics of the original guard condition. Concretely, we asked for finding a formula rewriting that

1. is equivalent to the original guard over all ground-truth instances,
2. reaches the societally given safety target in that its rate of false positives, i.e. of suggesting the critical manoeuvre when it is undue, remains below the societal acceptance threshold, and
3. optimizes performance in that it yields maximal true positive rate, i.e. actually enables the critical manoeuvre when it is due.

We showed that this problem has a constructive solution by reducing it to integer-linear programming, thereby automatically synthesizing a formula rewriting satisfying the aforementioned three requirements.

Practical implementation of the algorithm and of modifications leveraging symmetries as well as symbolic reasoning for enhancing scalability are underway as a student project. Future work will deal with a spatio-temporal rather than just propositional analysis, refining our analysis by topological and geometrical properties induced by the grid structure and temporal correlations induced by the dynamics. Obviously, a slight misplacement of a detected object both is more likely to happen and more unlikely to change a guard's perceived truth value than a large displacement. Similar arguments apply in the temporal domain, where true positives and true negatives tend to have a higher temporal persistence than false positives or false negatives, respectively. While these properties have extensively been studied for effectively filtering atomic mispercepts, their impact on the evaluation of complex spatio-temporal conditions serving as guards remains a subject of ongoing research [4].

Another interesting question concerns transfer of the results to other reference architectures than occupancy grids, especially to object list representations of the environment. These do classify objects, locate them at an environmental position, and add a bounding box characterizing their geometric extent. The latter is often inexact, giving rise to quality measures like the (relative) area of intersection over union for the perception. Related perception problems are detecting an object part (e.g., a car backlight) instead of the full object (the car) due to visibility problems like occlusion, then attributing a subcomponent label and an accordingly smaller bounding box. Both phenomena fit our analysis in principle, as again a reasonable guard condition will talk about a non-trivial Boolean combination of more than percept (now in terms of class, relative position and size of the bounding box), and rephrasing it to increase resilience is as relevant. Object lists add convexity properties and shape constraints due to the bounding boxes, but can otherwise be seen as mostly isomorphic to 2.5-dimensional occupancy grids, locating the bounding boxes in the perspective plane rather than the street plane. It will again be an issue of future research to fill in the details.

**Acknowledgements.** The research reported herein has been supported by the State of Lower Saxony within the Zukunftslabor Mobilität as well as by Deutsche Forschungsgemeinschaft under grant no. DFG FR 2715/5-1 “Konfliktresolution und kausale Inferenz mittels integrierter sozio-technischer Modellbildung”. It furthermore benefited from technical discussions with Krzysztof R. Apt, Werner Damm, Willem Hagemann, Hardi Hungar, and Paul Kröger, as well as from the comments of the anonymous reviewers. Their support is gratefully acknowledged.

## References

1. Baig, Q., Perrollaz, M., Laugier, C.: A robust motion detection technique for dynamic environment monitoring: a framework for grid-based monitoring of the dynamic environment. *IEEE Robot. Automat. Mag.* **21**(1), 40–48 (2014)
2. Bloem, R., Könighofer, B., Könighofer, R., Wang, C.: Shield synthesis: - runtime enforcement for reactive systems. In: Baier, C., Tinelli, C. (eds.) *Tools and Algorithms for the Construction and Analysis of Systems - 21st International Conference, TACAS 2015, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2015, London, UK, April 11-18, 2015. Proceedings.* *Lecture Notes in Computer Science*, vol. 9035, pp. 533–548. Springer, Cham (2015). [https://doi.org/10.1007/978-3-662-46681-0\\_51](https://doi.org/10.1007/978-3-662-46681-0_51)
3. Fawcett, T.: An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**(8), 861–874 (2006)
4. Finkbeiner, B., Fränzle, M., Kohn, F., Kröger, P.: A truly robust signal temporal logic: monitoring safety properties of interacting cyber-physical systems under uncertain observation. *Algorithms* **15**(4), 126 (2022). <https://doi.org/10.3390/A15040126>
5. Fouopi, P.P.: *Holistische Modellierung und Interpretation von Szenen und Situationen basierend auf symbolischen, probabilistischen und subsymbolischen Modellen.* Ph.D. Thesis, University of Oldenburg, Germany (2019). <http://oops.uni-oldenburg.de/4601>
6. Fränzle, M., Hagemann, W., Damm, W., Rakow, A., Swaminathan, M.: Safer than perception: Assuring confidence in safety-critical decisions of automated vehicles. In: Haxthausen, A.E., Huang, W., Roggenbach, M. (eds.) *Applicable Formal Methods for Safe Industrial Products - Essays Dedicated to Jan Peleska on the Occasion of His 65th Birthday.* *Lecture Notes in Computer Science*, vol. 14165, pp. 180–201. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-40132-9\\_12](https://doi.org/10.1007/978-3-031-40132-9_12)
7. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybernet. Part C (Appl. Rev.)* **42**(4), 463–484 (2011)
8. Geirhos, R., Janssen, D.H.J., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, F.A.: Comparing deep neural networks against humans: object recognition when the signal gets weaker. *CoRR* (2017). [arxiv:1706.06969](https://arxiv.org/abs/1706.06969)
9. Gliesche, P., Kowalski, C., Pflingsthor, M., Hein, A.: Geometry-based two-contact inverse kinematic solution for whole arm manipulation. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pp. 8269–8274. IEEE (2021). <https://doi.org/10.1109/IROS51168.2021.9636664>

10. Huber, L.S., Geirhos, R., Wichmann, F.A.: The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *CoRR* (2022). <https://doi.org/10.48550/ARXIV.2205.10144>
11. Khreich, W., Granger, E., Miri, A., Sabourin, R.: Iterative Boolean combination of classifiers in the ROC space: an application to anomaly detection with HMMs. *Pattern Recogn.* **43**(8), 2732–2752 (2010). <https://doi.org/10.1016/j.patcog.2010.03.006>
12. Könighofer, B., Rudolf, J., Palmisano, A., Tappler, M., Bloem, R.: Online shielding for reinforcement learning. *Innov. Syst. Softw. Eng.* **19**(4), 379–394 (2023). <https://doi.org/10.1007/S11334-022-00480-4>
13. Levinson, J., Montemerlo, M., Thrun, S.: Map-based precision vehicle localization in urban environments. In: *Proceedings of Robotics: Science and Systems*. Atlanta, GA, USA (2007). <https://doi.org/10.15607/RSS.2007.III.016>
14. Levinson, J., Thrun, S.: Robust vehicle localization in urban environments using probabilistic maps. In: *IEEE International Conference on Robotics and Automation*, pp. 4372–4378 (2010)
15. Maurer, M., Gerdes, J.C., Lenz, B., Winner, H.: *Autonomous Driving: Technical, 1st edn. Legal and Social Aspects*. Springer Publishing Company, Incorporated (2016)
16. Mitsch, S., Platzer, A.: Modelplex: verified runtime validation of verified cyber-physical system models. *Formal Methods Syst. Des.* **49**(1–2), 33–74 (2016). <https://doi.org/10.1007/S10703-016-0241-Z>
17. Moras, J., Cherfaoui, V., Bonnifait, P.: Moving objects detection by conflict analysis in evidential grids. In: *IEEE Intelligent Vehicles Symposium (IV 2011)*, pp. 1120–1125 (2011)
18. Palleschi, A., Hamad, M., Abdolshah, S., Garabini, M., Haddadin, S., Pallottino, L.: Fast and safe trajectory planning: solving the cobot performance/safety trade-off in human-robot shared environments. *IEEE Robot. Autom. Lett.* **6**(3), 5445–5452 (2021). <https://doi.org/10.1109/LRA.2021.3076968>
19. Petrovskaya, A., Thrun, S.: Model based vehicle detection and tracking for autonomous urban driving. *Auton. Robots* **26**(2–3), 123–139 (2009)
20. Powers, D.: Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2**(1), 37–63 (2011)
21. Sagi, O., Rokach, L.: Ensemble learning: a survey. *WIREs Data Min. Knowl. Discov.* **8**(4), e1249 (2018). <https://doi.org/10.1002/widm.1249>
22. Seto, D., Krogh, B., Sha, L., Chutinan, A.: The simplex architecture for safe online control system upgrades. In: *Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No.98CH36207)*, vol. 6, pp. 3504–3508 (1998). <https://doi.org/10.1109/ACC.1998.703255>
23. Thrun, S., Burgard, W., Fox, D.: *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press (2005)
24. Zanchettin, A.M., Messeri, C., Cristantielli, D., Rocco, P.: Trajectory optimisation in collaborative robotics based on simulations and genetic algorithms. *Int. J. Intell. Robotics Appl.* **6**(4), 707–723 (2022). <https://doi.org/10.1007/S41315-022-00240-4>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.







# Towards ML-Integration and Training Patterns for AI-Enabled Systems

Sven Peldszus<sup>1</sup>(✉) , Henriette Knopp<sup>1</sup> , Yorick Sens<sup>1</sup> ,  
and Thorsten Berger<sup>1,2</sup> 

<sup>1</sup> Ruhr University Bochum, Bochum, Germany  
sven.peldszus@rub.de

<sup>2</sup> Chalmers | University of Gothenburg, Gothenburg, Sweden

**Abstract.** Machine learning (ML) has improved dramatically over the last decade. ML models have become a fundamental part of intelligent software systems, many of which are safety-critical. Since ML models have complex lifecycles, they require dedicated methods and tools, such as pipeline automation or experiment management. Unfortunately, the current state of the art is *model-centric*, disregarding the challenges of engineering systems with multiple ML models that need to interact to realize complex functionality. Consider, for instance, robotics or autonomous driving systems, where perception architectures can easily incorporate more than 30 ML models. Developing such multi-ML model systems requires architectures that can integrate and chain ML components. Maintaining and evolving them requires tackling the combinatorial explosion when re-training ML components, often exploring different (hyper-)parameters, features, training algorithms, or other ML artifacts. Addressing these problems requires *systems-centric* methods and tools. In this work, we discuss characteristics of multi-ML-model systems and challenges of engineering them. Inspired by such systems in the autonomous driving domain, our focus is on experiment-management tooling, which supports tracking and reasoning about the training process for ML models. Our analysis reveals their concepts, but also their limitations when engineering multi-ML-model systems, especially due to their model-centric focus. We discuss possible integration patterns and ML training to facilitate the effective and efficient development, maintenance, and evolution of multi-ML-model systems. Furthermore, we describe real-world multi-ML-model systems, providing early results from identifying and analyzing open-source systems from GitHub.

**Keywords:** ML-Enabled Systems · ML Asset Management · ML Training · Maintenance · Evolution

## 1 Introduction

Many recent advances in artificial intelligence (AI) allow building software systems for tasks that seemed impossible before. Especially machine learning

© The Author(s) 2025

B. Steffen (Ed.): AISoLA 2023, LNCS 14129, pp. 434–452, 2025.

[https://doi.org/10.1007/978-3-031-73741-1\\_26](https://doi.org/10.1007/978-3-031-73741-1_26)

(ML) has been the driving force behind these advances. Consider the field of autonomous driving, which benefited greatly from these advances. Modern autonomous driving systems use AI in a variety of ways. For instance, AI perceives traffic lights, detects lanes, helps avoid obstacles on the road, and predicts traffic. Modern software systems can have many AI components, up to a whopping number of 28 ML models in the autonomous driving system Baidu Apollo [4, 27]. Since fully autonomous driving is still far away according to experts [12, 22], software systems incorporating even more ML models in different topologies and architectures can be expected.

Integrating one or many models in software systems poses new software-engineering challenges, since ML models differ significantly from traditional software artifacts. Models are developed (i.e., trained) highly iteratively based on data—the respective data-science methods and tools are entirely *model-centric* and *data-driven*. Traditional software-engineering methods and tools have not been designed for these circumstances [18]. This challenges managing different models in ML-enabled software systems, especially integrating the models and quality-assuring them. Training the models in such multi-ML-model systems gives rise to combinatorial explosion due to interactions [2], also since the models themselves are trained highly iteratively in so-called experiments. As such, managing different models [16] requires managing different experiments.

The integration of multiple ML models poses novel challenges, particularly for safety-critical systems. Consider again autonomous driving. Spectacular attacks—such as hackers being able to steer a Tesla into oncoming traffic using an attack as simple as putting some white stickers on the road [19]—illustrate the limitations in properly applying traditional software-engineering technologies in real, safety-critical systems [8]. Such systems typically use various sensors to perceive their environment and process the captured data using ML models to make safety-critical decisions. Each model used in such a system is trained for a specific task, such as detecting the road markings in a camera image. The performance of the final system however, does not depend on a single model, but on multiple, potentially interacting models and non-ML parts, each of which is potentially the weakest link in the chain. Furthermore, in practice, models are likely to be exposed to inputs that were not considered during training, raising the risk of incidents. To this end, it is unclear how we can systematically improve the overall performance and robustness of such AI-enabled systems [30].

Addressing these challenges requires novel *system-centric* methods that go beyond the *model-centric* methods from data science, which focus on individual ML models in isolation. We need best practices on how to integrate multiple ML models and how to decide which combination of models to improve (e.g., by retraining) to improve the overall system performance. Such methods need to support managing the necessary ML experiments, integrate them into a *system-centric* workflow, and consider the benefits to the whole system when training certain combinations of ML models.

As a first step towards such a *system-centric* method, we investigate the interaction between multiple ML models in AI-enabled systems. We discuss their

characteristics and derive possible training patterns for *system-centric* training of ML models. These training patterns allow not only to consider the overall system performance as an essential aspect while training ML models, but also to consider the interactions between ML models while training them. We discuss these training patterns, their potential benefits, but also their limitations in detail using the example of an autonomous racing car as described above. To enable the transfer to any AI-enabled system, we performed an exploratory survey on open-source software systems from GitHub, in which we identified different integration patterns for ML models. For each of these integration patterns, we relate the identified training patterns to them. We derive open challenges from our discussions and conclude with a research agenda.

## 2 Background

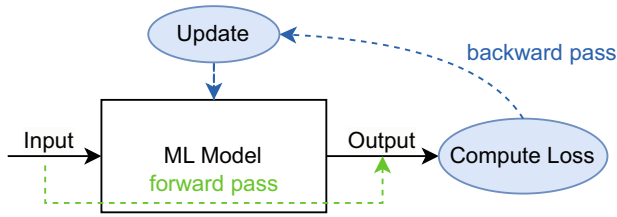
Before discussing how we can evolve from a *model-centric* to a *system-centric* development process, we provide an overview of current practices in ML model development and challenges discussed in the field of AI engineering.

To illustrate these challenges our running example is a Formula Student autonomous race car [1,33]. It is similar to full-featured autonomous driving systems, such as Baidu Apollo. These race cars have to drive on a track marked by yellow and blue cones, so they do not have to deal with all the complexities of real traffic. However, such a race car still uses various sensors, such as cameras and LiDAR to perceive the track. These sensor information process is using multiple ML models [10]. Typically, the outputs of these models are fused into a world representation that the car can use to make driving decisions. Currently, each of the individual models is trained in isolation for a specific task, such as detecting the cones in the camera images.

### 2.1 Machine Learning

Recall that, machine Learning (ML) approximates a function from data, the so-called ML model. There are several subcategories of ML that differ in how said pattern or function is learned. For example, in supervised learning, a model is trained from a dataset that consists of input data and a set of labels on that data from which the ML model is approximated. The ML model can then make predictions about the labels of unseen input data [21].

ML models are approximated iteratively from a set of training data in a process called ML model training that is realized in so-called experiments [21]. As shown in Fig. 1, a training algorithm is implemented as a training loop that consists of the following steps. First, in a so-called *forward pass*, the current configuration of the ML model is used to compute the output (i.e., the labels) on a batch of the training dataset. Then, the deviation of the computed output from the expected output (i.e., labels in the dataset), called the loss, is computed. The loss is then used to update the model in a *backward pass*, with the goal of iteratively optimizing the ML model towards the function to be approximated.



**Fig. 1.** High-level view on the steps of training an ML model

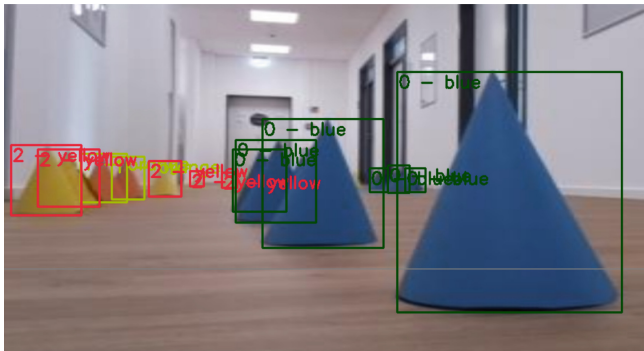
This training is repeated until an acceptable loss is achieved, i.e., a sufficiently low rate of false positive and false negative outputs.

In practice, models are trained in so-called experiments. An experiment consists of multiple runs where every run optimizes a model with differing sets of parameters. The resulting models from each of the runs are compared until a final model is selected to be used. Managing and comparing different experiment runs is a significant development effort and requires tools, called experimentation management tools, to adequately support it [14, 16, 17].

## 2.2 ML Frameworks

It is common to rely on ML frameworks to create and train ML models [31]. For this purpose, these frameworks provide a variety of support, ranging from implementing the most basic algorithms to providing fully functional ML models.

While frameworks such as PyTorch [28] and TensorFlow provide developers with the means to train ML models by providing implementations of common algorithms used in ML, the developer still needs to do a fair amount of development to create a working ML model. In contrast, Keras [9] is a high-level framework based on TensorFlow that allows easy implementation of ML models without too much background knowledge.



**Fig. 2.** Cones detected by an image-based perception model

Other frameworks provide pre-trained models and code to run them for specific tasks. For example, in autonomous driving, an essential task is to detect objects in the environment, such as the blue and yellow cones in the Formula Student example. One framework that focuses on efficiently implementing the task of object detection in images is YOLO [29]. Figure 2 shows an example of the output of such a YOLOv5 ML model for an indoor track oriented on the Formula Student. The objects—i.e., the cones—are detected in the form of bounding boxes that define the area in the image occupied by the detected objects. Usually, pre-trained models such as YOLO have to be trained on the concrete data and labels, e.g., the bounding boxes, they are supposed to detect. In this case, we trained YOLO version 5 (YOLOv5) [20, 29] to detect the cones.

### 2.3 Development of ML-Enabled Systems

The development of traditional software and training of ML models is very different. In traditional software system development functionality is programmed by hand and features are linked closely to the code. When training ML models their code is visible while the internals of ML models which realize the functionality are merely black boxes. ML model training is an iterative process that goes through different stages, including requirements analysis, data-oriented work, model-oriented work, and the DevOps phase. Several tools support the different phases of ML model development, including asset and experiment management [16, 17]. However, integrating the training of ML models with system development has been identified as a challenge among researchers [3, 5]. Especially the integration of a ML model into a systems environment is often difficult [24].

In particular, the quality of the training data is essential in ML development [24]. If the training data is not properly prepared, the training algorithm may not learn the correct pattern and make incorrect predictions on new, unseen data. Especially in safety-critical domains such as autonomous driving, where the system is frequently exposed to new inputs and new data, the development of robust ML models is critical [19].

### 2.4 ML Techniques Using Multi-ML-Model Systems

Multiple ML models in a system have been considered in related work, but with a focus on the robustness of a single task implemented using ML. The two techniques that come to mind are *Ensemble Learning* and *Federated Learning*.

*Ensemble Learning* is a technique based on multiple ML models to obtain classifications that are more robust than those of single ML models [34]. The basic strategy is to train multiple classifiers, possibly using different configurations or frameworks, on the same problem and determine a final output using the outputs of all classifiers. While ensemble learning also uses multiple ML models, the main difference with multi-ML-model systems as we consider them is that multi-ML-model systems use multiple ML models to solve different and sometimes independent problems.

*Federated Learning* is a technique for decentralized training of ML models. Users or clients (i.e., mobile devices) train the ML model themselves with their own data. The individually trained models are then aggregated into a new model, on which training is then continued in a distributed fashion. Particularly relevant in the context of our running example of a race car is research on the benefits of federated learning for ML-based object recognition, especially how Yolov5 (cf. Sect. 2.2) can be implemented or even improved [11].

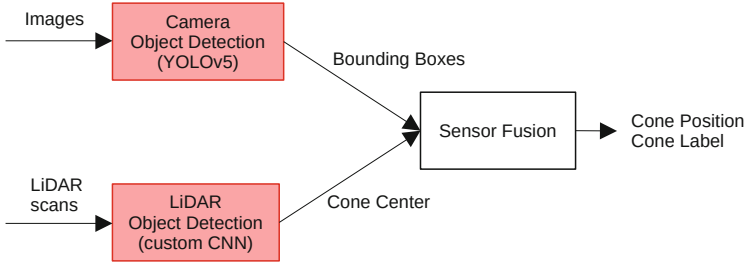
### 3 Integration of ML Models in ML-Enabled Systems

Multi-ML-model systems are ML-enabled systems that incorporate more than one ML model. While ML-enabled systems are on the rise in a variety of domains, multi-ML-model systems are often found in safety-critical domains such as autonomous driving. Systems that integrate ML models pose new challenges for developers. Bosch et al. discuss challenges related to processes and workflows and outline a research agenda that includes the architecture, development and processes involved in AI engineering and also highlight domain-specific challenges [5]. Nazir et al. present insights on challenges related to the design and discuss different architectural patterns in engineering ML-enabled systems [25]. Lastly, Apel et al. discuss the issue of unexpected interactions of ML models when they are being reassembled after an initial decomposition of a complex task and the resulting integration testing [2]. Altogether, there is a need to obtain a better understanding of multi-ML-model systems.

#### 3.1 Examples from the Autonomous Driving Domain

Baidu Apollo [4] is a real-world example of a multi-ML-model system from the domain of autonomous driving. A case study showed that the Apollo perception and prediction pipeline consists of 28 ML models [27]. For presentation purposes, however, we use a less complex running example: an autonomous Formula Student race car. Like the full-fledged autonomous driving system Baidu Apollo, the race car drives autonomously in its environment, but the latter is less complex, and therefore, requires fewer sensors and less ML models to process the sensor readings. In the remainder, we mainly focus on the ML-based perception architecture of our race car example.

Similar to Baidu Apollo, the autonomous race car uses a camera and a LiDAR to perceive its environment, i.e., the race track. For this task, it uses the YOLO model introduced above to detect cones that mark the race track in the camera images. Figure 3 shows a sketch of the race car’s perception architecture. Sensor data flows from left to right, with the system receiving two inputs, *Images* and *LiDAR scans*, which are processed independently of each other using ML models. Later, the data is fused in a sensor fusion, as shown further to the right, resulting in the detected *cone positions* and corresponding *cone labels*, i.e., their colors, as the pipeline output.



**Fig. 3.** ML-components of the running example

To train and manage the individual ML models, we used the open-source experiment-management tool MLFlow, which provides a Python API to log various metrics, which measure the models performance (i.e., recall or loss), during the training process. The results are stored and can be visualized using the MLFlow web interface. MLFlow is model-centric and does not support managing multiple ML models in systems, however. As we will show, this tool helps to keep track of individual training iterations of single models, it does not help to improve the overall system.

### 3.2 Integration Patterns of Multiple ML Models

The perception architecture of the race car illustrates a multi-ML model-system that integrates two ML models that process different inputs by merging their outputs. However, it is not representative of all AI-enabled systems employing multiple ML models, and it is unclear in what other ways ML models can be integrated. In the remainder, we will refer to the ways in which multiple ML models are integrated as *integration patterns*.

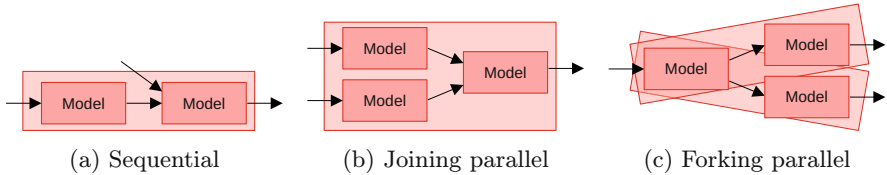
To obtain a broader overview of how ML models are integrated in practice, we searched on GitHub for repositories that use one of the three major ML libraries TensorFlow, PyTorch, and Scikit-learn [31]. We obtained an initial sample of 763,820 repositories from which we extracted 359 applications by a series of automated filtering steps followed by a manual review of the remaining repositories. Roughly 40% of these systems use more than one ML library. A list of these repositories and the data collected about them is available on GitHub<sup>1</sup>.

To identify the systems that use multiple ML models, we analyzed how many trained models are stored in binary files in the repositories. However, only few repositories (17%) contained such model files. We learned that in practice, ML models are often stored externally and have to be manually downloaded from platforms such as Google Drive. However, we also observed systems that automatically download the ML models from a server. In most cases, however, custom code for training ML models is provided with the system instead of already

<sup>1</sup> <https://github.com/isselab/ML-Systems-Datasets>.

trained models. Of those projects that store their model files in the repository, a median of 2 and an average of 5 ML models can be found.

Although most repositories include training code for their models, the training of individual models appears to be independent of each other. Therefore, we focused our further analysis on the concrete implementations of the ML models and their integration. Source code belonging to model implementations was found in almost all subject applications, although the number of model implementations varied greatly.



**Fig. 4.** ML integration patterns observed in multi-ML model systems

We manually examined a sample of these systems, focusing on their integration of multiple ML models. We identified the three underlying integration patterns for interacting ML models shown in Fig. 4.

**Sequential:** This integration pattern describes an integration of ML models in which one model processes the output of another one, potentially, including additional information (Fig. 4a).

**Joining parallel:** Following this integration pattern, two or more models are deployed in parallel, and their outputs are merged either by a traditional component or another ML model (Fig. 4b). Our running example represents a system that implements this integration pattern.

**Forking parallel:** As shown in Fig. 4c, we found parallel models, that process the output of another ML model. While this third pattern could be seen as two overlapping sequential patterns, it comes with particular challenges when considering training on a system-level scope as outlined above.

An example of the *forking parallel* deployment pattern is a system that detects traffic lights and road signs for autonomous driving, which we found on GitHub<sup>2</sup>. The detection is implemented in two steps and involves three different ML libraries. In a first step, a YOLOv5 model is used to detect bounding boxes around potential objects of interest. Thereafter, the forking happens and different models are applied in parallel for classifying these objects of interest. A Scikit-learn model is used to detect traffic lights and their state, and a TensorFlow model created using Keras is used to classify road signs.

<sup>2</sup> <https://github.com/JeffWang0325/Image-Identification-for-Self-Driving-Cars>.



## 4 Training Patterns for Multi-ML Model Systems

In this section, we discuss the interaction of ML models in multi-ML-model systems and possible training patterns for improving the training of such multiple, interacting ML models. When considering the training of AI-enabled systems consisting of multiple ML models that are integrated with each other, such as the perception pipeline of our race car example, one can try to improve the training efficiency and effectiveness by not considering the different models independently. The idea is to make the training process more effective and more focused on the actual output of the system. This mainly concerns how the training process is structured and implemented, and what datasets need to be managed.

### 4.1 ML Model Interactions

To begin with, we will discuss interactions between ML models in our running example to identify potentials for improving their training. Usually, the individual models of our race car example would be trained in isolation and only thereafter be considered in combination in traditional software engineering tasks such as system design or quality assurance. If each of the included ML models is trained individually, as it is currently the case, we need two training datasets that consist of different data. First, we need images from the camera that are labeled with cones, as described in detail above. Here, each label is specified by five data points, first the color of the cone, which is encoded as the *class* of the bounding box, and the bounding box itself, which is described by four numeric values ( $x\_center$ ,  $y\_center$ , *width*, and *height*). Second, we need LiDAR scans, again manually annotated with labels representing cones. In our case, each LiDAR scan is a 1D vector of distance measurements, and the cones are encoded by labeling each distance measurement as belonging to a cone or not. In total, cones have to be labeled in two datasets with six different values for each cone.

Not only is it a substantial effort to create the two required datasets, but by training in this way we are trying to optimize each of the models to its best performance in isolation. However, it is unclear how this affects the final output of the system. There are two factors to consider.

**System performance:** Although each model may perform optimally in isolation, we have no guarantee that the final output of the system will meet the performance requirements of the system, i.e., in terms of precision and recall of the cone detection needed by the race car for autonomous driving. Therefore, we need additional test cases to evaluate whether the performance requirements are met. However, if these are not met, it is not clear which model needs to be improved to meet the performance requirements. On the other hand, for optimal system performance, it may not be necessary for each model to perform optimally in isolation. For example, in our case, we have precise positions for each cone in the LiDAR scans, and as long as the sensor fusion operates properly, it may not be necessary to optimize the sizes and positions of the bounding boxes for maximum accuracy. Also, false positives may not be a significant issue in the outlined application, since cones

not detected by either ML model will not be included in the final output, while false positives can simply be discarded in the sensor fusion. Therefore, it would be beneficial to prioritize recall over precision in training. While such considerations are obvious in our simplified example, they may not be in more complex systems.

**Training cost:** Building up on the considerations above, if we do not need optimal performance of the individual ML models, we could even reduce the effort needed for training and only fit the ML models until they reach the required performance. Unfortunately, however, we are not aware of any systematic way to determine this threshold upfront and one has to fall back to systematic and extensive integration tests.

According to the above considerations, it would be beneficial to train both models not against their individual outputs, but against the overall output of the system. First, it allows for a better integration of the two models as described above. Second, it could even help to reduce the maintenance effort of the required datasets. While in the isolated training of the models against their individual outputs we need two datasets described by six values together, we can reduce this to a single dataset if we would train against the final system output. Furthermore, in our running example, we could even describe the cones in this dataset by only three values, the  $x$  and  $y$  positions of the detected cones, and *labels* representing their colors. These three values are even easy to measure.

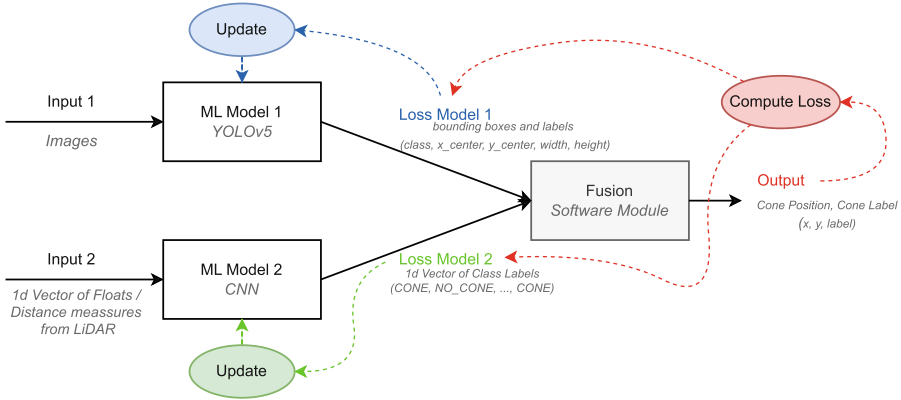
## 4.2 Training Patterns

In the following, we introduce training patterns that allow for a system centric training of the ML models. Based on the structure of the perception pipeline of our running example, we have identified four different training patterns for training multi-ML-model systems. We briefly describe each training pattern and what data is used where in the training process of our running example. Thereafter, we discuss the application to the other identified integration patterns.

Figure 5 shows in detail how the ML models of our running example are integrated and interact during training, and which data is relevant in which step of these interactions. The dashed lines represent the backward pass of the training, where data is propagated back through the system. First, according to the right part of the backward pass highlighted in red, the system output is first used to compute the loss of the pipeline, which we call the *end loss*. Then, the blue part at the top of the figure and the green part at the bottom show how the loss is traced back to each of the ML models, which are then updated. We can compute a deviation from expected outputs and predicted outputs, similar to a typical training algorithm. The figure also shows how the end loss is propagated back through the system. It consists of a combination of the error in predicting the position and color of the cones.

Our four identified training patterns are:

**Individual Model Training:** This training pattern represents the state of the art, where each ML model is optimized independently of its integration into



**Fig. 5.** Interactions among the ML models in our running example. (blue lines: update of ML model 1/Yolov5 using loss 1; green lines: update of ML model 2/CNN using loss 2; red lines: back propagation of the *end loss*) (Color figure online)

the system. It is useful for ML models that are used as stand-alone models, such as the model used in the YOLO library. However, when the model is integrated into the system, it may not be necessary to fine-tune every aspect. For example, in the case of the YOLOv5 model integrated into our race car, the training optimizes the position of the bounding boxes as well as the labeling. In this case, the exact position of the bounding box on the image data does not affect the overall system performance because the exact cone position is determined based on the more accurate LiDAR measurements.

**Partial End-Loss Training:** This training pattern puts a single ML model into the context of the system during training. It differs from the state of the art by considering its integration into a system. The idea is that instead of updating the ML model based on its direct output, as we did in *Individual End Loss Training*, we will follow the data through the system, compute the end loss and use this to update the model. This means that any software components used for pre- or post-processing must be integrated into the training algorithm. In our example, this would mean that the data would be fed into the system and processed by both ML models. After sensor fusion, the final result is used to update the LiDAR model, which is indicated by the backward pass consisting of the red and green markers in Fig. 5. Alternatively, we could also update the Yolov5 model in this way.

**Simultaneous End-Loss Training:** This training pattern extends the previous training pattern to include all ML models in the system. This means that all ML models are trained in the context of the system. The training infrastructure is the same as in the previous training pattern. The difference is that now all ML models are updated using this end loss. In each training iteration, data is passed through the system and processed by the ML models. The data is then further processed and the final result is used to update

the models using the backward pass through the system. In Fig. 5, we update Yolov5 by following the red and the blue arrows and we update the CNN by following the red and the green arrows.

**Alternating End-Loss Training:** This training pattern is a modification of the previous training pattern. Instead of updating all ML models in each training iteration, we will update only one ML model. This variation reduces the computation time in each steps and allows for better traceability between the improvement in the end loss and model update. The idea is that in some cases, one of the ML models will perform quite well, while the other will not. In our example, the end loss includes different features, such as the position and color of the cones. At least in our example, each of these features can be traced back to a different ML model. If the position of the cone is predicted incorrectly but the color is correct, it may make more sense to update the LiDAR ML model instead of the Yolov5 model. For reference, in Fig. 5, we would follow the red arrows and then either the blue or green arrows, alternating the two updates.

Until now, we only considered our running example that implements a joining parallel integration of multiple ML models. In our investigation of multi-ML model systems at GitHub however, we also identified sequential and forking parallel ML model integration patterns. Like in our running example, the overall system performance of these systems depends on the interaction of the different ML models. For example, in the system that classifies traffic lights and road signs using models deployed following the forking parallel integration pattern, a traffic light or road sign that is missed by the first model, cannot be recovered by the subsequent models executed in parallel, but they can discard potential objects that are not traffic lights or road signs. Therefore, to improve safety, the first model should be trained mainly targeting a high recall or as proposed by us immediately concerning the overall system output. While these considerations are obvious for this example, they may be challenging for different application scenarios or more complicated tasks.

Since the sequential integration pattern can be seen as a subset of the discussed joining parallel integration pattern, the training patterns outlined for the joining parallel pattern can be applied without changes to train multiple ML models that are integrated sequentially. However, when ML models are integrated following the forking parallel integration pattern, the output of the first model is processed by multiple subsequent models, and therefore, the training patterns outlined above cannot be applied immediately to this integration pattern. In principal, the alternating end loss training could be applied and the training process iterates through all subsequent models. While this might work sufficiently well for an initial training, it would imply a huge effort if only one of the subsequent models does not perform well. Furthermore, since the first model would go through much more training iterations than the subsequent models, this may will lead to over fitting. Alternatively, we would need for example a loss function that can merge the outputs of all subsequent models for training the preceding model. However, this would still not solve the issue of

only optimizing one of the subsequent models while not impacting the other ones. To overcome this issue, one could fall back to performing multiple experiments again for the contained sequential patterns and, thereafter, use federated learning [35] for merging the multiple, individually trained, models of the forking parallel pattern into one that fits the needs of all subsequent models.

## 5 Research Directions

The detailed challenges outlined above give rise to the following research directions. We focus on the actual implementation of multi-ML systems intended to operate in safety-critical domains, such as autonomous driving.

### 5.1 Research Challenges

We now summarize the challenges we identified and concrete research directions.

**Challenge 1:** *Integrating multiple ML models into an AI-enabled system.*

In our preliminary study of multi-ML-model systems on GitHub, we have identified underlying patterns for integrating multiple ML models into an AI-enabled system, but it is unclear whether the identified list is complete. Furthermore, best practices on how to actually realize these integration patterns as well as what are the advantages and disadvantages of concrete integration patterns are not systematically captured, yet. To effectively implement multi-ML model systems, developers need such best practices for integrating the multiple models with each other.

**Challenge 2:** *Robustness of AI-enabled systems against adversarial input.*

Many AI-enabled systems operate in critical domains where it is essential to ensure their robustness against adversarial inputs to avoid dangerous behaviors such as steering into oncoming traffic [19]. While securing a system against adversaries is a well-known challenge [23, 26, 32], it becomes even more challenging due to the black box behavior of ML models [13]. Individual patterns, such as redundancy, have proven effective in improving robustness to adversarial input. However, we lack a systematic overview of such patterns as well as general techniques and processes for improving the robustness of AI-enabled systems. In particular, it is unclear how optimized training and data management help improve the robustness of AI-enabled systems.

**Challenge 3:** *Managing experiments in multi-ML model systems.*

As discussed above, following the current state of practice in developing AI-enabled systems, one creates one training dataset per ML model and manages its experiments individually. However, the integration of multiple ML models leads to many interactions that need to be considered in the experiments. In practice, we have many training datasets that may be related, such as LiDAR data and camera images of the cones in our running example. It is unclear if and how this relationship needs to be reflected in the experiments. Furthermore, these interactions lead to significant scalability issues, since in the worst case we have to consider the cross product of all experiment runs for all ML models.

**Challenge 4:** *Training ML models on a system-wide scope.*

Although ML models can be an important part of an AI-enabled system, they are not an end in themselves. Instead, requirements always target concrete functionalities of a system, and ML models are only one technical solution to realize such functionalities. As discussed above, in many systems it is necessary to integrate multiple ML models, but training is still done on the basis of individual models against unclear internal requirements. Therefore, training should be shifted from a model-centric to a system-centric scope. Despite the outline of concrete training patterns that aim at this goal, how to actually achieve it remains an open challenge. This includes how each ML model influences the overall system performance, and how each ML model should be updated from a system perspective, particularly how to calculate a meaningful loss for each model.

**Challenge 5:** *Optimization of the system performance.*

The development of complex software systems is usually not a one-time task, but involves frequent iterations of implementing new functionality, testing and validating that functionality, and then improving it. Once ML models are integrated into such a software system, improving the functionality provided by the system may involve improving one or more of the models. However, pinpointing specific models that need to be retrained, especially in terms of what is missing, is currently a labor-intensive, manual task that requires tracing system-level tests back to the corresponding models. Providing support for this task remains an open challenge. Furthermore, whenever we retrain a model, it is currently unclear whether the entire system will benefit from this improvement, or whether it remains a local improvement that does not affect the entire system at all. Overall, the challenge is twofold. First, we need to identify architectures that support the overall system benefiting from model improvements, and second, we need to estimate this benefit in advance. The *Alternating End Loss Training* proposed above could help address this challenge by considering models in combination and scheduling retraining passes according to which improvement would benefit the entire system. However, its feasibility and effectiveness remain to be demonstrated.

**Challenge 6:** *Testing of multi-ML-model systems.*

Testing of multi-ML-model systems is a frequently mentioned challenge. Because ML-models are trained deductively from examples, instead of inductively from fixed specifications, they are inherently difficult to test. Related works have addressed this issue by proposing ML-specific testing methods [6,7], but these are not yet widely adopted. More work is needed in that direction, especially to evaluate the feasibility of these methods in actual software systems. Integrating multiple ML models in a software system further complicates the issue of testing, as additional integration tests are needed. In sequentially applied models, errors are propagated through the entire pipeline, which makes it hard to trace their origin.

## 5.2 Research Agenda

For addressing the challenges and the corresponding research opportunities summarized above, building up on the early ideas presented in this paper, we identified two concrete research directions that we will target next.

**Integration Patterns in Practice.** To obtain a better overview of the state of the art regarding the development of ML-enabled software, we are going to extend the mining study already discussed. Currently, the dataset used still contains a number of relatively small and immature software systems, many of which use only a single ML model. This could potentially be improved, by adapting the filtering criteria, such as filtering them by their popularity at GitHub. Furthermore, we will broaden the study by taking projects implemented in other programming languages into account.

Furthermore, although the integration patterns observed in our preliminary study are promising, we have to go deeper into detail. By extracting integration patterns from additional systems we could potentially find additional patterns. Also, by now we only have a high-level overview but have extract details on how to actually implement these patterns. We must assess how often each integration pattern is used in practice and in which context it is used. Based on this information, we can derive advantages and disadvantages of each pattern, especially in terms of performance and robustness as considered in Challenges 1, 2, and 5. Additionally, we can analyze the subject systems for the usage of further techniques to increase adversarial robustness, such as safeguards.

To get insights on how interactions among models are considered during their training and testing, we will study training code and datasets in depth. To this end, we can build upon the work of Idowu et al. [15] and their techniques to identify code that belong to different steps of the training process. This way, we aim to identify whether experiments are conducted together with system development or in separate repositories and how testing at a system-wide scope takes place (Challenge 3). The mapping of API calls to stages of ML development provides a basis for characterizing the relationship between these stages. We can, for example, determine if experiments are conducted on the entire pipeline or only the single models are refined in isolation, or if the output of one model is used as training data for the next one. Furthermore we can analyze the systems for testing practices, for example, the test coverage of ML-and non-ML code or the prevalence of model and data validation. We aim for identifying practices to address the issue of testing pipelines with multiple ML models (Challenge 4). This provides an empirical basis for addressing Challenge 6.

Building on top of this extended mining study, we could potentially conduct a survey with developers regarding the most pressing challenges they face and the reasoning behind the design choices we uncovered in the mining study. This could also help to put the discovered architectures and practices into context, separating good from bad practices.

**Training (Multiple) ML Models on a System-Centric Scope.** In this paper, we have already identified opportunities and discussed initial ideas for shifting the training of multiple ML models from a model-centric to a system-centric scope. In particular, the integrated training of multiple ML models following the outlined training patterns would allow to consider a system-centric scope also in the training of multi-ML systems, as targeted by Challenge 4. In this way, the testing associated with the training process of ML models would be further shifted towards a system scope, thus also addressing Challenge 6. Finally, as outlined above, system-centric training could help to reduce the effort needed to create and maintain training datasets, as well as to reduce the number of individual experiments needed to train the ML models of multi-ML systems (Challenge 3). To this end, we also want to assess how this idea is perceived by developers and how it improves their systems and training processes.

Building on the ideas presented in this paper, the next step is a prototypical implementation of the training patterns to evaluate the advantages and disadvantages of the proposed training patterns, particularly also concerning the robustness of the AI-enabled system to adversarial inputs (Challenge 1) and the impact on improving the system performance (Challenge 5). Here, the first challenge is that the ML model frameworks are not designed to be integrated into a system as we envision it. In other words, for our pipeline to work properly and for our ML models to be able to be trained, we need to merge the typical ML training loop with the execution of our system.

Next, we need to properly define the metrics and losses, and come up with a baseline against which to compare the effectiveness of the training patterns. We propose to use the *Individual Model Training* as a baseline. The losses need to be carefully defined and we will need to track what data each ML model needs and outputs as well as what information might be lost due to further processing. The question is how much of the data is left in the final result. One risk to consider is that the end loss may also not accurately reflect the performance of an individual model. Consider the following example. The end loss is the sum of the deviation in cone position and cone color. The Yolov5 model made an incorrect prediction, while the LiDAR model is already performing quite well and the cone position is predicted close to the expected output. If we update both models with the same loss, we are penalizing the LiDAR model for a good prediction. This is something that *Alternating End Loss Training* could fix and might be a peril in *Simultaneous End Loss Training*.

After properly defining the metrics and losses, we need to evaluate the benefits of this system-centric training. Assuming that we can practically train ML models that have the same performance and robustness as those trained using traditional training, the aspects we need to evaluate and compare are mainly the number of iterations used for training and the computational time. Another factor is the amount of data required. In other words, how much is the dataset reduced when training the pipeline as a whole. We believe that the reduction is significant, but we do not have anything tangible. Another aspect is the gen-



eralization of our proposed training patterns to other integration patterns that differ from our example.

## 6 Conclusion

We presented an overview of how multiple ML models are integrated in AI-enabled systems and how we can shift their training from a model-centric to a system-centric scope. Our long-term goal is to gain a better understanding of how we can rethink the development and training process for multi-ML model systems and enable software engineers without extensive background in data science to develop high-quality multi-ML systems. To this end, we want to be able to make recommendations on how to simplify the training process based on the context of the system. As a first step, we identified integration patterns of ML models in AI-enabled systems, as well as training patterns for shifting the training of ML models from a model-centric to a system-centric technique. In doing so, we identified open challenges in systematically integrating multiple ML models into a robust AI-enabled system and ensuring its performance and robustness that need to be addressed by future research. To this end, in our upcoming research, we will evaluate the identified training patterns and ML model interactions in real-world multi-ML systems. In addition, we will further evaluate the underlying integration patterns in ML-enabled software systems. We will analyze a large set of ML-enabled systems and present our results in a separate mining study.

## References

1. FSG Competition Handbook 2024. Tech. rep., Formula Student Germany (2023)
2. Apel, S., Kästner, C., Kang, E.: Feature interactions on steroids: on the composition of ML models. *IEEE Softw.* **39**(3), 120–124 (2022)
3. Arpteg, A., Brinne, B., Crnkovic-Friis, L., Bosch, J.: Software engineering challenges of deep learning. In: *Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 50–59 (2018). <https://doi.org/10.1109/SEAA.2018.00018>
4. Baidu: Apollo Auto: An Open Autonomous Driving Platform (2024). <https://github.com/ApolloAuto>
5. Bosch, J., Olsson, H.H., Crnkovic, I.: Artificial Intelligence Paradigms for Smart Cyber-Physical Systems, chap. *Engineering AI Systems: A Research Agenda*, pp. 1–19. IGI Global (2020). <https://doi.org/10.4018/978-1-7998-5101-1.ch001>
6. Braiek, H.B., Khomh, F.: On testing machine learning programs. *J. Syst. Softw. (JSS)* **164**, 110542 (2020). <https://doi.org/10.1016/j.jss.2020.110542>
7. Cheng, C.H., Huang, C.H., Yasuoka, H.: Quantitative projection coverage for testing ML-enabled autonomous systems. In: *International Symposium on Automated Technology for Verification and Analysis (ATVA)*, pp. 126–142 (2018)
8. Chernikova, A., Oprea, A., Nita-Rotaru, C., Kim, B.: Are self-driving cars secure? Evasion attacks against deep neural networks for steering angle prediction. In: *IEEE Security and Privacy Workshops (SPW)*, pp. 132–137 (2019)
9. Chollet, F., et al.: Keras (2015). <https://github.com/fchollet/keras>

10. Gong, H., Feng, Y., Chen, T., Li, Z., Li, Y.: Fast and accurate: the perception system of a formula student driverless car. In: International Conference on Robotics, Control and Automation (ICRCA), pp. 45–49 (2022). <https://doi.org/10.1109/ICRCA55033.2022.9828892>
11. Hegiste, V., Legler, T., Ruskowski, M.: Federated Ensemble YOLOv5 – A Better Generalized Object Detection Algorithm (2023). arXiv (arXiv:2306.17829). <https://doi.org/10.48550/ARXIV.2306.17829>
12. Heineke, K., Kampshoff, P., Mkrtchyan, A., Shao, E.: Self-Driving Car Technology: When Will the Robots Hit the Road? Tech. rep., McKinsey & Company (2017). <https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/self-driving-car-technology-when-will-the-robots-hit-the-road>
13. Hu, Y., et al.: Artificial intelligence security: threats and countermeasures. ACM Comput. Surv. **55**(2), 20:1–20:36 (2023). <https://doi.org/10.1145/3487890>
14. Idowu, S., Osman, O., Strueber, D., Berger, T.: Machine learning experiment management tools: a mixed-methods empirical study. Empir. Softw. Eng. (EMSE) **29**, 1 (2024)
15. Idowu, S., Sens, Y., Berger, T., Krüger, J., Vierhauser, M.: A large-scale study of ML-related python projects. In: Symposium On Applied Computing (SAC) (2024). <https://api.semanticscholar.org/CorpusID:267375897>
16. Idowu, S., Strueber, D., Berger, T.: Asset management in machine learning: state-of-research and state-of-practice. ACM Comput. Surv. **55**, 1 (2022)
17. Idowu, S., Strueber, D., Berger, T.: EMMM: a unified meta-model for tracking machine learning experiments. In: Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 48–55 (2022). <https://doi.org/10.1109/SEAA56994.2022.00016>
18. Idowu, S., Strüber, D., Berger, T.: Asset management in machine learning: a survey. In: International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), pp. 51–60 (2021). <https://doi.org/10.1109/ICSE-SEIP52600.2021.00014>
19. Jing, P., et al.: Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations. In: USENIX Security Symposium, pp. 3237–3254 (2021). <https://www.usenix.org/conference/usenixsecurity21/presentation/jing>
20. Jocher, G.: YOLOv5 by Ultralytics (2020). <https://doi.org/10.5281/zenodo.3908559>. <https://github.com/ultralytics/yolov5>
21. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015). <https://doi.org/10.1126/science.aaa8415>
22. Liu, L., et al.: Computing systems for autonomous driving: state of the art and challenges. IEEE Internet Things J. **8**(8), 6469–6486 (2021). <https://doi.org/10.1109/JIOT.2020.3043716>
23. McGraw, G.: Software security. IEEE Secur. Privacy **2**(2), 80–83 (2004). <https://doi.org/10.1109/MSECP.2004.1281254>
24. Nahar, N., Zhang, H., Lewis, G., Zhou, S., Kästner, C.: A meta-summary of challenges in building products with ml components - collecting experiences from 4758+ practitioners. In: International Conference on AI Engineering – Software Engineering for AI (CAIN), pp. 171–183 (2023). <https://doi.org/10.1109/CAIN58948.2023.00034>
25. Nazir, R., Bucaioni, A., Pelliccione, P.: Architecting ML-enabled systems: challenges, best practices, and design decisions. J. Syst. Softw. (JSS) **207**, 111860 (2024). <https://doi.org/10.1016/J.JSS.2023.111860>

26. Peldszus, S.: Security Compliance in Model-driven Development of Software Systems in Presence of Long-term Evolution and Variants. Springer, Cham (2022). <https://doi.org/10.1007/978-3-658-37665-9>
27. Peng, Z., Yang, J., Chen, T.H., Ma, L.: A first look at the integration of machine learning models in complex autonomous driving systems: a case study on Apollo. In: Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), pp. 1240–1250 (2020)
28. PyTorch: Open Source Machine Learning Framework (2023). <https://pytorch.org>
29. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: unified, real-time object detection. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016). <https://doi.org/10.1109/CVPR.2016.91>
30. Shafique, M., Naseer, M., Theocharides, T., Kyrkou, C., Mutlu, O., Orosa, L., Choi, J.: Robust machine learning systems: challenges, current trends, perspectives, and the road ahead. *IEEE Des. Test* **37**(2), 30–57 (2020). <https://doi.org/10.1109/MDAT.2020.2971217>
31. Stancin, I., Jovic, A.: An overview and comparison of free Python libraries for data mining and big data analysis. In: International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 977–982 (2019). <https://doi.org/10.23919/MIPRO.2019.8757088>
32. Tuma, K., Peldszus, S., Strüber, D., Scandariato, R., Jürjens, J.: Checking Security Compliance between Models and Code. *Softw. Syst. Model. (SoSyM)* **22**(1), 273–296 (2023). <https://doi.org/10.1007/S10270-022-00991-5>
33. Valls, M.I., Hendrikx, H.F., Reijgwart, V.J., Meier, F.V., Sa, I., Dubé, R., Gawel, A., Bürki, M., Siegwart, R.: Design of an autonomous racecar: perception, state estimation and system integration. In: International Conference on Robotics and Automation (ICRA), pp. 2048–2055 (2018). <https://doi.org/10.1109/ICRA.2018.8462829>
34. Webb, G.I., Zheng, Z.: Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. *IEEE Trans. Knowl. Data Eng. (TKDE)* **16**(8), 980–991 (2004). <https://doi.org/10.1109/TKDE.2004.29>
35. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A Survey on Federated Learning. *Knowl. Based Syst.* **216**, 106775 (2021). <https://doi.org/10.1016/J.KNOSYS.2021.106775>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.



The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# **Safety Verification of DNNs**



# The Reachability Problem for Neural-Network Control Systems

Christian Schilling<sup>(✉)</sup> and Martin Zimmermann<sup>ID</sup>

Aalborg University, Aalborg, Denmark  
{christianms,mzi}@cs.aau.dk

**Abstract.** A control system consists of a plant component and a controller which periodically computes a control input for the plant. We consider systems where the controller is implemented by a feedforward neural network with ReLU activations. The reachability problem asks, given a set of initial states, whether a set of target states can be reached. We show that this problem is undecidable even for trivial plants and fixed-depth neural networks with three inputs and outputs. We also show that the problem becomes semi-decidable when the plant as well as the input and target sets are given by automata over infinite words.

## 1 Introduction

Cyber-physical systems consist of digital (cyber) and physical components. A common instance of this paradigm is a control system, consisting of a physical *plant* and a *controller* whose purpose is to steer the plant to a desired behavior [7]. Control theory studies the automatic synthesis of such controllers, which, for nonlinear systems, is a difficult task. Machine learning has long been successfully applied to tackle this task, where the learned controller was first represented by a (shallow) feedforward neural network [19] and more recently by a *deep neural network* (DNN) [15]. We call a control system with a DNN controller a *neural-network control system* (NNCS). Due to their black-box nature, DNNs have raised concerns about their correctness and safety, in particular in terms of the worst-case behavior [28]. However, as they are deployed in safety-critical applications, proving machine-learned NNCS correct is of utmost importance, and considerable resources have been invested into their verification [16, 17].

In this paper, we are concerned with the fundamental problem of safety for NNCS: given a set of initial states and a set of bad states of the plant, does the controller prevent the plant to reach a bad state when started in an initial state? Note that the failure of safety is captured by a reachability property: does there exist an initial state from which a bad state is reachable? Thus, in the following, we study the reachability problem for NNCS.

Recall that an NNCS is a combination of a DNN (the controller) and a plant. It is known that the reachability problem is already undecidable for sufficiently complex plants, even without any controller [13]. So the question becomes: is there a simple but expressive class of plants for which the reachability problem

is tractable? Inspired by similar results for recurrent neural networks [10, 27], we show in Sect. 3 that the answer is negative: the reachability problem is undecidable even for trivial plants. Intuitively, a DNN can simulate one computational step of a two-counter machine. Thus, a recurrent neural network can simulate a two-counter machine. As a DNN controlling a plant is essentially recurrent (as it bases its control decisions on the current state of the plant), undecidability follows.

On the positive side, we show in Sect. 4 that the reachability problem is at least semi-decidable for plants whose behavior can be captured by automata over infinite words: Sälzer et al. showed that the behavior of DNNs can be captured by such automata [23]. Hence, relying on standard automata-theoretic constructions, the composition of a DNN and an automata-definable plant can also be captured by automata. The class of automata-definable plants includes, for instance, plants that are described by multi-mode linear maps. Such maps are able to express, for example, the dynamics of adaptive cruise controls [14].

## 1.1 Related Work

Reachability in NNCS is generally challenging. Existing approaches typically combine techniques developed for dynamical systems (the plant) [2] and neural networks [16]. Tools such as CORA [1, 12], JuliaReach [5, 25], and NNV [18] compete in the ARCH-COMP friendly competition, and we refer to the report [17] for typical examples of NNCS.

Undecidability of questions about unbounded computations with piecewise-linear (PWL) functions is long known, e.g., periodicity in iterated 2D maps [22] or reachability for linear hybrid automata [9]. Similar results have been shown for DNNs. Siegelmann and Sontag showed undecidability for unbounded computations in DNNs with activations given by a PWL approximation of the sigmoid function (which effectively is the ReLU function truncated at 1) [27]. Later, Hyotyniemi showed an encoding of two-counter machines in recurrent neural networks (RNNs) with ReLU activations [10]. While an RNN can conceptually be seen as the special case of an NNCS without a plant, the formalism differs. We thus consider our encoding of two-counter machines in NNCS of independent (yet mainly pedagogical) value. Cabessa showed an encoding of two-counter machines in a variant of RNNs with conditional weights called *spike-timing dependent plasticity* [6]. Recently, we also showed an encoding of two-counter machines in decision-tree control systems [26], where the DNN is replaced by a decision tree with simple conditions  $x \leq c$  for some variable  $x$  and constant  $c$ .

Katz et al. studied the problem of reachability in a ReLU DNN without iteration and showed that, given polyhedral (i.e., described by linear constraints) input and output sets, the reachability problem is NP-complete [11]. Sälzer and Lange later fixed some issues in the proof, mainly related to the effective representation of real numbers [24].

Sälzer et al. recently presented an encoding of a DNN in a weak Büchi automaton [23]. We build on this encoding for the analysis of semi-decidability.

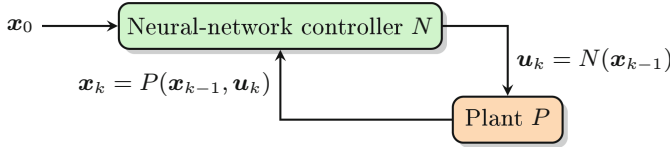


Fig. 1. Neural-network control system

## 2 Preliminaries

We start by formally introducing the type of DNN under study.

**Definition 1 (Deep neural network).** A neuron is a function  $\nu: \mathbb{R}^m \rightarrow \mathbb{R}$  with  $\nu(\mathbf{x}) = \sigma(\sum_{i=1}^m c_i x_i + b)$ , where  $m$  is the input dimension, the  $c_i \in \mathbb{Q}$  are the weights,  $b \in \mathbb{Q}$  is the bias, and  $\sigma: \mathbb{R} \rightarrow \mathbb{R}$  is the activation function of  $\nu$ , which is either the identity function or the rectified linear unit (ReLU)  $y \mapsto \max\{y, 0\}$ .

A layer is a sequence of neurons  $(\nu_1, \dots, \nu_n)$ , all of the same input dimension  $m$ , computing the function  $\ell: \mathbb{R}^m \rightarrow \mathbb{R}^n$  given by  $\ell(\mathbf{x}) = (\nu_1(\mathbf{x}), \dots, \nu_n(\mathbf{x}))$ . The dimensions  $m$  and  $n$  are the input resp. output dimension of the layer.

A deep neural network (DNN)  $N$  is a sequence of layers  $(\ell_1, \dots, \ell_k)$  such that the output dimension of  $\ell_i$  is the input dimension of  $\ell_{i+1}$  for all  $i = 1, \dots, k-1$ . The last layer is the output layer and all other layers are called hidden layers. If  $m$  is the input dimension of  $\ell_1$  and  $n$  is the output dimension of  $\ell_k$ , then the DNN computes the function  $N: \mathbb{R}^m \rightarrow \mathbb{R}^n$  defined as

$$N(\mathbf{x}) = \ell_k(\ell_{k-1}(\dots \ell_1(\mathbf{x}) \dots)).$$

Next, we define control systems as depicted in Fig. 1. The system consists of a plant and a controller (here: a DNN) and acts in iterations: first, the controller computes a control input  $\mathbf{u}$  for the plant based on the current state  $\mathbf{x}$  of the plant. Next, from its inputs  $\mathbf{x}$  and  $\mathbf{u}$ , the plant computes a new state. Then the process repeats. For the plant, we restrict ourselves to discrete time, i.e., we are only interested in its output and not its intermediate states. For now, we also abstract from the concrete type of plant and just view it as a general function.

**Definition 2 (Neural-network control system).** A neural-network control system (NNCS) is a tuple  $(P, N)$  with a plant  $P: \mathbb{R}^{d+c} \rightarrow \mathbb{R}^d$  and a controller given by a DNN  $N: \mathbb{R}^d \rightarrow \mathbb{R}^c$ , i.e.,  $d$  is the dimension of the states of  $P$  and  $c$  is the dimension of the control vectors computed by  $N$ .

The semantics of an NNCS are given as a sequence of states  $\mathbf{x}_k$  and control inputs  $\mathbf{u}_k$ , induced by some initial state  $\mathbf{x}_0 \in \mathbb{R}^d$  via

$$\begin{aligned} \mathbf{u}_k &= N(\mathbf{x}_{k-1}) \\ \mathbf{x}_k &= P(\mathbf{x}_{k-1}, \mathbf{u}_k) \end{aligned}$$

We introduce a shorthand to express one iteration of the control loop in Fig. 1, i.e., the composition of the DNN followed by the plant, to compute  $\mathbf{x}_k$  from  $\mathbf{x}_{k-1}$ :

$$C_{P,N}(\mathbf{x}_{k-1}) = P(\mathbf{x}_{k-1}, N(\mathbf{x}_{k-1}))$$

We will focus on sets of states represented by linear constraints. Given  $a \in \mathbb{Q}^n$ ,  $b \in \mathbb{Q}$ , the set  $H_{a,b} = \{\mathbf{x} \in \mathbb{R}^n \mid \langle a, \mathbf{x} \rangle \leq b\}$  is a linear constraint, where “ $\langle \cdot, \cdot \rangle$ ” denotes the scalar product. A polyhedron is a finite intersection of linear constraints. Let  $\mathcal{P}(n)$  denote the set of all polyhedra in  $n$  dimensions.

We are now ready to define the reachability problem for NNCS.

*Problem 1.* (Reachability problem for NNCS). Given a DNN  $N: \mathbb{R}^d \rightarrow \mathbb{R}^c$ , a plant  $P: \mathbb{R}^{d+c} \rightarrow \mathbb{R}^d$ , a polyhedron  $X_0 \in \mathcal{P}(d)$  of initial states, and a polyhedron  $\varphi \in \mathcal{P}(d)$  of target states, does there exist an initial state  $\mathbf{x}_0 \in X_0$  and a  $k \in \mathbb{N}$  such that  $(C_{P,N})^k(\mathbf{x}_0) \in \varphi$ ?

### 3 Undecidability

In this section, we prove that the NNCS reachability problem is undecidable. The proof is by a reduction from the halting problem for two-counter machines.

Formally, a two-counter machine  $\mathcal{M}$  is a sequence

$$(0 : \mathbf{I}_0)(1 : \mathbf{I}_1) \cdots (k-2 : \mathbf{I}_{k-2})(k-1 : \text{STOP}),$$

where the first element of a pair  $(\ell : \mathbf{I}_\ell)$  is the line number and  $\mathbf{I}_\ell$  for  $0 \leq \ell < k-1$  is an instruction of the form

- $\text{INC}(i)$  with  $i \in \{0, 1\}$ ,
- $\text{DEC}(i)$  with  $i \in \{0, 1\}$ , or
- $\text{JZ}(i, \ell')$  with  $i \in \{0, 1\}$  and  $\ell' \in \{0, \dots, k-1\}$ .

A configuration of  $\mathcal{M}$  is of the form  $(\ell, c_0, c_1)$  with  $\ell \in \{0, \dots, k-1\}$  (the current value of the program counter) and  $c_0, c_1 \in \mathbb{N}$  (the current contents of the two counters). The initial configuration is  $(0, 0, 0)$  and the unique successor configuration of a configuration  $(\ell, c_0, c_1)$  is defined as follows:

- If  $\mathbf{I}_\ell = \text{INC}(i)$ , then the successor configuration is  $(\ell+1, c'_0, c'_1)$  with  $c'_i = c_i + 1$  and  $c'_{1-i} = c_{1-i}$ .
- If  $\mathbf{I}_\ell = \text{DEC}(i)$ , then the successor configuration is  $(\ell+1, c'_0, c'_1)$  with  $c'_i = \max\{c_i - 1, 0\}$  and  $c'_{1-i} = c_{1-i}$ .
- If  $\mathbf{I}_\ell = \text{JZ}(i, \ell')$  and  $c_i = 0$ , then the successor configuration is  $(\ell', c_0, c_1)$ .
- If  $\mathbf{I}_\ell = \text{JZ}(i, \ell')$  and  $c_i > 0$ , then the successor configuration is  $(\ell+1, c_0, c_1)$ .
- If  $\mathbf{I}_\ell = \text{STOP}$ , then  $(\ell, c_0, c_1)$  has no successor configuration.

The unique run of  $\mathcal{M}$  (starting in the initial configuration) is defined as the maximal sequence  $\gamma_0 \gamma_1 \gamma_2 \cdots$  of configurations  $\gamma_j \in \mathbb{N}^3$  where  $\gamma_0$  is the initial configuration and  $\gamma_{j+1}$  is the successor configuration of  $\gamma_j$ , if  $\gamma_j$  has a successor configuration. This run is either finite (line  $k-1$  is reached) or infinite (line  $k-1$  is never reached). In the former case, we say that  $\mathcal{M}$  terminates. The halting problem for two-counter machines asks, given a two-counter machine  $\mathcal{M}$ , whether  $\mathcal{M}$  terminates when started in the initial configuration.



**Proposition 1** ([20]). *The halting problem for two-counter machines is undecidable.*

In the following, we show that the halting problem for two-counter machines can be reduced to the reachability problem for NNCS by simulating the semantics of a two-counter machine by a NNCS.

**Theorem 1.** *The reachability problem for NNCS is undecidable.*

*Proof.* Fix some two-counter machine  $\mathcal{M}$  with  $k$  instructions. We show how to construct a gadget for each instruction of  $\mathcal{M}$  (except for the STOP instruction), which we then combine into a DNN simulating one configuration update of  $\mathcal{M}$ . Thus, the reachability problem for NNCS (which involves the iterated application of the DNN) then allows to simulate the full run of  $\mathcal{M}$ .

Formally, the DNN implements a function from  $\mathbb{R}^3 \rightarrow \mathbb{R}^3$  with the following property: If the three inputs encode a non-stopping configuration of the two-counter machine, then the three outputs encode the successor configuration. Note that, since the weights and biases of the DNN we construct are integral, the outputs given integral inputs are also integral. In the following, we often implicitly assume that inputs are integral when we explain the intuition behind our construction.

Our construction of the DNN fits into the common architecture [8] that all hidden neurons use ReLU activations and all output neurons use identity activations. We let the plant component simply turn the control input into the new state ( $P(\mathbf{x}, \mathbf{u}) = \mathbf{u}$ ), as the DNN already simulates  $\mathcal{M}$ .

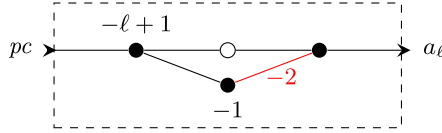
In some more detail, for every instruction  $(\ell; \mathbb{I}_\ell)$  of  $\mathcal{M}$ , we construct one gadget simulating this instruction. All these gadgets will be executed in parallel in one iteration of the DNN, but only one of them (determined by the current value of the program counter) will actually perform a computation. The other gadgets just compute the identity function for each of their inputs. Thus, in the end we need to subtract  $(k - 2) \cdot v$  from each output  $v$ .

All gadgets have inputs named  $pc$  (representing the current value of the program counter), and  $c_0$  and  $c_1$  (representing the current counter values), as well as three outputs named  $pc'$  (representing the value of the program counter of the successor configuration), and  $c'_0$  and  $c'_1$  (representing the counter values of the successor configuration). To simplify our construction, we use an additional gadget that conceptually checks whether the value of the program counter is equal to some fixed line number  $\ell$ . This gadget is shown in Fig. 2. The output  $a_\ell$  of this gadget (which has only one input  $pc$ ) satisfies

$$a_\ell = \begin{cases} 1 & pc = \ell, \\ 0 & pc \neq \ell. \end{cases}$$

The outputs  $a_\ell$  of these auxiliary gadgets (we have one for each line number  $\ell$ ) will be fed into the other gadgets simulating the instructions.

Next, we describe the instruction gadgets, where we restrict ourselves to the counter with index zero; the counter with index one is treated in the analogous



**Fig. 2.** Auxiliary gadget for instruction  $\ell$ . Here and in all later illustrations of DNNs, dots denote neurons, where filled dots use ReLU activations and empty dots can use either identity or ReLU activations (the choice is irrelevant since the value before the activation is nonnegative). Sometimes, as in this case, the empty dots are only present for a fully-connected architecture. Edge colors only serve the visual association with the weights. We omit weight 1 and bias 0 as well as connections with weight 0

way. Figure 3 shows these gadgets together with the possible outputs. It is easy to verify that each gadget performs the corresponding computation whenever the input  $pc$  is equal to  $\ell$ , and the identity function otherwise. Let us stress that each gadget we construct depends both on the line number and the instruction.

The final layout of the gadgets is shown in Fig. 4. Essentially, each auxiliary gadget is wired to the corresponding instruction gadget, and at the end we need to subtract the inputs  $k - 2$  times as described above. Note that there is no gadget for the STOP instruction (instruction  $k - 1$  in  $\mathcal{M}$ ). When  $pc$  is equal to  $k - 1$ , then the DNN computes the identity function: First, all  $a_\ell$  are equal to 0; Hence, each of the  $k - 1$  instruction gadgets  $I_\ell$  computes the identity function; after the subtraction, we are indeed left with the identity.

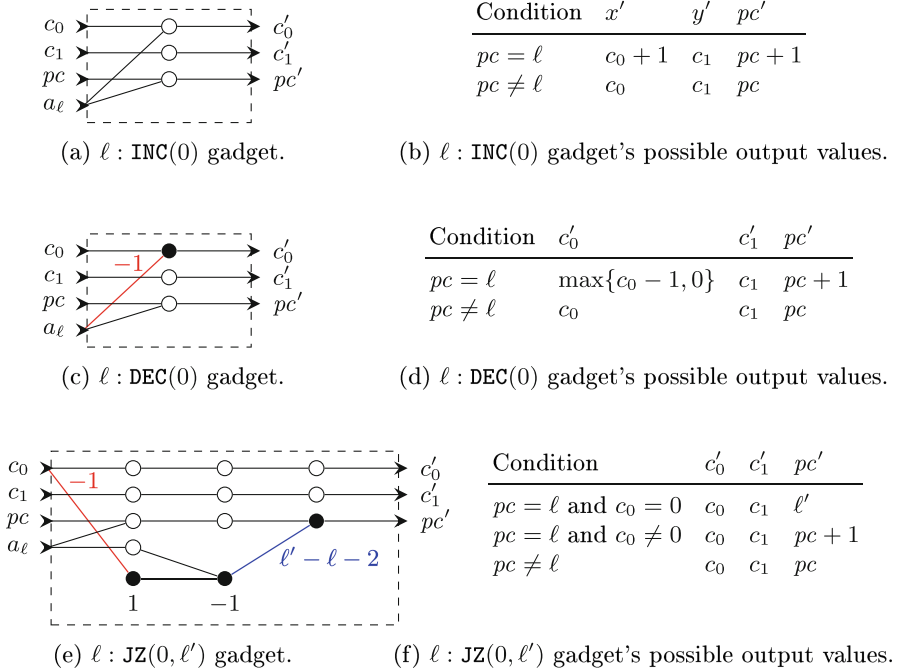
Finally, the initial input to the DNN is  $\mathbf{x}_0 = (0, 0, 0)$  (representing the initial configuration) and the target set is  $\varphi = \{(k - 1, c_0, c_1) \mid c_0, c_1 \geq 0\}$ , where  $k - 1$  is the last instruction number (STOP) of  $\mathcal{M}$ . Clearly,  $\mathcal{M}$  terminates if and only if the NNCS reaches a state satisfying  $\varphi$  when started in  $X_0 = \{\mathbf{x}_0\}$ .  $\square$

We note that the DNNs simulating two-counter machines are rather simple.

**Corollary 1.** *The NNCS reachability problem remains undecidable for DNNs with integral weights, 3 input and output dimensions, 6 hidden layers, a singleton initial set, and a target set  $o = v$  for some output neuron  $o$  and constant  $v \in \mathbb{N}$ .*

One may wonder whether the six hidden layers are necessary. In general, one cannot hope to obtain a small neural network when removing layers [4]. However, since we can iterate the NNCS, and the plant model is not interfering, we can reduce one iteration of a DNN  $N$  with six hidden layers (constructed in the proof above) to seven iterations of a DNN  $N'$  with one hidden layer<sup>1</sup>. Figure 5 shows a sketch of the construction idea. Essentially, we take the hidden layers of  $N$  and stack them as one wide hidden layer in  $N'$ . (For instance, layer  $\ell_1$  has width  $k + 1$ .) We refer to each of these hidden layers as a *track*. For each track, we need to add input and output dimensions corresponding to the number of neurons in the respective previous and next hidden layers. The output of track  $j$  is fed to track  $j + 1$  (and the output of the last track is fed to the first track).

<sup>1</sup> Typically, neural networks with only one hidden layer are not called *deep*.



**Fig. 3.** The gadgets for the three instructions. See Fig. 2 for further explanations

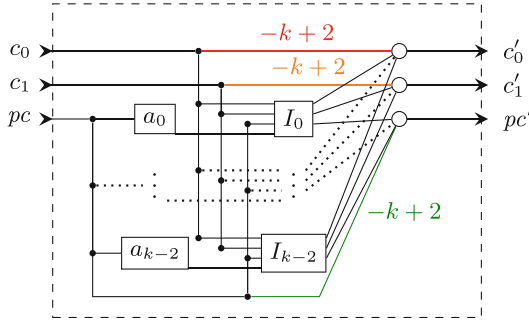
When presented with an input vector  $\mathbf{x}_0$  of appropriate size, the first track performs the computation of the first hidden layer and feeds its output to the second track, and so on. After seven iterations, the output of the last track will equal the output of the seven-layer DNN  $N$  after the first iteration. This output is then used as the input of the first track again and the process continues.

Finally, we need to make sure that the other tracks do not accidentally produce an output that leads to a target state between multiples of seven iterations. In order to only consider outputs in every seventh iteration, we use the additional gadget shown in Fig. 5(b). This gadget has seven inputs and outputs and is to be stacked below the other DNN. When the initial input is  $(1, 0, 0, 0, 0, 0, 0)$ , the 1 is propagated to the second index, and so on, until it arrives back at the first index after seven iterations.

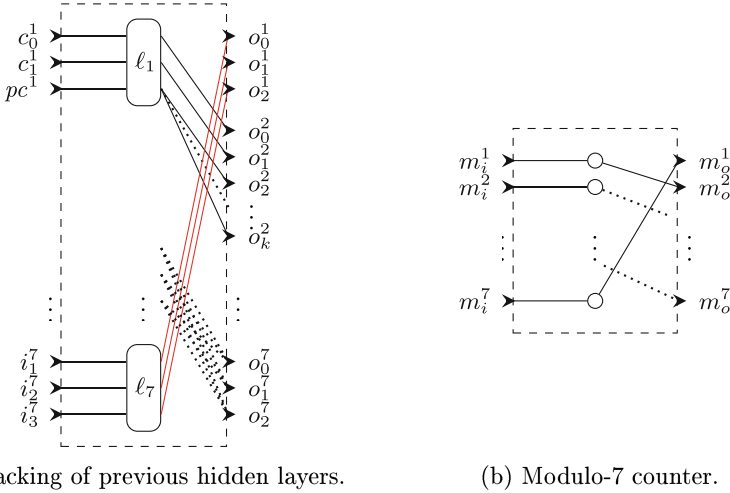
The target set  $\varphi$  now simply needs to get extended to arbitrary values in the auxiliary dimensions, except for the seven-last entry ( $m_i^1$ ), which has to equal one. Formally:  $\varphi' = (\varphi \wedge m_i^1 \leq 1 \wedge -m_i^1 \leq -1)$ .

In summary, by scaling the number of inputs and outputs with  $\mathcal{M}$ , we obtain a DNN with one (wide) hidden layer.

**Corollary 2.** *The NNCS reachability problem remains undecidable for DNNs with integral weights, one hidden layer, a singleton initial set, and a target set  $o = v$  for some output neuron  $o$  and constant  $v \in \mathbb{N}$ .*



**Fig. 4.** Complete construction. Each box represents an auxiliary gadget  $a_\ell$  resp. an instruction gadget  $I_\ell$ . Small dots denote junctions of connections and have no further semantics. The last layer is the output layer (with identity activations)



**Fig. 5.** Construction with a single hidden layer

### 4 Semi-decidability

In this section, we show that the NNCS reachability problem is semi-decidable for a particular class of plants. Indeed, from a single initial state  $\mathbf{x}_0$ , we can enumerate all states  $C_{P,N}(\mathbf{x}_0)^k$  reachable in  $k$  iterations and for each of them decide membership in the target polyhedron  $\varphi$ . However, since we allow for an initial set  $X_0$ , this algorithm is not effective.

The image of a polyhedron under a ReLU DNN is a (finite) union of polyhedra [21]. If we choose a class of plants with the same property, we obtain an effective algorithm again. In what follows, we show a more general result by using an automaton encoding of DNNs from [23]. This will allow us to more abstractly consider a class of plants that is definable in the same automaton formalism.

We slightly deviate from the original approach by Sälzer et al. [23] in that we use a more expressive automaton model, as we are (unlike them) not bothered with efficiency considerations (since our problem is undecidable).

**Definition 3 (Büchi Automaton).** A (nondeterministic) Büchi automaton (NBA)  $\mathcal{A} = (Q, \Sigma, q_0, \delta, F)$  consists of a finite set  $Q$  of states, a finite alphabet  $\Sigma$ , an initial state  $q_0 \in Q$ , a transition relation  $\delta \subseteq Q \times \Sigma \times Q$ , and a set of accepting states  $F \subseteq Q$ .

A run on an infinite word  $w = a_0a_1\dots$  is an infinite sequence of states  $q_0, q_1, \dots$  starting in the initial state and satisfying  $(q_i, a_i, q_{i+1}) \in \delta$  for all  $i \geq 0$ . A run is accepting if  $q_i \in F$  for infinitely many  $i$ . The language of  $\mathcal{A}$  is

$$L(\mathcal{A}) = \{w \in \Sigma^\omega \mid \mathcal{A} \text{ has an accepting run on } w\}.$$

A language is  $\omega$ -regular if there exists an NBA that accepts it.

In the following, we recall an effective encoding of real numbers in NBA from [23]. Let  $\Sigma = \{+, -, 0, 1, .\}$ . A word  $w = sa_n \dots a_0.b_0b_1\dots$  with  $n \geq 0$ ,  $s \in \{+, -\}$ ,  $a_i, b_i \in \{0, 1\}$  encodes the real value

$$dec(w) = (-1)^{\text{sign}(s)} \cdot \left( \sum_{i=0}^n a_i \cdot 2^i + \sum_{i=0}^{\infty} b_i \cdot 2^{-(i+1)} \right)$$

where  $\text{sign}(s) = 0$  if  $s = +$  and  $\text{sign}(s) = 1$  if  $s = -$ . As usual, the word encoding is not unique, but the decoding is [23, Page 5].

Now, we switch to a word encoding of multiple numbers by using a product alphabet. A symbol over this product alphabet  $\Sigma^k$  is a  $k$ -vector of symbols. A word over  $\Sigma^k$  is well-formed if it is of the form

$$w = \begin{bmatrix} s_1 \\ \vdots \\ s_k \end{bmatrix} \begin{bmatrix} a_{1,n} \\ \vdots \\ a_{k,n} \end{bmatrix} \cdots \begin{bmatrix} a_{1,0} \\ \vdots \\ a_{k,0} \end{bmatrix} \begin{bmatrix} . \\ \vdots \\ . \end{bmatrix} \begin{bmatrix} b_{1,0} \\ \vdots \\ b_{k,0} \end{bmatrix} \begin{bmatrix} b_{1,1} \\ \vdots \\ b_{k,1} \end{bmatrix} \cdots$$

where  $s_i \in \{+, -\}$ ,  $a_{i,j}, b_{i,h} \in \{0, 1\}$  for  $i = 1, \dots, k$ ,  $j = 0, \dots, n$ , and  $h = 0, 1, \dots$ . In other words, the signs and the point are aligned, which can be achieved by filling up with leading zeros. The language  $WF_k$  of well-formed words is  $\omega$ -regular [23]. The selection of a single component  $i \in \{1, \dots, k\}$  is obtained in the obvious way:

$$w_i = s_i a_{i,n} \dots a_{i,0} . b_{i,0} b_{i,1} \dots$$

If an NBA over  $\Sigma^k$  accepts only well-formed words, then we can understand its language as a relation over  $\mathbb{R}^k$ . Furthermore, linear constraints are also  $\omega$ -regular [23]. Thus, as NBA are closed under intersection and union, (finite unions of) polyhedra are also  $\omega$ -regular. Finally, we can also use NBAs to encode functions  $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$  via their graphs, which are relations over  $\mathbb{R}^{m+n}$ .

Sälzer et al. showed that every function computable by a DNN can be represented by an NBA<sup>2</sup>.

**Proposition 2 (Theorem 1 in [23]).** *Let  $N: \mathbb{R}^d \rightarrow \mathbb{R}^c$  be a DNN. There exists an NBA  $\mathcal{A}_N$  over  $\Sigma^{d+c}$  with*

$$L(\mathcal{A}_N) = \{w \in WF_{d+c} \mid N(\text{dec}(w_1), \dots, \text{dec}(w_d)) = (\text{dec}(w_{d+1}), \dots, \text{dec}(w_{d+c}))\}.$$

For our application, we need to slightly modify the automaton  $\mathcal{A}_N$  from Proposition 2 so that it also copies its input for further use. This modification can be implemented by replacing each transition label  $(a_1, \dots, a_n, a'_1, \dots, a'_m)$  by  $(a_1, \dots, a_n, a_1, \dots, a_n, a'_1, \dots, a'_m)$ .

**Corollary 3** *Let  $N: \mathbb{R}^d \rightarrow \mathbb{R}^c$  be a DNN. There exists an NBA  $\widehat{\mathcal{A}}_N$  over  $\Sigma^{d+d+c}$  with*

$$L(\widehat{\mathcal{A}}_N) = \{w \in WF_{d+d+c} \mid (w_1, \dots, w_d) = (w_{d+1}, \dots, w_{d+d}) \text{ and} \\ P((\text{dec}(w_1), \dots, \text{dec}(w_d)) = (\text{dec}(w_{d+d+1}), \dots, \text{dec}(w_{d+d+c})))\}.$$

Thus, the semantics of DNNs can be captured by NBAs. This is in general not true for plants. Hence, in the following, we restrict ourselves to plants that can also be captured by NBAs.

**Definition 4 ( $\omega$ -Regular Plant).** *A plant  $P: \mathbb{R}^{d+c} \rightarrow \mathbb{R}^d$  is  $\omega$ -regular if there exists an NBA  $\mathcal{A}_P$  over  $\Sigma^{d+c+d}$  such that*

$$L(\mathcal{A}_P) = \{w \in WF_{d+c+d} \mid \\ P(\text{dec}(w_1), \dots, \text{dec}(w_{d+c})) = (\text{dec}(w_{d+c+1}), \dots, \text{dec}(w_{d+c+d}))\}.$$

Now, both the DNN and the plant are given by NBA. Hence, we can apply standard automata-theoretic constructions to capture a bounded number of applications of the control loop by repeatedly composing the NBA for the DNN and the NBA for the plant. To this end, we introduce the (parametric) composition operator  $\circ_k$  constructing from two NBAs  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , which accept the graphs of two functions  $f_1: \mathbb{R}^{k_1} \rightarrow \mathbb{R}^k$  and  $f_2: \mathbb{R}^k \rightarrow \mathbb{R}^{k_2}$ , an NBA  $\mathcal{A}_1 \circ_k \mathcal{A}_2$  accepting the graph of  $\mathbf{x} \mapsto f_2(f_1(\mathbf{x}))$ .

**Lemma 1 (Lemma 4 of [23]).** *Let  $k, k_1, k_2 \geq 0$  and let  $\mathcal{A}_1$  and  $\mathcal{A}_2$  be two NBAs over  $\Sigma^{k_1+k}$  and  $\Sigma^{k+k_2}$ , respectively. Then, there exists an NBA  $\mathcal{A}_1 \circ_k \mathcal{A}_2$  over  $\Sigma^{k_1+k_2}$  accepting the language*

$$\{(u_1, \dots, u_{k_1}, w_{k+1}, \dots, w_{k+k_2}) \mid \exists (v_1, \dots, v_k) \text{ s.t.} \\ (u_1, \dots, u_{k_1}, v_1, \dots, v_k) \in L(\mathcal{A}_1) \text{ and} \\ (v_1, \dots, v_k, w_{k+1}, \dots, w_{k+k_2}) \in L(\mathcal{A}_2)\}.$$

---

<sup>2</sup> They actually proved the result for the more restrictive class of eventually-always weak NBA. But for us it is more prudent to consider the more general class of NBA.

Now we are ready to prove our main result of this section: NNCS reachability restricted to  $\omega$ -regular plants is semi-decidable. Note that this is tight, as the problem is undecidable as shown in Theorem 1: the plant just returning its control input as output is  $\omega$ -regular.

**Theorem 2** *The NNCS reachability problem is semi-decidable when restricted to  $\omega$ -regular plants.*

*Proof* We are given a problem instance  $(N, P, X_0, \varphi)$  and need to (semi)-decide whether there exists a  $k \geq 0$  such that  $(C_{P,N})^k(\mathbf{x}_0) \in \varphi$  for some  $\mathbf{x}_0 \in X_0$ . Let  $d$  be the dimension of the states of  $P$  and  $c$  be the dimension of the control vectors computed by  $N$ , respectively.

Let  $\mathcal{A}_N$  (over  $\Sigma^{d+d+c}$ ) and  $\mathcal{A}_P$  (over  $\Sigma^{d+c+d}$ ) be the NBAs as in Corollary 3 and Definition 4. Then, we define  $I_0$  to be an NBA accepting the graph of the  $d$ -ary identity function

$$L(I_0) = \{w \in WF_{d+d} \mid (w_0, \dots, w_d) = (w_{d+1}, \dots, w_{d+d})\}$$

and, for  $k \geq 1$ ,  $I_k = I_{k-1} \circ_d (\widehat{\mathcal{A}}_N \circ_{d+c} \mathcal{A}_P)$ .

By construction, we have

$$(w_1, \dots, w_d, w_{d+1}, \dots, w_{d+d}) \in L(I_k)$$

if and only if

$$(dec(w_{d+1}), \dots, dec(w_{d+d})) \in (C_{P,N})^k(dec(w_1), \dots, dec(w_d)).$$

There are NBAs  $\mathcal{A}_0$  and  $\mathcal{A}_\varphi$  accepting  $X_0$  and  $\varphi$ , as they are polyhedra. Both these NBAs have alphabet  $\Sigma^d$ , while each  $I_k$  has alphabet  $\Sigma^{d+d}$  where the first  $d$  components encode the inputs and the last  $d$  components encode the outputs. Hence, to restrict  $\mathcal{A}_0$  and  $\mathcal{A}_\varphi$  to  $X_0$  and  $\varphi$ , we need to widen  $\mathcal{A}_0$  and  $\mathcal{A}_\varphi$  to NBA with alphabet  $\Sigma^{d+d}$ . Formally, let NBAs  $\widehat{\mathcal{A}}_0$  and  $\widehat{\mathcal{A}}_\varphi$  (both over  $\Sigma^{d+d}$ ) such that

- $L(\widehat{\mathcal{A}}_0)$  contains the encodings of all vectors  $(x_1, \dots, x_{d+d}) \in WF_{d+d}$  such that  $(x_1, \dots, x_d)$  is in  $X_0 \subseteq \mathbb{R}^d$  and  $(x_{d+1}, \dots, x_{d+d}) \in \mathbb{R}^d$  is arbitrary, and
- $L(\widehat{\mathcal{A}}_\varphi)$  contains the encodings of all vectors  $(x_1, \dots, x_{d+d}) \in WF_{d+d}$  such that  $(x_1, \dots, x_d) \in \mathbb{R}^d$  is arbitrary and  $(x_{d+1}, \dots, x_{d+d})$  is in  $\varphi \subseteq \mathbb{R}^d$ .

Now, there exist an  $\mathbf{x}_0 \in X_0$  and a  $k \geq 0$  such that  $(C_{P,N})^k(\mathbf{x}_0) \in \varphi$  if and only if the language of  $\widehat{\mathcal{A}}_0 \cap I_k \cap \widehat{\mathcal{A}}_\varphi$  is nonempty.

In summary, to semi-decide the NNCS reachability problem for  $\omega$ -regular plants, we iteratively construct  $I_k$  for  $k \geq 0$  and check  $\widehat{\mathcal{A}}_0 \cap I_k \cap \widehat{\mathcal{A}}_\varphi$  for nonemptiness. □

Let us remark that the construction in Theorem 2 does not require the initial set  $X_0$  and the target set  $\varphi$  to be polyhedral. It is sufficient that they are  $\omega$ -regular to effectively decide (non)emptiness of the intersection. The class of  $\omega$ -regular languages is strictly more expressive than polyhedral sets<sup>3</sup>. Thus, our result is more general than the statement of Theorem 2.

<sup>3</sup> For example, the set of natural numbers is  $\omega$ -regular (in the encoding used here), but it is not a polyhedron.

### 4.1 Multi-mode Linear Plants

In this subsection, we give an example of a plant model that falls into the class of  $\omega$ -regular languages. Our example is inspired by linear hybrid automata [3], which are finite state machines with constant-term ordinary differential equations (ODEs) in the modes (states) and guard conditions on the transitions.

Hybrid automata have two sources of nondeterminism: an enabled transition need not be taken (may-semantics), and multiple transitions may be enabled at the same time. Because we have restricted ourselves to deterministic plants in this work, we need to introduce some restrictions. First, we assume a fixed rational control period, and a transition can only be taken at the end of such a period. Then, the solution of the ODEs is a linear map, which can be analytically computed, and our system becomes discrete-time. Second, we require that exactly one guard is enabled, i.e., in each mode, all guards are pairwise-disjoint and their union is the universe. To simplify the presentation, we do not include discrete updates with the transitions but note that these can easily be added. We call the resulting model a multi-mode linear map.

**Definition 5 (Multi-Mode Linear Map).** A multi-mode linear map is a tuple  $\mathcal{H} = (M, E, d, c, F, G)$  consisting of a finite set  $M \subseteq \mathbb{N}$  of modes, a set of edges  $E \subseteq M \times M$ , input and control dimensions  $d$  and  $c$ , a flow function  $F: M \rightarrow \mathbb{Q}^{d \times d} \times \mathbb{Q}^{d \times c} \times \mathbb{Q}^d$  (mapping a mode to two matrices and a vector), and a guard function  $G: E \rightarrow \mathcal{FUP}(d+c)$  (where  $\mathcal{FUP}$  denotes the set of finite unions of polyhedra), satisfying

- if  $(m, m') \in E$  and  $(m, m'') \in E$ , then  $G(m, m') \cap G(m, m'') = \emptyset$ , and
- $\bigcup_{m' \in M} G(m, m') = \mathbb{R}^{d+c}$  for all  $m \in M$ .

The function  $f_{\mathcal{H}}: M \times \mathbb{R}^{d+c} \rightarrow M \times \mathbb{R}^d$  computed by  $\mathcal{H}$  is defined as

$$f_{\mathcal{H}}(m, x_1, \dots, x_d, u_1, \dots, u_c) = (m', x'_1, \dots, x'_d)$$

where  $F(m) = (A, B, c)$ ,

$$(x'_1, \dots, x'_d) = A \cdot (x_1, \dots, x_d)^T + B \cdot (u_1, \dots, u_c)^T + c,$$

and  $m'$  is the unique mode such that

$$(x'_1, \dots, x'_d, u_1, \dots, u_c) \in G(m, m').$$

Note that the first component of inputs for  $f_{\mathcal{H}}$  is restricted to modes of  $\mathcal{H}$ , not arbitrary reals as stipulated by the definition of plants. However, this is not an issue as long as the initial input has a mode in the first component, as  $f_{\mathcal{H}}$  also returns only outputs that have a mode in the first component.

**Lemma 2** Multi-mode linear maps are  $\omega$ -regular plants.

*Proof* (Sketch) The following operations can be implemented by NBAs [23]:



- Multiplication of real inputs with constants in  $\mathbb{Q}$  and addition of reals. These two operations allow us to compute the output  $(x'_1, \dots, x'_d)$  from  $(x_1, \dots, x_d)$  and  $(u_1, \dots, u_c)$ .
- Checking membership of a vector of reals in a fixed polyhedron. This allows us to compute the next mode  $m'$  from the current mode  $m$ , the current state  $(x_1, \dots, x_d)$ , and the current input  $(u_1, \dots, u_c)$ , as  $m'$  is determined by the membership of  $(x_1, \dots, x_d, u_1, \dots, u_c)$  in a finite union of polyhedra.

This allows us to build an NBA that accepts the graph of  $f_{\mathcal{H}}$  for every given multi-mode linear map  $\mathcal{H}$ .  $\square$

**Corollary 4** *The NNCS reachability problem is semi-decidable when the plant is restricted to multi-mode linear maps.*

## 5 Conclusion

In this paper, we studied the reachability problem for dynamical systems controlled by deep neural networks. We showed that, for the common ReLU activations, the problem is undecidable even when the plant is trivial and the network is restricted to integral weights and a singleton initial set; furthermore, we can either fix the input and output dimensions to 3 and the number of hidden layers to 6, or use a single hidden layer. We then turned to the question when the problem can be semi-decided; here we extended a recent encoding of neural networks in Büchi automata and showed that  $\omega$ -regular plants as well as input and target sets are sufficient for a semi-decision procedure; as an example, we demonstrated that a model akin to linear hybrid automata is  $\omega$ -regular.

**Acknowledgments.** We thank the participants of AISoLA 2023 for suggesting to study the NNCS reachability problem with one hidden layer.

This research was partly supported by the Independent Research Fund Denmark under reference number 10.46540/3120-00041B, DIREC - Digital Research Centre Denmark under reference number 9142-0001B, and the Villum Investigator Grant S4OS under reference number 37819.

## References

1. Althoff, M.: An introduction to CORA 2015. In: ARCH. EPiC Series in Computing, vol. 34, pp. 120–151. EasyChair (2015). <https://doi.org/10.29007/zbkv>
2. Althoff, M., Frehse, G., Girard, A.: Set propagation techniques for reachability analysis. *Annu. Rev. Control Robot. Auton. Syst.* **4**, 369–395 (2021). <https://doi.org/10.1146/annurev-control-071420-081941>
3. Alur, R., Courcoubetis, C., Henzinger, T.A., Ho, P.: Hybrid automata: An algorithmic approach to the specification and verification of hybrid systems. In: *Hybrid Systems. LNCS*, vol. 736, pp. 209–229. Springer, Cham (1992). [https://doi.org/10.1007/3-540-57318-6\\_30](https://doi.org/10.1007/3-540-57318-6_30)
4. Arora, R., Basu, A., Mianjy, P., Mukherjee, A.: Understanding deep neural networks with rectified linear units. In: *ICLR. OpenReview.net* (2018). [https://openreview.net/forum?id=B1J\\_rgWRW](https://openreview.net/forum?id=B1J_rgWRW)

5. Bogomolov, S., Forets, M., Frehse, G., Potomkin, K., Schilling, C.: JuliaReach: a toolbox for set-based reachability. In: HSCC, pp. 39–44. ACM (2019). <https://doi.org/10.1145/3302504.3311804>
6. Cabessa, J.: Turing complete neural computation based on synaptic plasticity. PLoS ONE **14**(10), e0223451 (2019). <https://doi.org/10.1371/journal.pone.0223451>
7. Doyle, J.C., Francis, B.A., Tannenbaum, A.R.: Feedback Control Theory. Dover Publications (2013)
8. Goodfellow, I.J., Bengio, Y., Courville, A.C.: Deep Learning. In: Adaptive Computation and Machine Learning. MIT Press (2016). <http://www.deeplearningbook.org/>
9. Henzinger, T.A., Kopke, P.W., Puri, A., Varaiya, P.: What’s decidable about hybrid automata? J. Comput. Syst. Sci. **57**(1), 94–124 (1998). <https://doi.org/10.1006/jcss.1998.1581>
10. Hyötyniemi, H.: Turing machines are recurrent neural networks. STeP **96**, 13–24 (1996)
11. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: CAV. LNCS, vol. 10426, pp. 97–117. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-63387-9\\_5](https://doi.org/10.1007/978-3-319-63387-9_5)
12. Kochdumper, N., Schilling, C., Althoff, M., Bak, S.: Open- and closed-loop neural network verification using polynomial zonotopes. In: NFM. LNCS, vol. 13903, pp. 16–36. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-33170-1\\_2](https://doi.org/10.1007/978-3-031-33170-1_2)
13. Koiran, P., Moore, C.: Closed-form analytic maps in one and two dimensions can simulate universal turing machines. Theor. Comput. Sci. **210**(1), 217–223 (1999). [https://doi.org/10.1016/S0304-3975\(98\)00117-0](https://doi.org/10.1016/S0304-3975(98)00117-0)
14. Larsen, K.G., Mikucionis, M., Taankvist, J.H.: Safe and optimal adaptive cruise control. In: Correct System Design. LNCS, vol. 9360, pp. 260–277. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-23506-6\\_17](https://doi.org/10.1007/978-3-319-23506-6_17)
15. Le, D.M., Greene, M.L., Makumi, W.A., Dixon, W.E.: Real-time modular deep neural network-based adaptive control of nonlinear systems. IEEE Control. Syst. Lett. **6**, 476–481 (2022). <https://doi.org/10.1109/LCSYS.2021.3081361>
16. Liu, C., Arnon, T., Lazarus, C., Strong, C.A., Barrett, C.W., Kochenderfer, M.J.: Algorithms for verifying deep neural networks. Found. Trends Optim. **4**(3–4), 244–404 (2021). <https://doi.org/10.1561/24000000035>
17. Lopez, D.M., Althoff, M., Forets, M., Johnson, T.T., Ladner, T., Schilling, C.: ARCH-COMP23 category report: Artificial intelligence and neural network control systems (AINNCS) for continuous and hybrid systems plants. In: ARCH. EPiC Series in Computing, vol. 96. EasyChair (2023). <https://doi.org/10.29007/x38n>
18. Lopez, D.M., Choi, S.W., Tran, H., Johnson, T.T.: NNV 2.0: The neural network verification tool. In: CAV. LNCS, vol. 13965, pp. 397–412. Springer, Cham (2023). [https://doi.org/10.1007/978-3-031-37703-7\\_19](https://doi.org/10.1007/978-3-031-37703-7_19)
19. Miller, W.T., Sutton, R.S., Werbos, P.J.: Neural Networks for Control. MIT Press (1995)
20. Minsky, M.L.: Computation: Finite and Infinite Machines. Prentice-Hall (1967)
21. Montúfar, G., Pascanu, R., Cho, K., Bengio, Y.: On the number of linear regions of deep neural networks. In: NeurIPS, pp. 2924–2932 (2014). <https://proceedings.neurips.cc/paper/2014/hash/109d2dd3608f669ca17920c511c2a41e-Abstract.html>
22. Moore, C.: Unpredictability and undecidability in dynamical systems. Phys. Rev. Lett. **64**(20), 2354–2357 (1990). <https://doi.org/10.1103/PhysRevLett.64.2354>

23. Sälzer, M., Alsmann, E., Bruse, F., Lange, M.: Verifying and interpreting neural networks using finite automata. CoRR (2022). <https://doi.org/10.48550/arXiv.2211.01022>
24. Sälzer, M., Lange, M.: Reachability in simple neural networks. *Fundam. Informaticae* **189**(3–4), 241–259 (2022). <https://doi.org/10.3233/FI-222160>
25. Schilling, C., Forets, M., Guadalupe, S.: Verification of neural-network control systems by integrating Taylor models and zonotopes. In: AAAI, pp. 8169–8177. AAAI Press (2022). <https://doi.org/10.1609/aaai.v36i7.20790>
26. Schilling, C., Lukina, A., Demirović, E., Larsen, K.G.: Safety verification of decision-tree policies in continuous time. In: NeurIPS. vol. 36, pp. 14750–14769. Curran Associates, Inc. (2023). [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/2f89a23a19d1617e7fb16d4f7a049ce2-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/2f89a23a19d1617e7fb16d4f7a049ce2-Abstract-Conference.html)
27. Siegelmann, H.T., Sontag, E.D.: Turing computability with neural nets. *Appl. Math. Lett.* **4**(6), 77–80 (1991). [https://doi.org/10.1016/0893-9659\(91\)90080-F](https://doi.org/10.1016/0893-9659(91)90080-F)
28. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014). <http://arxiv.org/abs/1312.6199>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Author Index

## A

Aichernig, Bernhard K. 385

## B

Bantry White, Eleanor 132

Berger, Thorsten 434

Bernasconi, Eleonora 144

Bertl, Markus 165, 203

Beyer, Dirk 351

Black, Ashton 333

Bongo, Lars Ailo 30

Brandon, Colm 245

Breathnach, Ciara 3, 65

Brehmer, Kai 225

Bucciarelli, Nicholas 203

## C

Chiffi, Daniele 284

Clancy, Stuart 104

Cohen, Itay 362

Conte, Marina Marchisio 48

## D

Di Pierro, Davide 144

Draheim, Dirk 203

## F

Fatemi, Bahareh 187

Ferilli, Stefano 144

Fissore, Cecilia 48

Floris, Francesco 48

Fränzle, Martin 415

## G

Garrett, Eilidh 30

Gasperini, Anna 119

## H

Havelund, Klaus 385

Hein, Andreas 415

Hintersdorf, Dominik 269

## I

Islam, Maisha 30

## K

Kersting, Kristian 269

Knopp, Henriette 434

Kristoffersen, Doris Tove 30

Krumrey, Marco 84

## L

Lamo, Yngve 165, 187

Langer, Markus 293

Le Roux, Maelle 119

Leucker, Martin 165, 225

## M

Margaria, Tiziana 3, 65, 84, 165, 245

Mitwalli, Daniel Sami 84

Mohammadi, Esfandiar 165

Mourits, Rick J. 10

Mukhiya, Suresh Kumar 165, 187

Murphy, Rachel 65

## N

Nowotka, Dirk 225

## O

O'Shea, Enda 84

## P

Pechmann, Ludwig 165, 203, 225

Pedersen, Bjørn-Richard 30

Peldszus, Sven 434

Peled, Doron 362

Pesch, Dirk 132

Piho, Gunnar 165, 203

Potdevin, Yannik 225

## Q

Qirtas, Malik Muhammad 132

**R**

Rabbi, Fazle 165, 187  
Redavid, Domenico 144  
Reid, Alice 30  
Riswick, Tim 10  
Ross, Peeter 203  
Ruscheimer, Hannah 318

**S**

Sacchet, Matteo 48  
Schiaffonati, Viola 284  
Schieweck, Alexander 65  
Schilling, Christian 455  
Schwammberger, Maike 307  
Sens, Yorick 434  
Sharma, Rahul 203  
Singh, Amandeep 245

Sommereth, Hilde 30  
Stapel, Rombert 10  
Struppek, Lukas 269

**T**

Teumert, Sebastian 84

**V**

Valera, Isabel 293

**W**

Walsh, Oonagh 104

**Z**

Zafeiridi, Evi 132  
Zanotti, Giacomo 284  
Zimmermann, Martin 455