CDSMS

EDITED BY

PIETER VERDEGEM

AI FOR EVERYONE?

CRITICAL PERSPECTIVES

AI for Everyone?

Critical Perspectives

Edited by
Pieter Verdegem

Critical, Digital and Social Media Studies

Series Editor: Christian Fuchs

The peer-reviewed book series edited by Christian Fuchs publishes books that critically study the role of the internet and digital and social media in society. Titles analyse how power structures, digital capitalism, ideology and social struggles shape and are shaped by digital and social media. They use and develop critical theory discussing the political relevance and implications of studied topics. The series is a theoretical forum for internet and social media research for books using methods and theories that challenge digital positivism; it also seeks to explore digital media ethics grounded in critical social theories and philosophy.

Editorial Board

Published

Critical Theory of Communication: New Readings of Lukács, Adorno, Marcuse, Honneth and Habermas in the Age of the Internet
Christian Fuchs
https://doi.org/10.16997/book1

Knowledge in the Age of Digital Capitalism: An Introduction to Cognitive Materialism
Mariano Zukerfeld
https://doi.org/10.16997/book3

Politicizing Digital Space: Theory, the Internet, and Renewing Democracy
Trevor Garrison Smith
https://doi.org/10.16997/book5

Capital, State, Empire: The New American Way of Digital Warfare
Scott Timcke
https://doi.org/10.16997/book6

The Spectacle 2.0: Reading Debord in the Context of Digital Capitalism
Edited by Marco Briziarelli and Emiliana Armano
https://doi.org/10.16997/book11

The Big Data Agenda: Data Ethics and Critical Data Studies
Annika Richterich
https://doi.org/10.16997/book14

Social Capital Online: Alienation and Accumulation
Kane X. Faucher
https://doi.org/10.16997/book16

The Propaganda Model Today: Filtering Perception and Awareness
Edited by Joan Pedro-Carañana, Daniel Broudy and Jeffery Klaehn
https://doi.org/10.16997/book27
Critical Theory and Authoritarian Populism
Edited by Jeremiah Morelock
https://doi.org/10.16997/book30
Peer to Peer: The Commons Manifesto
Michel Bauwens, Vasilis Kostakis and Alex Pazaitis
https://doi.org/10.16997/book33
Bubbles and Machines: Gender, Information and Financial Crises
Micky Lee
https://doi.org/10.16997/book34
Cultural Crowdfunding: Platform Capitalism, Labour and Globalization
Edited by Vincent Rouzé
https://doi.org/10.16997/book38
The Condition of Digitality: A Post-Modern Marxism for the Practice of Digital Life
Robert Hassan
https://doi.org/10.16997/book44
Incorporating the Digital Commons: Corporate Involvement in Free and Open Source Software
Benjamin J. Birkinbine
https://doi.org/10.16997/book39
The Internet Myth: From the Internet Imaginary to Network Ideologies
Paolo Bory
https://doi.org/10.16997/book48
Communication and Capitalism: A Critical Theory
Christian Fuchs
https://doi.org/10.16997/book45
Marx and Digital Machines: Alienation, Technology, Capitalism
Mike Healy
https://doi.org/10.16997/book47
The Commons: Economic Alternatives in the Digital Age
Vangelis Papadimitropoulos
https://doi.org/10.16997/book46
Intellectual Commons and the Law: A Normative Theory for Commons-Based Peer Production
Antonios Broumas
https://doi.org/10.16997/book49
The Fight Against Platform Capitalism: An Inquiry into the Global Struggles of the Gig Economy
Jamie Woodcock
https://doi.org/10.16997/book51

AI for Everyone?

Critical Perspectives

Edited by
Pieter Verdegem

University of Westminster Press
www.uwestminsterpress.co.uk

Published By
University of Westminster Press
115 New Cavendish Street
London W1W 6UW
www.uwestminsterpress.co.uk

First published 2021

Cover design: www.ketchup-productions.co.uk
Series cover concept: Mina Bach (minabach.co.uk)
Print and digital versions typeset by Siliconchips Services Ltd.

ISBN (Paperback): 978-1-914386-16-9
ISBN (PDF): 978-1-914386-13-8
ISBN (EPUB): 978-1-914386-14-5
ISBN (Kindle): 978-914386-15-2

DOI: https://doi.org/10.16997/book55

The full text of this book has been peer-reviewed to ensure high academic standards. For full review policies, see: http://www.uwestminsterpress.co.uk/site/publish.

Competing interests: The editors and contributors have no competing interests to declare.

Suggested Citation: Verdegem, P. (ed.) 2021. *AI for Everyone? Critical Perspectives*. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55. License: CC-BY-NC-ND 4.0

To read the free, open access version of this book online, visit https://doi.org/10.16997/book55 or scan this QR code with your mobile device

Contents

CHAPTER 1

Introduction: Why We Need Critical Perspectives on AI

Pieter Verdegem

Introduction

The renewed interest in Artificial Intelligence (AI) has made it the most recent hype in the world of technological innovation. In the business world, AI is seen as a catalyst for growth, which will manifestly transform the economy and the future of work (Agrawal, Gans and Goldfarb 2018; Lee 2018; McAfee and Brynjolfsson 2017). Policymakers and civil society are putting their hopes on AI for tackling global challenges such as pandemics and even climate change (Dobbe and Whittaker 2019; Dananjayan and Raj 2020). AI also seems to be the subject of an arms race between China, Russia and the USA for equipping their armies with automated weaponry (Asaro 2018).

Whenever we are confronted with a hype, it is of utmost importance to untangle what exactly is at stake and who is behind the discourses and myths created. We are being told stories about AI as the ultimate innovation, transforming the ways we live and work – often started in corporate circles and distributed by their supportive popular outlets. At the same time, however, analysis is revealing that AI itself is one reason behind intensifying societal problems and harms. Researchers and thinkers have observed and/or predicted that AI leads to discrimination (Zuiderveen Borgesius 2018), is the engine behind growing inequalities (Korinek and Stiglitz 2017), can bring about technological unemployment (Ford 2015) and may even contribute to the end of humanity (Bostrom 2014).

How to cite this book chapter:
Verdegem, P. 2021. Introduction: Why We Need Critical Perspectives on AI. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 1–18. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.a.

Amidst this doom and gloom, what we desperately need is a more nuanced debate about AI's risks and opportunities. This can – must – be a serious and informed discussion that goes beyond hyperbole and polarisation, fuelled by popular media and thus feeding into public debate. What we need is critical perspectives on AI: what it is and what it is not; what type of AI we need, what visions exist about this and who is behind them; and ultimately, how to think and talk about AI power and inequalities.

In one word, it is *power* that must be at the centre of our conversations about AI and that is what this book is about. If we want to talk about critical perspectives on AI, formulating a critique on AI, how it is currently being developed and discussed, and yes, if we are serious about making sure that AI will benefit everyone, we need to talk about power. Power refers to the capacity to influence the actions, beliefs or behaviour of others. Ultimately, this comes down to 'the question of who can influence what society looks like and who controls the means that allow such influence' (Fuchs 2017: 86). Power decides who will benefit from new technologies such as AI, but a concentration of power will likely result in growing inequalities and other negative outcomes. The current critiques about AI centre on AI ethics (Coeckelbergh 2020), which is valuable and important to shape policy discussions. AI ethics, however, also has serious limitations when it comes to bringing about real change and making sure that the benefits of AI are accessible for everyone. Further in this introduction, I elaborate on this and I make the case for a radical democratisation of AI, and why we need to put power at the centre for achieving this.

The contributions in this book braid discussion of power and critique with three strands: AI – Humans vs. Machines, Discourses and Myths About AI and AI Power and Inequalities.

Part 1: *AI – Humans vs. Machines* – deals with the history and conceptualisation of AI and what is at stake in its development. This section looks at different perspectives about what characterises machine intelligence and how it might be important to further radical humanism in the era of automation and AI.

Part 2: *Discourses and Myths About AI* – analyses how AI is framed in popular and scholarly discussions and investigates the normative projections of what AI should be and what it should do. This section poses critical questions about how AI needs to debunk the myths surrounding it.

Part 3: *AI Power and Inequalities* – advances the debate around AI by critically examining what 'AI for Everyone?' means. This is dealing with the root of the problem: who will benefit from AI is ultimately down to who has the power to decide. These contributions look at how AI capitalism is organised, what (new) inequalities it might bring about and how we can fight back.

Why do we need a book on AI for Everyone? and why do we need it now? The 2007–2008 financial crisis, and the resulting global economic crisis, has not only brought about a decade of austerity in large parts of the Western world; it has also been the context in which social media and digital platforms have transformed into behemoths. Tech companies are now dominating the top 10

of the most valuable companies in the world (Verdegem 2021). Austerity has also led to growing inequalities and political polarisation, bringing right-wing authoritarian politics into power in a number of countries (Fuchs 2018). A world already cracked by economic uncertainty and the looming threat of climate change was then shaken in 2020 by a global pandemic. COVID-19 has massively impacted the global economy, on a much larger scale than the 2007–2008 crisis. On top of this, the pandemic has also resulted in an even bigger dependence and dominance of tech platforms such as Amazon, Alibaba, Google and Tencent. These companies are, not surprisingly, also leading AI companies. Only a small number of corporations have the necessary computational power to develop AI systems, are financially strong enough to hire the brightest AI talent and have access to the gigantic datasets that are needed to train machine learning and deep learning (AI) models. This context makes it very clear why we need to ask critical questions about AI and power.

Conceptualising AI – What AI Are We Talking About?

Before understanding what type of AI we *want*, we need to understand what AI we *have*. This is an area of significant debate, and the book opens by exploring the varying approaches to how we define AI.

The Origins of AI

It is easy to forget that AI has been with us for more than 60 years. Despite the flash of excitement and anxiety that feels so recent, AI itself is not a new phenomenon. The name *Artificial Intelligence* (AI) was coined in the mid-1950s at a series of academic workshops organised at Dartmouth College, New Hampshire (USA). A group of scientists, led by mathematics professor John McCarthy, gathered to investigate the ways in which machines could simulate aspects of human intelligence: the ability to learn and make decisions. Their core assumption was that human reasoning could be reconstructed using mathematical techniques and, as a consequence, problem-solving could be formalised into algorithms (McCarthy et al. 1955/2006).

What is more recent is a reflexive, if not critical, and social-scientific, understanding of not just AI's capabilities, but its impacts on human life and social organisation (Elliott 2019). It took decades for AI research to move from what it could do *for* us to what it could do *to* us, or enable us to do to each other. These first critical insights came along with observations that AI can not only supercharge innovation and bring about economic prosperity but also lead to inequalities and unfairness.

This book contributes to this debate by critically reflecting on how we should think about AI and the relationship between humans and machines. It analyses the discourses and myths that exist around AI; what it will enable

and what not. And it looks at issues about AI, power and inequalities, investigating where the risks of exclusion are and how we should deal with this.

The book also brings diverse and critical voices to this debate. Whereas AI as a discipline has been dominated by white, male, predominantly older scientists from mathematical disciplines, this collection brings perspectives that are characterised by a strong diversity in authorship and discipline. And threading through all, the contributions offer a discussion of different tangents of power and political economy in the field of AI and society.

The first task is to name our terms. For a concept that has been with us for so long, there is little consensus on how to define it. The history of debating AI is almost as old as AI itself. There is more debate than agreement about what AI is and what it is not, and the only thing generally agreed is that there is no widely accepted definition (Russell and Norvig 2016). The first definition comes from that gathering of scientists in 1955: McCarthy et al. (1955/2006) then defined AI as: 'Making a machine behave in ways that would be called intelligent if a human were so behaving'. This only raises the challenge of how exactly to define *intelligence*. Russell and Norvig (2016: 2) define different approaches to AI to serve different goals. AI can refer to systems that: (1) think like humans; (2) think rationally; (3) act like humans; and (4) act rationally. Each of the approaches requires different disciplinary expertise, thus requiring an inter-, or at least cross-disciplinary discussion. The human-centred approaches will depart from social science studying human behaviour, while the rationalist approaches will involve a combination of mathematics and engineering. From the four approaches, *acting like humans* is closest to how we define and understand contemporary AI.

We can see the roots of *acting like humans* in the *Turing test*, developed by Alan Turing in 1950. This test, originally designed to provide a satisfactory definition of intelligence, has been central to conceptualising AI. According to the test, if a human interrogator cannot distinguish a machine from a human through conversation, then the machine can be considered *intelligent*. Russell and Norvig (2016) argue that for a computer to be intelligent – to pass the Turing test – it needs to possess the following capabilities: *natural language processing* (being able to communicate successfully), *knowledge representation* (being able to store what it knows or hears), *automated reasoning* (being able to use the stored information to answer questions and to draw new conclusions) and *machine learning* (being able to adapt to new circumstances and to detect and extrapolate patterns).

Towards an Operational Definition – For Now

It is helpful to first distinguish between *strong* and *weak* AI (Bostrom 2014). Strong AI, also called *AGI* (Artificial General Intelligence) refers to computational systems with *general* cognitive abilities which have the future

potential to surpass human intellectual capacities. This can be seen as the attempt to mechanise human-level intelligence. Computer scientists and philosophers disagree on whether this is at all possible (Coeckelbergh 2020): some directly reject this scenario while others think if theoretically possible, it is not likely to happen (soon) in practice (Boden 2016). This is why it might be better to focus on advancements in weak AI or *ANI* (Artificial Narrow Intelligence), as this is the type of AI already impacting everyday life on a massive scale. Weak/narrow AI performs *specific* tasks which would normally require intelligence in a human being – machines aiding human thought and action. This type of AI is a mathematical method for prediction (Agrawal et al. 2018). Such systems can be extremely powerful but are limited in the range of tasks they can perform.

Russell and Norvig (2016) see *machine learning* as a prerequisite for intelligent machines. Machine learning is a paradigm that allows programs to automatically improve their performance on a particular task by learning from vast amounts of data (Alpaydin 2016). It seeks and uses statistical patterns and correlation in enormous datasets. Unlike older types of AI (e.g. expert systems, that are based on rules which are inputted by humans), machine learning algorithms learn not from humans but from data. The availability of significant amounts of real-world data (that we produce by using the internet, social media, sensors or other Internet-of-Things applications), combined with the availability of powerful and almost limitless computing capacity and advancements in machine learning and deep learning is why we are currently in another period of AI optimism and hype (Elish and boyd 2018).

Given the concepts and the brief discussion above, how can we agree on an operational definition of AI? A basic definition would be to refer to AI as computer programming that learns from and adapts to data. A more elaborate version of this, as Elliott (2019: 4) puts it, defines AI as 'any computational system that can sense its relevant context and react intelligently to data' in order to perform highly complex tasks effectively and to achieve specific goals, thereby mimicking intelligent human behaviours. The discussion about how to define AI cannot be settled in one definition, let alone one book. It is an important starting point, however, and Part 1 and Part 2 of this book will unpack several approaches to defining AI.

The Realities of AI for Some vs. the Ideals of AI for Everyone

Visions of AI in Policies and Ethics

Examining AI policies and ethics helps us to explore questions of what type of AI we want/need, how its development should look like and how we deal with its impact. Policy development happens at several levels and includes a number of stakeholders: national governments, intergovernmental organisations, corporations, professional associations and academics.

While AI policies reflect the priorities of the stakeholders involved, ethical guidelines project a vision of what type of AI is preferred, what benefits it should deliver and how we should deal with potential risks. Obviously, this is part of a normative debate but we can learn a lot from who is involved in these discussions and how they aim to shape the future of AI.

Given the projections about the role of AI in economic development, AI is high on the policy agenda. Putin famously said that the nation that leads in AI would be the ruler of the world (Vincent 2017). Major nations are rushing to create AI initiatives, unsurprisingly led by China and the USA (Lee 2018). What is surprising, however, is how much overlap there is in their strategic vision.

China's national strategy for AI, the *New Generation Artificial Intelligence Development Plan* (NGAIDP), was released in 2017 (State Council of China 2017). China wants to become the world leader in AI by 2030 and has formulated strategic goals to achieve this, such as making China the superpower of fundamental and applied AI research and development in order to dominate the global AI market. The main focus of China's AI policy is on economic development and competition, even though it also discusses some concerns in terms of economic security and social stability.

The Trump administration launched the *American AI Initiative* in 2019 (White House 2019). This strategic policy is all about a nationalist vision of American leadership in AI. The US government wants to invest in AI R&D, set AI standards and build the AI workforce. The Trump AI strategy not surprisingly has an intense national focus, highlighting AI for American innovation, industry, workers and values, aimed at promoting and protecting national AI technology and innovation. There is some discussion of public trust and confidence in AI as well as the protection of civil liberties, privacy and American values but this is subordinated to leadership and protecting American AI technology. With the election of Biden, it remains to be seen what the shift in AI policies will be, but given his track record the US will continue to pursue US capitalist interests, although maybe in a less outspoken nationalist way.

Most European nations where we see AI policy development, including France, Germany and the UK, are taking a different approach, which more explicitly offers a normative vision of how AI should contribute to social progress. France, for example, has entitled its vision *AI for Humanity* and aims for the development of an ethical framework for transparent and fair use of AI applications (Villani 2018). Germany also wants to guarantee responsible development and deployment of AI which serves the good of society. The UK sits somewhere between the continental European visions and the US vision, with goals contributing to global AI development, tempered with nationalist objectives focusing on specific benefits for the UK.

It is clear that China and the US are in an intense battle for global AI leadership and their policies are dominated by nationalist goals. European countries want to engage in AI innovation and boost their competitiveness while also

ensuring that the societal impact of AI is not forgotten. But still, this does not tell us a lot about what type of AI we want/need; it rather explains what countries expect AI to do for them. The European Union, however, has done more to develop a vision of what type of AI needs to be pursued and what aspects need to be dealt with in this.

The EU situates itself between China (state capitalism) and the US (market capitalism) and seeks to shift the debate in terms of the impact on society and its citizens. This positioning is aligned with how they have approached General Data Protection Regulation (GDPR) in the context of data protection and privacy. The EU has put forward *trustworthy AI* as the key term highlighting what type of AI it likes to see developed. This concept is the result of an open consultation and its ethics guidelines have been presented by the *High-Level Expert Group of AI*. According to these guidelines, trustworthy AI should be: (1) lawful (respecting all applicable laws and regulations); (2) ethical (respect ethical principles and values); and (3) robust (both from a technical perspective while taking into account its social environment) (European Commission 2019). These aspects are vague (*how* is something ethical or robust exactly?) as well as self-evident (very few people would favour *unlawful* AI). The EU, however, has made these guidelines more explicit by formulating specific aims: human agency/oversight, technical robustness/safety, privacy/data governance, transparency, diversity/non-discrimination/fairness, societal/environmental well-being and accountability. This is helpful as the list of specific aims can be read as values we would like to attribute to AI.

Not only governments or governmental organisations are active in putting forward a vision for AI. Companies also have a stake in this debate so it is instructive to examine how leading tech companies talk about what type of AI they want to build. Google (2020) has developed a vision it calls *Advancing AI for Everyone*, which can be summed up as applying AI to improve their products and developing tools to ensure that everyone can access AI. Google also has an *AI for Social Good* project, similar to Microsoft's *AI for Good* program. The latter aims to use AI expertise to solve humanitarian and environmental challenges: AI for earth, health, accessibility, humanitarian action and cultural heritage. While seemingly well-intended at first glance, these AI programs are carefully designed to support goals of *corporate social responsibility* (Sandoval 2014) and are undeniably textbook examples of what Morozov (2013) has called *techno-solutionism*. The problem with these corporate AI visions is that they lack substance and therefore do not reveal anything about what they intend specifically and how they actually can and should benefit society.

More substance can be found in how professional associations propose a vision of what *Good AI* exactly means. Organisations such as the Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM) have produced codes that propose ethical principles for computer science in general and AI in particular. ACM (2020), for example,

talks about AI that needs to 'contribute to society and to human well-being', while IEEE (2020) has come up with principles for *ethically aligned design*. General principles include human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse and competence.

Often cited are the *Asilomar AI Principles*. The *Asilomar Conference on Beneficial AI* was organised by the Future of Life Institute (2017) and brought together more than 100 AI researchers from academia and industry and thought leaders in economics, law, ethics and philosophy to address and formulate principles of *beneficial AI*. The resulting Asilomar AI principles are organised around (1) research issues, (2) ethics and values and (3) longer-term issues (Future of Life Institute 2017). The first category, *research issues*, sets out some guidelines in terms of research goals, funding and culture. Secondly, thirteen specific *ethics and values* are listed, dealing with transparency, safety, privacy, etc. and they also address aspects such as *shared benefit* ('AI technologies should benefit and empower as many people as possible') and *shared prosperity* ('the economic prosperity created by AI should be shared broadly, to benefit all of humanity'). Last, under *longer-term issues*, cautionary aspects and risks are addressed, including the principle of *common good*, which states: 'Superintelligence [Artificial General Intelligence, as discussed above] should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organisation.'

While the Asilomar AI principles are valid, they leave unclear who can and should take ownership and what mechanisms can be developed to enforce them. One specific concern of the Asilomar AI initiative is the heavy involvement of corporate stakeholders, given that it is backed by tech giants including Google, Facebook and Apple. These are not non-profit organisations but companies that are among the most wealthy and profitable organisations in the world. They might say they want to develop AI applications that are beneficial for society but can we trust them not to use their power to shift the direction of AI development to their corporate benefit and the return on investment for their investors and shareholders?

The *AI4People* initiative, set up by the non-profit organisation Atomium-EISMD (European Institute for Science, Media and Democracy), is the European response to the Asilomar AI initiative. AI4People also brings together academics, business partners (e.g. Facebook, Intel and Microsoft), and civil society organisations. The ambition of AI4People (Atomium-EISMD, 2020) is 'to draft a set of ethical guidelines aimed at facilitating the design of policies favourable to the development of a "Good AI Society"'.

AI4People has developed an ethical framework of principles that should underpin the adoption of AI and offers a list of specific recommendations and action points that should help to establish a *Good AI Society* (Floridi et al. 2018). AI4People proposes five core ethical principles: (1) *beneficence* (promoting well-being, preserving dignity and sustaining the planet); (2) *non-*

maleficence (privacy, security and capability caution); (3) *autonomy* (the power to decide/whether to decide); (4) *justice* (promoting prosperity and preserving solidarity); and (5) *explicability* (enabling the other principles through intelligibility and accountability). While the first four overlap with traditional bioethics principles, the last one is 'a new enabling principle for AI' (Floridi et al. 2018, 700). There might be overlap with the Asilomar AI principles, but AI4People has come up with a comprehensive list of ethical principles, recommendations and action points that can help policymakers to develop and support AI projects and initiatives that benefit society. However, they are not without gaps and flaws.

What is Missing in AI Policies and Ethics: Introducing Capitalism

The overview of AI policies and initiatives aimed at formulating AI ethics, helps us understand the debate about what AI we want/need and what it should deliver (or what should be avoided). However, something crucial is missing: power. This brings us to the crux of the book and the possibilities of critical analysis of AI. To bring power into the debate, we must first understand two points: (1) the problem of AI ideology and (2) the limitations of ethics.

Let me start with *AI ideology*. National policies clearly illustrate that AI is seen as an important instrument for positioning countries in terms of what type of future society they aim to develop. Here comes the role of ideology. While a contested notion, ideology can refer to: 'worldviews and ideas on the one end, to the process of the production of false consciousness on the other end of the spectrum' (Fuchs 2020, 180). In other words, it can have a neutral meaning but ideology can also be used to manipulate human consciousness. In the latter meaning, ideology is seen as a typical characteristic of capitalism and class societies, and it is being used to serve the material interests of the ruling class (Fuchs 2020). As discussions of AI often include visions about its potential to radically alter societies and economies, we need to be alert to and critical towards AI ideology.

Berman (1992) wrote almost three decades ago that the growing interest in AI in capitalist societies can be understood not only in terms of its practical achievements but also in the ideological role it plays as a technological paradigm for the continuation and reinvention of capitalism. AI as an ideology means that it can be seen as: 'a potential hegemonic principle within the sphere of formal organizations which facilitates the "fit" of human beings into the revised structures of a capitalism based on micro-electronic and information technology, and ideologically contains, and significantly mutes, resistance and social conflict' (Berman 1992, 104). The technological paradigm is thus a major component of hegemonic ideology that helps to maintain the essential structures of the current capitalist system and makes coherent and viable

alternatives increasingly difficult to envision. AI ideology thus propagates one specific vision of what AI is and what it should do – including serving the interests of the ruling class – and discourages alternative visions from materialising.

Second, we need to be aware of the limitations of AI ethics. Computer ethics, the broader field to which AI ethics belongs, is a philosophical field of study that deals with the question of 'how computer technology should be used' (Moor 1985, 266). It investigates social impact but also how policies for ethical use of computer technology can be formulated and justified. This is important and is why I discuss not only ethical guidelines but also AI policies. AI ethics are important as they let us think about what a *good society* constitutes, how we – as members of that society – can live a meaningful and fulfilling life, and especially what the role of technology, in general, and AI, in particular, in this is (Coeckelbergh 2020). There are, however, problems and limitations with AI ethics and how they get linked to policy.

When it comes to developing AI ethical guidelines, the first question to ask is: who is involved? The issue of diversity and inclusion plays out on multiple levels. Research by Jobin, Ienca and Vayena (2019). (2019) demonstrates that developing AI ethics is concentrated in North America, the European Union, Japan and a small handful of other countries. The absence or underrepresentation of countries from Africa, Central and South America and Central Asia means that large global regions are not invited to contribute to this debate, illustrative of a geopolitical power imbalance. There are also questions about who exactly is involved in developing the guidelines and whether the panels of experts who produce ethical guidelines, are – or are not – representative of society. This undermines the plurality that AI ethics aim for.

Another problem of establishing AI ethics is the speed at which technologies are developing (Boddington 2017). Formulating ethical guidelines takes time and there is a question of whether or not ethics can keep up with the rapid development of technologies. AI policies, just as any policies, face a similar challenge and as a consequence they are often reactive rather than proactive. AI ethical guidelines also face the problem of *ethics washing* (Wagner 2018). This refers to the practice of exaggerating a company's interest in promoting beneficial AI systems (Google's *Advancing AI for Everyone* (2020)and Microsoft's *AI for Good* (2020) programs, cfr. supra, are often used as examples for this) but also when ethics is used as a substitute for regulation, meaning that companies highlight how ethically they are acting while simultaneously abandoning their legal obligations (for example, not respecting principles in terms of data protection).

The vulnerability of ethics advocates and researchers is illustrated by the case of Timnit Gebru. Gebru is well-known for her work on racial bias in technology, such as facial recognition, and has criticised systems that fail to recognise black faces. She was fired by Google in December 2020 after sending an internal email that accused Google of silencing marginalised voices (Hao 2020).

It is clear that we need to be aware of AI ideology and acknowledge that AI ethics alone, despite their value and contributions, will not save the world. The other problem is about how we move from AI ethics to concrete policies. There is no roadmap for what exactly should be done, no precise course of action to be taken in policy development (Coeckelbergh 2020). It comes down to who has the capacity to influence the actions, beliefs or behaviour of others. Or who can influence what type of society we want, and what the role of technology such as AI should be in it. Ultimately, this is a question about power and who is in control to make decisions.

We Have to Talk about AI and Power

The problem of AI ideology and – more broadly – the question of whether we need AI and if so, what type of AI we need, illustrates why we need critical perspectives on AI. What do I mean by *critical*? The Frankfurt School has been pivotal in the development of critical thinking and theory. According to Max Horkheimer (2002), one of the leading figures of the Frankfurt School, critical theory distinguishes itself from traditional theory because of its focus on human emancipation. The goal of critical theory is to scrutinise and understand systems of domination and oppression and to look for ways of how to increase liberation and freedom.

If we make human emancipation central, we need to ask questions about AI and power. And this is exactly what is missing in AI policies and ethics: power. *Power* is a contested concept in social theory. In a pragmatic way, Wright (2010, 111) defines it as: 'the capacity of actors to accomplish things in the world'. This is a positive take on power, whereas a lot of definitions of power are negative – coercive power, preventing others to act in a certain way (Fuchs 2017). In addition to coercive power, Thompson (1995) also talks about economic power, political power and symbolic power. Economic power refers to how certain individuals and groups in society can accumulate resources for productive activity; political power is about the authority to coordinate individuals and their interaction; and symbolic power refers to meaning making and influencing the actions of others. AI ideology has raised issues of *symbolic* power, so I now turn to economic and political power in the context of AI.

We need to be aware that AI simultaneously refers to technical approaches, social practices and industrial infrastructures (Crawford 2018). The *technical approaches* are straightforward: these are computational systems that use data for training machine learning and deep learning algorithms (Alpaydin 2016). The other two elements need more clarification. The *social practices* of AI refer to the classification systems, developed by humans, which are behind the machine/deep learning algorithms and models. Political power asks who is involved in developing these classification systems and who decides what they

look like (Crawford 2018). Questions about inclusion and representation are inherently political questions. AI also refers to *industrial infrastructures*: the infrastructure does not only entail the possibilities of collecting vast amounts of data, but also the computational power needed to develop machine/deep learning models. Very few companies have simultaneously the computational power, access to data and AI expertise (human resources) at their disposal, which means that the economic power of these organisations is crucial for the development of AI and is highly concentrated (Dyer-Witheford, Kjøsen and Steinhoff 2019).

The Case for a Radical Democratisation of AI

Asking critical questions about AI with the objective to foster human emancipation requires us to investigate the political and economic power dynamics of AI. My point here is that we need to move beyond discussions of what beneficial AI means and what opportunities and risks exist in its development. We urgently need to think instead about what radical approaches to AI are and how we can enable them. Why *radical*, and what does that mean? *Radical* originates in the Latin word *radix*, which means *root* and that is why it has been popularised as *grasping things at the root*. Radical can mean many things but here I refer to it as in *radical politics* (Fenton 2016). Radical politics is characterised by its intention of transforming the fundamental principles of a political system or a society, often by making use of structural change or radical reform – change at the root.

A radical perspective to AI thus means we need to examine AI through the lens of power. Ultimately this comes down to the question of how AI is shifting power. This is about bringing real change for the better, disrupting power dynamics and avoiding an unequal power distribution. We could repeat (and slightly revise) William Gibson's (2003) seminal quote 'AI is already here; it's just not evenly distributed'. The question then remains: how can we redistribute power in AI?

My proposal is that if we want to establish AI that transforms society for the better and enables human emancipation, we need a radical democratisation of AI. This radical democratisation is necessary to avoid power inequalities, in other words, to avoid a situation whereby only a few organisations, whether governmental or corporate, have the economic and political power to decide what type of AI will be developed and what purposes it will serve.

This is vital in the data and AI sector, which is characterised by a strong tendency to establish monopolies. Network effects intensify competition between data platforms: the more users on their platform, the more valuable they become (Srnicek 2017). More data then also generates more users, which allows for the creation of better services. This is called a *data-feedback loop*. Data giants will therefore acquire competitors, which leads to a situation of

an oligopoly or even monopoly. This is even more crucial in the AI industry where few companies have access to data to train machine/deep learning algorithms, possess the computing power to deal with massive data sets and also to hire the AI talent that is necessary to build AI systems and applications (Dyer-Witheford et al. 2019).

So, what does a radical democratisation of AI actually mean? First, AI and the benefits it offers, should be *accessible to everyone*. Second, AI and the different services that are being developed should also *represent everyone*. Third and last, AI should be *beneficial to everyone*. These three principles are inspired by the late Erik Olin Wright's critique of capitalism. Wright (2019) proposes the principles of equality/fairness, democracy/freedom and community/solidarity as normative foundations for establishing a society that allows its members to live a decent life. In the following paragraphs, I briefly unpack these three guiding principles.

#Principle 1: AI Should Be Accessible to Everyone

This first principle proposes equal access to AI and the benefits it can offer. In a decent society, all persons should have broadly equal access to the advantages and possibilities being created by digital technologies such as AI. This means that we need to make sure that all groups in society have access to and can use AI. The egalitarian ideal is at the centre of nearly all concepts of social justice, including *data justice* (Taylor 2017), although there are different opinions about what it means exactly. An important nuance here is to distinguish between equal *access* and *opportunity*. The former is chosen over the latter as it 'is a sociologically more appropriate way of understanding the egalitarian ideal' (Wright 2019, 11). Given the current economic, social and environmental crisis we are living in, there should be particular attention to intergenerational and environmental justice. The first aspect points to the consequences of technological developments for the future generations, whereas the second aspect asks for attention for IT and sustainability. This is controversial as AI is both seen as a source of and solution for environmental degradation (Dauvergne 2020).

#Principle 2: AI Should Represent Everyone

The second principle is centred around democracy and inclusion. In a decent society, all members should have a say about what type of AI is being developed and what services are being offered. The production and implementation of AI must be democratised so that all groups in society are consulted and represented, avoiding exclusion. This element of democracy entails two aspects: everyone is involved and everyone is represented. The latter aspect highlights that when fairness fails, there is a risk of discrimination (Hoffmann 2019). The

history of AI is full of examples of how technology is being developed by (predominantly) white middle-class men, thereby excluding people of colour and minority communities. Wright (2019) also connects democracy and freedom in order to reflect the value of self-determination. In this sense, members of society should be given the possibility to participate meaningfully in decisions that affect their lives. As AI becomes more omnipresent, people should have a say about this. Principles such as fairness, accountability and transparency (ACM 2020) are key when we want technological development not only to represent the people but also guaranteeing control by the people to counterbalance the power of the state and corporations.

#Principle 3: AI Should Be Beneficial to Everyone

The third and last principle states that developments in AI should contribute to the well-being of everyone in society. This matches with Wright's (2019) ideas of community and solidarity, which are crucial because of their connection to human flourishing and of their role in fostering equality and democracy (see also principles 1 and 2). Central is the idea that if people cooperate, they can achieve more than if they compete, and cooperation also contributes to the well-being of all members of society. This means that AI development must be organised in such a way that all members of society are able to reap the benefits. Another aspect of this principle is the question about how to develop beneficial machines, in other words, how can we ensure that AI serves the objectives of humanity. Stuart Russell (2019, 11) states: 'machines are beneficial to the extent that their actions can be expected to achieve our objectives'. According to him, this is at the centre of the problem of control in AI and his interpretation focuses on the human–machine relationship, as part of being beneficial to everyone. Developing AI that is beneficial for everyone, thus includes thinking about how to create beneficial machines that serve humanity.

AI for everyone risks becoming yet another hype, if we let the tech giants take over the debate with their slogans such as *AI for social good*. What they are missing is a real vision of democratising technology because they fail to understand what *AI for everyone* really means: putting the human at the centre (Pasquale 2020). In one word, this is about power. If we are not talking about power, we are not talking about *AI for everyone*. Critical perspectives require us to talk about the human and society. By bringing together diverse critical contributions to the debate, this book presents one thing they have in common: the idea of putting society first.

Chapter Overview

Part 1: *AI – Humans vs. Machines* consists of four contributions. Andreas Kaplan (Chapter 2) goes deeper into the history and definition of AI and

elaborates on how humans and machines have to coexist in the age of AI. Wolfgang Hofkirchner (Chapter 3) continues the discussion about humans versus machines by analysing what Digital Humanism exactly entails. He proposes dialectical models in order to overcome the human–machine dualism. Jenna Ng (Chapter 4) adds to this discussion by elaborating on the rationalisation of AI and what this means for creativity. Dan McQuillan (Chapter 5) has a different take on humanism and proposes how people's councils for AI can serve solidarity and mutual aid in times of crisis.

Part 2: *Discourses and Myths About AI* is comprised of five chapters. Rainer Rehak (Chapter 6) stresses the importance but also limitations of metaphors when talking about AI and intelligent systems. Angela Daly, S. Kate Devitt and Monique Mann (Chapter 7) introduce and discuss their Good Data approach in order to overcome the limitations of AI ethics and governance. James Steinhoff (Chapter 8) critically analyses the social reconfiguration of AI and discusses the central questions about utility and feasibility. Benedetta Brevini (Chapter 9) analyses AI policies in Europe and unpacks some of the myths around AI that legitimate capitalism. Alkim Almila Akdag Salah (Chapter 10) reflects on how the discourses of artistic computational production have changed and how myths about AI need to be uncovered in this context.

Part 3: *AI Power and Inequalities* involves five contributions. Carrie O'Connell and Chad Van de Wiele (Chapter 11) revisit Wiener's cybernetic prediction as the theoretical foundation of AI and make a plea how we need to uncover the black box of what is behind prediction and simulation. Jernej A. Prodnik (Chapter 12) critically analyses algorithmic logic in digital capitalism, its characteristics and social consequences. Asvatha Babu and Saif Shahin (Chapter 13) investigate biometrics and biopolitics and apply their analysis to a case study of the ban on facial recognition in California. Rafael Grohmann and Willian Fernandes Araújo (Chapter 14) turn to a discussion of human labour that is behind global AI platforms and report about their empirical research on the Mechanical Turk in Brazil. Last, Lina Dencik (Chapter 15) also reflects on the relationship between labour and AI and proposes the concept of data justice unionism to rethink the governance of AI.

References

ACM. 2020. *ACM Code of Ethics and Professional Conduct*. Retrieved from: https://www.acm.org/code-of-ethics

Agrawal, A., Gans, J. and Goldfarb, A. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business School Publishing.

Alpaydin, E. 2016. *Machine Learning*. Cambridge, MA: MIT Press.

Asaro, P. 2018. What Is the Artificial Intelligence Arms Race Anyway? *I/S: A Journal of Law and Policy for the Information Society*, 15(1–2).

Atomium-EISMD. 2020. AI4People, Europe's First Global Forum on AI Ethics, Launches at the European Parliament. Retrieved from: https://www.eismd.eu/ai4people-europes-first-global-forum-ai-ethics-launches-at-the-european-parliament/

Berman, B.J. 1992. Artificial Intelligence and the Ideology of Capitalist Reconstruction. *AI & Society*, 6, 103–114.

Boden, M. 2016. *AI: Its Nature and Future*. Oxford: Oxford University Press.

Boddington, P. 2017. *Towards a Code of Ethics for Artificial Intelligence*. Cham: Springer.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Coeckelbergh, M. 2020. *AI Ethics*. Cambridge, MA and London: MIT Press.

Crawford, K. 2018. *The Politics of AI. Royal Society, You and AI*. Retrieved from: https://www.youtube.com/watch?v=HPopJb5aDyA

Dananjayan, S. and Raj, G.M. 2020. Artificial Intelligence During a Pandemic: The COVID-19 example. *The International Journal of Health Planning and Management*, 35(5), 1260–1262.

Dauvergne, P. 2020. *AI in the Wild. Sustainability in the Age of Artificial Intelligence*. Cambridge, MA: MIT Press.

Dobbe, R. and Whittaker, M. 2019. *AI and Climate Change: How They're Connected, and What We Can Do about It*. AI Now Institute. Retrieved from: https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c

Dyer-Witheford, N., Kjøsen, A.M. and Steinhoff, J. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. London: Pluto Press.

Elish, M.C. and boyd, d. 2018. Situating Methods in the Magic of Big Data and AI. *Communication Monographs*, 85(1), 57–80.

Elliott, A. 2019. *The Culture of AI. Everyday Life and the Digital Revolution*. London: Routledge.

European Commission. 2019. *Ethics Guidelines for Trustworthy AI*. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Fenton, N. 2016. *Digital, Political, Radical*. Cambridge: Polity Press.

Floridi, L., Cowls, J., Beltrametti, M. et al. 2018. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Mind and Machines*, 28, 689–707.

Ford, M. 2015. *The Rise of the Robots*. London: Oneworld Publications.

Fuchs, C. 2017. *Social Media: A Critical Introduction* (2nd Edition). London: SAGE.

Fuchs, C. 2018. *Digital Demagogue: Authoritarian Capitalism in the Age of Trump and Twitter*. London: Pluto.

Fuchs, C. 2020. *Marxism: Key Ideas in Media and Cultural Studies*. New York: Routledge.

Future of Life Institute. 2017. *Principles Developed in Conjunction with the 2017 Asilomar Conference on Beneficial AI*. Retrieved from: https://futureoflife.org/ai-principles

Gibson, W. 2003. The Future Is Already Here – It's Just Not Evenly Distributed. *The Economist*, 4 December 2003.

Google. 2020. *Advancing AI for Everyone*. Retrieved from: https://ai.google

Hao, K. 2020. *We Read the Paper that Forced Timnit Gebru Out of Google. Here's What it Says*. Retrieved from: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru

Hoffmann, A.L. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society*, 22(7), 900–915.

Horkheimer, M. 2002. *Critical Theory*. New York: Continuum.

IEEE. 2020. *IEEE Ethics in Action in Autonomous and Intelligent Systems*. Retrieved from: https://ethicsinaction.ieee.org/#resources

Jobin, A., Ienca, M. and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1, 389–399.

Korinek, A. and Stiglitz, J.E. 2017. *Artificial Intelligence and Its Implications for Income Distribution and Unemployment*. Working Paper 24174, National Bureau of Economic Research. Retrieved from: https://www.nber.org/papers/w24174

Lee, K.-F. 2018. *AI Superpowers: China, Silicon Valley and the New World Order*. Boston, MA: Houghton Mifflin.

McAfee, A. and Brynjolfsson, E. 2017. *Machine, Platform, Crowd: Harnessing the Digital Revolution*. New York: WW Norton & Company.

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. 1955/2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, 27(4), Winter 2006.

Microsoft. 2020. *AI for Good*. Retrieved from: https://www.microsoft.com/en-us/ai/ai-for-good

Moor, J.H. 1985. What is Computer Ethics? *Metaphilosphy*, 16(4).

Morozov, E. 2013. *To Save Everything, Click Here*. London: Penguin Books.

Pasquale, F. 2020. *New Laws of Robotics. Defending Human Expertise in the Age of AI*. Cambridge, MA: Harvard University Press.

Russell, S. 2019. *Human Incompatible. AI and the Problem of Control*. London: Allen Lane.

Russell, S. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*. Harlow: Pearson Education Limited.

Sandoval, M. 2014. *From Corporate to Social Media. Critical Perspectives on Corporate Social Responsibility in Media and Communication Industries*. London and New York: Routledge.

Srnicek, N. 2017. *Platform Capitalism*. Cambridge: Polity Press.

State Council of China. 2017. *New Generation of Artificial Intelligence Development Plan* (translated by F. Sapio, W. Chen and A. Lo). Retrieved from: https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf

Taylor, L. 2017. What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally. *Big Data & Society*, 4(2), 1–14.

Thompson, J. 1995. *The Media and Modernity. A Social Theory of the Media*. Cambridge: Polity Press.

Turing, A.M. 1950. Computing Machinery and Intelligence. *Mind*, LIX(236), 433–460.

Verdegem, P. 2021. Social Media Industries and the Rise of the Platform. In P. McDonald (Ed.) *Routledge Companion to Media Industries*. New York: Routledge.

Villani, C. 2018. *AI for Humanity. French Strategy for Artificial Intelligence*. Retrieved from: https://www.aiforhumanity.fr/en/

Vincent, J. 2017. *Putin Says the Nation that Leads in AI 'Will be the Ruler of the World'*. Retrieved from: https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world

Wagner, B. 2018. Ethics as an Escape from Regulation. From 'Ethics-Washing' to 'Ethics-Shopping'? In E. Bayamlioglu, I. Baraliuc, L.A.W. Janssens and M. Hildebrandt (Eds.) *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, pp. 84–89. Amsterdam: Amsterdam University Press.

White House. 2019. *Artificial Intelligence for the American People*. Retrieved from: https://www.whitehouse.gov/ai

Wright, E.O. 2010. *Envisioning Real Utopias*. London and New York: Verso.

Wright, E.O. 2019. *How to be an Anti-Capitalist in the 21st Century?* London: Verso.

Zuiderveen Borgesius, F. 2018. *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. Strasbourg: Council of Europe, Directorate General of Democracy. Retrieved from: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

PART I

AI – Humans vs. Machines

CHAPTER 2

Artificial Intelligence (AI): When Humans and Machines Might Have to Coexist

Andreas Kaplan

Introduction

Artificial intelligence (AI), defined as a 'system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation' (Kaplan and Haenlein 2019, 17), will likely have a deep impact on human beings and society at large. The recent COVID-19 pandemic has particularly accelerated and accentuated society's digitalisation and strongly influences the future relationship between human beings and AI-driven machines (Haenlein and Kaplan 2021).

Various opinions and viewpoints on the future altered by advances in AI exist, ranging from horror scenarios as stated by Tesla CEO Elon Musk, to utopian scenarios like the vision of Google Chief Engineer Raymond Kurzweil. While Musk fears that AI might lead to nothing less than a third world war, Kurzweil believes that AI will enhance humans instead of replacing them. Expressing these opposing views, in 2018, theoretical physicist Stephen Hawking proclaimed that AI can 'either be the best, or the worst thing, ever to happen to humanity'.

How to cite this book chapter:
Kaplan, A. 2021. Artificial Intelligence (AI): When Humans and Machines Might Have to Coexist. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 21–32. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.b.

Clearly, humans will need to coexist with machines. Jobs traditionally done by humans will be shifted towards AI systems. Artificial intelligence is already able to translate languages, diagnose illnesses, assist in retail (Kaplan 2020c), and the like – in several cases, better than the human workforce. Human jobs might be created in the future that are unimaginable now, similar to the fact that nobody really predicted the job of mobile app designers just a few years ago.

In this world, AI would rather be augmenting and complementing – rather than replacing – humans in their work. In the pessimistic case, i.e., massive unemployment, ideas such as universal basic income are already being discussed. Fundamental philosophical questions would need to be answered surrounding life for humans when most of our work is done by AI systems. In any case, the State will certainly have to come up with a set of rules governing this human < > machine coexistence and interdependence. Society overall is thus challenged.

This chapter has a look at artificial intelligence, its history and its evolutionary stages. Furthermore, what challenges might arise in the future when humans will have to learn to live among machines and robots will be discussed. This will be done by analysing challenges concerning algorithms and organisations, challenges with respect to (un)employment, and looking at democracy and freedom potentially jeopardised due to AI progress.

Artificial Intelligence: Definition and Classification

Artificial intelligence is a rather fuzzy concept, and quite difficult to define. At least two reasons can be proposed for the difficulty in formulating a definition therefore: firstly, it is not easy to find a clear definition for what intelligence in general is, as it depends largely upon the context. Thus intelligence is described in several different ways such as the capacity for learning, reasoning, planning, understanding, critical thinking, creativity, and last but not least, problem solving.

Secondly, artificial intelligence is a moving target: advances previously considered to AI with time will not be considered as such as soon as we get used to them. This phenomenon is known as the *AI effect*. As McCordick (2004, 204) formulated it: 'It's part of the history of the field of artificial intelligence that every time somebody figured out how to make a computer do something – play good checkers, solve simple but relatively informal problems – there was chorus of critics to say, "that's not thinking".' Or as Rodney Brooks, MIT's Artificial Intelligence Laboratory director, explains, 'Every time we figure out a piece of it, it stops being magical; we say, "Oh, that's just a computation", and will not count as artificial intelligence any longer' (Kahn 2002).

One of the prevailing definitions of artificial intelligence, as aforementioned, characterises AI as 'a system's ability to correctly interpret external data, to learn from such data and to use those learnings to achieve specific goals and tasks through flexible adaptation' (Kaplan and Haenlein 2019, 17). Several further

definitions exist and experts disagree on how to best characterise artificial intelligence. By analysing different AI definitions, Russell and Norvig (2016), e.g., concluded that there are four main approaches for defining AI, i.e., see it as systems that (1) think like humans, (2) act like humans, (3) think rationally and (4) act rationally.

Often terms such as big data, machine learning or the Internet-of-Things (IoT) are incorrectly applied as synonyms for artificial intelligence, yet they are indeed differing concepts and terms. An AI-driven system needs big data from which to learn, which essentially are 'datasets made up by huge quantities (volume) of frequently updated data (velocity) in various formats, such as numeric, textual or images/videos (variety)' (Kaplan and Haenlein 2019, 17). Again, a variety of different definitions for big data exists: while one group of them focuses on what big data is, a second group stresses what big data actually does (Gandomi and Haider 2015). Such big data sets can derive from an organisation's internal databases, third-party data or social media applications (Kaplan 2012; Kaplan and Haenlein 2010b).

Another possibility for obtaining big data is via the Internet-of-Things (Krotov 2017; Saarikko, Westergren and Blomquist 2017), which basically is an extension of internet connectivity into physical devices and everyday objects such as a refrigerator or a heater, equipped with sensors and software to collect and exchange data.

Machine learning, simply put, is 'methods that help computers learn without being explicitly programmed' (Kaplan and Haenlein 2019, 17), and is applied in order to identify underlying patterns within the big data, and as such is an essential element of artificial intelligence. A more elaborated definition comes from Mitchell (1997, 2) stating 'A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.' AI is much broader than machine learning, as it additionally comprises such abilities as the perception of data (e.g., voice/image recognition, natural language processing, etc.) or the control and movement of objects (robotics or cybernetics).

Artificial intelligence can be classified into three types of systems: analytical, human-inspired and humanised (Kaplan and Haenlein 2019). *Analytical* AI contains characteristics consistent with cognitive intelligence only: generating cognitive representation of the world and using learning based on past experience to inform future decisions. *Human-inspired* AI contains elements of cognitive and emotional intelligence: understanding human emotions, in addition to cognitive elements, and considering them in their decision-making. *Humanised* AI contains characteristics of all types of competencies (i.e., cognitive, emotional and social intelligence), is able to be self-conscious, and is self-aware in interactions with others.

A robot driven by analytical artificial intelligence would be capable of answering queries concerning restaurant recommendations based on certain

objective characteristics. Human-inspired AI robots could additionally read a human's emotional state via facial recognition or tone of voice, and adapt its suggestions, e.g., a human who appears sad or depressed would not enjoy a restaurant with a lively atmosphere, whereas a happy human might totally enjoy such an environment. Finally, a humanised robot would understand when it was appropriate for it to offer to accompany the human or whenever this would not be appreciated, e.g., a couple insanely in love who would rather spend the time in intimate togetherness.

Finally, we must distinguish AI on the lower spectrum from so-called expert systems, often wrongly associated with artificial intelligence, as well as on the higher spectrum from skills that remain only possible for human beings: Expert systems are 'collections of rules programmed by humans in the form of if > then statements' (Kaplan and Haenlein 2019, p. 18). As these systems lack the ability to learn autonomously from external data, they should definitely not be counted as AI. Expert systems reconstruct human intelligence in a top-down manner (also called the knowledge-based or symbolic approach), considering that it can be codified as a set of predefined rules. In contrast, AI applies a bottom-up approach (also called the behaviour-based or connectionist approach) and imitates a brain's set-up (e.g., through neural networks) by using large quantities of data to infer knowledge independently.

The question that arises is what will remain human in the future and what cannot be imitated by AI systems, which is quite a tough question to answer. Most likely, humans will always have exclusivity when it comes to artistic creativity, Albert Einstein having pointed out that 'creativity is intelligence having fun'. Currently, it seems very improbable that AI systems will be able to be truly creative. But then again, the question is what exactly true creativity is, and who will be the judge of it?

Artificial Intelligence: History and Evolution

To structure AI's history, we'll use an analogy of the four seasons: spring, summer, autumn and winter (Haenlein and Kaplan 2019). AI's birth period, i.e., spring, took place both in fiction as well as non-fiction. Regarding the former, Isaac Asimov, an American writer and professor of biochemistry at Boston University, published 'Runaround', a story revolving around an AI-driven robot, in 1942. In this story, Asimov's (1950, 40) three laws of robotics explicitly appear for the first time:

1. 'A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.'

These three laws already hint at the difficulty of humans and robots coexisting. In any case, the robot in Asimov's story freezes in a loop of repetitive behaviour, as it doesn't find a solution for obeying laws 2 and 3 at the same time. 'Runaround' is therefore a cornerstone in the history of artificial intelligence, as it inspired generations of academics and researchers in the domain of AI.

Regarding the real world, we can refer to computer scientist Alan Turing's seminal paper 'Computing Machinery and Intelligence', published in 1950. Therein, Turing describes what now is known as the Turing test, or a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human. AI spring's climax can be pinpointed to the 1956, when Marvin Minsky and John McCarthy organised the Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) at Dartmouth College. It was at this workshop that the term *artificial intelligence* was coined.

After spring, there followed a couple of hot AI summers and very cold AI winters. While AI summers were characterised by huge enthusiasm and financing of AI, winters were marked by reduced funding and interest in artificial intelligence research. The first summer period lasted nearly 20 years. One of its successes was certainly ELIZA: Developed in 1966 by German-American computer scientist Joseph Weizenbaum, a professor at MIT, this computer program was so good at conversing with a human being that it appeared to pass the aforementioned Turing test.

General hype around AI and its development followed. However, this hype was soon replaced by disappointment and disenchantment. AI winter somehow had already begun when Marvin Minsky supposedly still contended that artificial intelligence could attain a human being's general average intelligence within three to eight years from that moment (Darrach 1970). As we all know, this did not occur; AI funding was heavily reduced and another AI summer did not happen until the 1980s, when the Japanese government decided to massively invest in AI and consequently the US DARPA followed. Success again was scarce, and summer was again followed by another cold winter.

We might have reached AI's autumn, completing the four seasons of artificial intelligence (Haenlein and Kaplan 2019), as a result of computational strength having constantly increased over recent years, rendering deep learning and artificial neural networks possible (Libai et al. 2020). This new era of AI is said to have begun in 2015 when AlphaGo, a computer program designed by Google, beat a (human) world champion in the Chinese board game Go. This event made the news around the world, and regenerated hype around the domain of artificial intelligence.

This hype might continue for quite some time, as we are currently only experiencing so-called first-generation AI applications, usually referred to as artificial narrow intelligence (ANI). Within such systems, AI is only applied to very specific tasks such as choosing which news items it will tell an individual during his or her morning before-work routine based on the individual's intellectual preferences.

Second-generation AI applications will be able to plan, solve and reason problems independently, even for actions for which they have not been programmed initially. Such artificial general intelligence (AGI) will thus be able to broaden its horizons autonomously, entering new areas and domains. For example, an AGI-powered system could, on top of conveying news headlines during one's morning routine, also learn to make coffee for the aforementioned individual preparing for work.

Finally, we might potentially even experience artificial super intelligence (ASI), the third generation of AI. Such truly self-conscious and self-aware AI systems, outperforming humans in (nearly) all domains, capable of general wisdom, scientific creativity and social skills, could render human beings redundant. As such, in our above example, the individual would not need to prepare for work anymore, as this could be done entirely by the ASI-powered machine or robot (Kaplan and Haenlein 2019). For a detailed discussion on the evolution of AI systems, we refer to Huang and Rust (2018).

Artificial Intelligence: Machines and Humans

In the future, artificial intelligence will raise several challenges, and humans will have to learn to coexist with machines and robots. Pushed by the global COVID-19 health crisis, it is clear that AI will deeply impact societies around the world (Kaplan 2021). We will discuss some of these questions, looking at challenges in terms of algorithms and individual organisations; the employment market; and last but not least, democracy and human freedom potentially at stake due to advances in AI.

About Algorithms and Organisations

When machines and humans coexist, it is important that both do what they are good at. As an illustration, let's have a look at a study by researchers from MIT's Computer Science and Artificial Intelligence Laboratory in cooperation with the machine-learning startup PatternEx (Conner-Simons 2016). AI systems and humans scored far better in identifying cyber-attacks when collaborating than when trying to do so separately. While the AI systems could crawl through enormous quantities of big data, humans were better at detecting anomalies, playing those back into the system. This iterative and collaborative approach was optimal.

Also, humans are better in behaving ethically and morally, while algorithms have problems doing so, as the notion of ethics and morals is difficult to program. Machines, however, are better at, e.g., utilitarian, repetitive tasks. While most humans would not consciously discriminate another individual for gender, sexual orientation, social background, or race, machines, not having a

conscience, are more likely to be biased, essentially because the data on which they were trained was biased. A study by Wilson, Hoffman and Morgenstern (2020) illustrates that several decision-support systems applied by judges may be racially biased (as a result of past rulings); and self-driving cars better detected lighter skin than darker tones, since their algorithm was trained using pictures among which were few people of colour.

Regulation and guidance is definitely needed in order to avoid such bias, to establish a good foundation for machine < > human collaboration. The development of specific requirements with respect to the testing and training of AI is likely the preferred approach, as opposed to regulating artificial intelligence itself. In addition, we could require AI warranties, consistent with safety testing in the case of physical goods. Thus, AI regulation could be stable over time even if the actual aspects of AI technology change (Kaplan and Haenlein 2020).

About (Un)Employment

A tough challenge when human beings coexist with machines might be the evolution of the job market. Already, automation in manufacturing has led to a significant decrease in blue-collar jobs; advances in AI could lead to a similar decrease in white-collar jobs. AI systems already outperform medications in the identification of skin cancer and other tasks (Welch 2018).

For the moment, it appears that the time gain through AI's application is used for other tasks within the job, and does not necessarily lead to a human being's replacement. The Swedish bank, SEB, e.g., developed AIDA, an AI-driven virtual assistant responding to a vast range of customers' queries, such as how to make overseas payments or how to proceed when opening a bank account. AIDA is even capable of detecting a customer's mood by the tone of her or his voice and adapting its recommendations and suggestions thereto. In around 30% of situations, AIDA is not able to respond or help. In this case, the customer is transferred to a human. AIDA's implementation freed up human employees' time, which they then use for more complex demands, i.e., the 30% that exceeded AIDA's limitations.

A study by Wilson and Daugherty (2018, 117) suggested that it is in companies' interest not to replace employees with AI, as this would not be a long-term strategy. Looking at 1,500 corporations, they identified the best improvements in performance when machines and human beings work together, and concluded: 'Through such collaborative intelligence, humans and AI actively enhance each other's complementary strengths: the leadership, teamwork, creativity, and social skills of the former, and the speed, scalability, and quantitative capabilities of the latter' (Wilson and Daugherty 2018).

However, with advances in artificial intelligence, machines improve, and might indeed replace humans in their jobs. It is uncertain that enough new jobs at the right skill levels will evolve for everybody, similar to previous shifts in

job markets such as the Industrial Revolution. The demanded skill level might just be too high for all human beings to be able to find a job not yet done by a machine. Or, there just might not be enough jobs left, as more jobs are replaced by machines than are newly created. Massive unemployment would result.

In the short to medium term, regulation could certainly help to avoid mass unemployment, at least for a transitional period. Examples are the requirement for companies to spend a certain amount of their budgets saved via the help of AI on training their workers for higher-skilled jobs; or the restriction of the number of hours worked per day in order to distribute the available work across the entire population. However, in the longer run, if machines replace humans as workers, the idea of a universal basic income will be put back on the table. This would trigger a series of fundamental philosophical but also religious debates: questions such as the purpose of life, how to feel useful and what to strive for, are some issues for which society would have to find answers. Ethics and education will play an important role in order to tackle these societal challenges and questions (Kaplan 2020a).

About Democracy and Freedom

Finally, AI progress could represent nothing less than a danger to peace and democracy (Kaplan 2020b). There are at least two ways in which artificial intelligence might constitute a threat to democracy and its mechanisms, endangering the peaceful coexistence of humans and machines: supervision and manipulation.

Using the example of China, we will provide an illustration as to how far the possibilities of artificial intelligence reach with respect to control and supervision. AI is largely embraced by the Chinese government, which uses it to track and monitor its citizens and inhabitants. For each individual, the Chinese government calculates a so-called 'social credit score' based on (big) data coming from various different sources such as health and tax records, social media activity, purchasing behaviour, criminal records and so forth. The system also uses facial recognition and images of the 200 million surveillance cameras mounted across the country for data collection and respective score calculation. Good behaviour such as volunteering at an orphanage leads to higher scores; bad behaviour such as littering leads to lower scores. In order to fulfil the score's aim, i.e., to encourage good behaviour and citizenship, bad scores result in punishments such as not being eligible for bank loans, not being allowed to fly or not being hired by public agencies (Marr 2019).

In addition to control possibilities, artificial intelligence also allows for manipulation, as we now constantly experience with the dissemination of fake news and disinformation on the various social media platforms (Deighton et al. 2011; Kaplan and Haenlein 2010a; Kaplan 2018). Especially in election campaigns, social media are heavily used to manipulate voters. For example, in the

final three months of the 2016 US presidential election, the top 20 false news items on only one social medium – Facebook – led to more comments, likes and shares than did the 20 most influential news stories from approximately 20 major actors in the news sector together (including such outlets as the *New York Times* and the *Washington Post*; Silverman 2016).

This alone gives enough food for thought regarding the manipulative power of AI-based systems. And yet, the next, bigger thing is just around the corner: deepfakes, which are 'AI-based technology used to produce or alter audio or video content so that it presents something that did not, in fact, occur' (Kaplan 2020b). This technology allows inserting words in audio or even video format in an individual's speech that s/he never actually uttered. Thus, one could make a seemingly authentic video of the Pope stating that monogamy is overrated and that everybody should have open relationships. What this means for future elections and other phenomena is indeed difficult to imagine.

The above two examples clearly show that artificial intelligence potentially leads to issues that do not stop at countries' borders, with Russia having knowingly been deeply involved in the aforementioned 2016 US presidential election. Regulation that applies to some countries only will most likely be ineffective in governing the coexistence of humans and machines. Intensive international coordination and cooperation in regulation is clearly needed, whenever feasible.

Such international cooperation might be a challenge. While China and the United States are considered as the AI superpowers, they are less known for their implementation of AI regulations (Kaplan 2020a). The development of regulation as well as ethics guidelines falls rather within the expertise of the European Union. The EU, however, has far less influence in the actual development and elaboration of artificial intelligence. Nevertheless, spill-over effects are possible. The EU's strict General Data Protection Regulation (GDPR), effective since May 2018, applies to any corporation that markets products to EU residents, regardless of its location. Thus, GDPR influences data protection requirements worldwide. As such, the California Consumer Privacy Act (CCPA), which governs the most populous US state's data protection since January 2020, is recurrently referred to as California's GDPR. Government regulation is certainly a necessary step. Most likely, whenever society realises the topic's importance, companies will feel obliged to go into the direction of self-regulation, similarly to the worldwide impact of citizens' increased commitment and desire for sustainability and a stronger protection of the environment.

Conclusion: Only Time Will Tell

In this chapter, we introduced the concept of artificial intelligence and how it differs from related concepts such as big data, the Internet-of-Things, and machine learning. We also surveyed AI's history and evolution before discussing the relationship between humans and machines from various angles.

Future research will be needed to address the various challenges with regards to the development of artificial intelligence. Which formal method can be used to test for algorithmic bias? Can we identify simple to use measures to assess bias, similar to the way we assess reliability and validity? What is the best way to bridge (deep) learning and privacy? Should learning be conducted on the user side (with algorithms requiring new data)? Or should data be transferred to a trusted intermediary who performs the analysis on behalf of firms? Do users need to be compensated in one way or another for data or resources provided? Moreover, how can the refusal to share data lead to biases in the data available for learning? Which data sources can and should be used for algorithmic learning? Are there certain types of data that should be 'off-limits'? What role will interdisciplinary AI teams play in establishing coexistence between humans and machines? To mention just a few of the potential future research questions, which, in the light of the unprecedented global COVID-19 pandemic and its acceleration of society's digitalisation, become of vital importance.

At least for the moment, it looks as if AI-driven machines will enhance human work instead of replacing it. This is also the opinion of John Kelly, vice president of IBM, who stated, 'Man and machine working together always beat or make a better decision than a man or a machine independently' (Waytz 2019). Moreover, according to a recent Accenture study, more than 60% of employees believe that AI will have a beneficial impact on their work and jobs (Shook and Knickrehm 2017).

COVID-19 impressively showed that artificial intelligence has played an important role in tackling this unprecedented health crisis on a global level. As such, researchers worldwide made use of AI to efficiently identify potentially infected humans, analyse the virus, test possible treatments and therapies, and more generally to find strategies to fight the pandemic. AJ Venkatakrishnan, e.g., applying AI, discovered that a mutation of the original virus would mimic a protein which the human body uses to regulate its fluid and salt equilibrium (Cha 2020). However, the application of artificial intelligence also showed its connected impact on individuals' daily lives as well as on such questions as data security and privacy. Regulation for the human-machine entanglement is clearly needed.

Furthermore, an example at Mercedes-Benz clearly shows that the replacement of the human workforce is still not as easy as sometimes claimed, and that indeed, currently, human < > machine coexistence is here. Normally, in the automobile manufacturing process, robots and automation are common. However, Mercedes-Benz key accounts increasingly demand more customisation – which the robots were not able to deliver.

Therefore, the German automobile giant decided to replace the fully automated process with 'cobots', or collaborative robots, which are robots designed to physically interact with human beings in a shared workspace. These cobots are controlled by humans, and are to be considered an extension of the human's body, facilitating the carrying and moving of heavy car parts. This form of human < > machine collaboration enables an efficient and productive

customization process, responding in real time to customers' precise choices with regard to leather seats, tyre caps, and so forth.

As in the automotive sector, AI will certainly trigger changes and evolutions in the upcoming years in many sectors. Without a crystal ball, it will be difficult to know where and how the coexistence of humans and machines will evolve. However, it is crystal clear that the business world (and society at large) will need to constantly adapt to advances in AI in order to keep up with the pace (Kaplan and Haenlein 2020), or, to quote Benjamin Franklin: 'When you're finished changing, you're finished.'

References

Asimov, I. 1950. Runaround in *I, Robot: The Isaac Asimov Collection*. New York: Doubleday.

Cha, A. 2020. Artificial Intelligence and Covid-19: Can the Machines Save Us? *Wall Street Journal*, 1 November 2020.

Conner-Simons, A. 2016. System predicts 85% of cyber-attacks using input from human experts. *MIT News*. Retrieved from: http://news.mit.edu/2016/ai-system-predicts-85-percent-cyber-attacks-using-input-human-experts-0418.

Darrach, B. 1970. Meet Shaky, the first electronic person, *Life Magazine*, 69(21), pp. 58–68.

Deighton, J., Fader, P., Haenlein, M., Kaplan, A.M., Libei, B. and Muller, E. 2011. Médias Sociaux et Entreprise, une Route Pleine de Défis. *Recherche et Applications en Marketing*, 26(3), 117–124.

Gandomi, A. and Haider, M. 2015. Beyond the hype: Big Data Concepts, Methods, and Analytics. *International Journal of Information Management*, 35(2), 137–144.

Haenlein, M. and Kaplan, A. 2019. A Brief History of AI: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, 61(4), 5-14.

Haenlein, M. and Kaplan, A. 2021. Artificial Intelligence And Robotics: Shaking Up the Business World and Society at Large. *Journal of Business Research*, 124, 405–407.

Huang, M.-H. and Rust, R.T. 2018. Artificial Intelligence in Service. *Journal of Service Research*, 21(2), 155–172.

Kahn, J. 2002. It's alive. *Wired*, 1 March.

Kaplan, A. 2012. If you Love Something, Let it Go Mobile: Mobile Marketing and Mobile Social Media 4×4. *Business Horizons*, 55(2), 129–139.

Kaplan, A. 2018. Social Media. In: B. Warf (Ed.), *The Sage Encyclopedia of the Internet*. London: Sage Publications Ltd.

Kaplan, A. 2020a. Artificial Intelligence: Emphasis on Ethics and Education. *International Journal of Swarm Intelligence and Evolutionary Computation*, 9(3).

Kaplan, A. 2020b. Artificial Intelligence, Social Media, and Fake News: Is This the End of Democracy? In A.A. Gül, Y.D. Ertürk and P. Elmer (Eds.), *Digital Transformation in Communication and Media Studies*. Istanbul: Istanbul University Press.

Kaplan, A. 2020c. Retailing and the Ethical Challenges and Dilemmas behind Artificial Intelligence. In E. Pantano (Ed.), *Retail Futures: The Good, the Bad, and the Ugly of the Digital Transformation*. Bingley: Emerald Publishing.

Kaplan, A. 2021. *Higher Education at the Crossroads of Disruption: The University of the 21st Century, Great Debates in Higher Education*. Bingley: Emerald Publishing.

Kaplan, A. and Haenlein, M. 2010a Uitdagingen en kansen rond social media. *Management Executive*, 8(3), 18–19.

Kaplan, A. and Haenlein, M. 2010b. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68.

Kaplan, A. and Haenlein, M. 2019. Siri, Siri in my hand, who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25.

Kaplan, A. and Haenlein, M. 2020. Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, 63(2).

Krotov, V. 2017. The Internet of Things and New Business Opportunities, *Business Horizons*, 60(6), 831–841.

Libai, B., Bart, Y., Gensler, S., Hofacker, C., Kaplan, A., Köttenheinrich, K. and Kroll, E. 2020. A brave new world? On AI and the management of customer relationships. *Journal of Interactive Marketing*, 51(C), 44–56.

Marr, B. 2019. Chinese social credit score: Utopian Big Data – Bliss? Or black mirror on steroids? *Forbes*, 21 January.

McCordick, P. 2004. *Machines Who Think* (2nd ed.). Natick, MA: A. K. Peters, Ltd.

Mitchell, T. 1997. *Machine Learning*. New York: McGraw Hill.

Russell, S.J. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*. Malaysia: Pearson Education Limited.

Saarikko, T., Westergren, U. H., and Blomquist, T. 2017. The Internet of Things: Are You Ready for What's Coming? *Business Horizons*, 60(5), 667–676.

Shook, E. and Knickrehm, M. 2017. *Harnessing Revolution: Creating the Future Workforce*. Accenture Strategy.

Silverman, C. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *Buzzfeed*, 16 November.

Waytz, A. 2019. How humans and machines can live and work together. *CNN Business Perspectives*, 1 May.

Welch, A. 2018. AI better than dermatologists at detecting skin cancer, study finds. *CBS News*, 29 May.

Wilson, B., Hoffman, J. and Morgenstern, J. 2020. Predictive Inquiry in Object Detection. Working paper.

Wilson, J. and Daugherty, P. 2018. Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review,* 1 July.

CHAPTER 3

Digital Humanism: Epistemological, Ontological and Praxiological Foundations

Wolfgang Hofkirchner

Introduction

It seems a common agreement that due to certain progress made in Artificial Intelligence (AI) and related fields mankind is facing a blurring of the human and the machine such that humanism is put under pressure. Is humanism outdated and can it be renounced? Or does it only need an update? And if so, an update in which direction?

There is discussion abound with pros and cons concerning technological, military, sociological and philosophical aspects of AI, Trans- and Post Humanism (Hofkirchner and Kreowski 2020). And there is a candidate for updating humanism – Digital Humanism.

This term popped up in a Gartner Special Report published in April 2015. The report had the title 'Digital Business: Digital Humanism Makes People Better, Not Technology Better' and its summary makes clear what Digital Humanism was supposed to be about and what it is was not supposed to be about: 'Digital humanism is the recognition that digital business revolves around people, not technology. CIOs and business leaders who recognise that digital business revolves around people's value will see employee capabilities translate

How to cite this book chapter:
Hofkirchner, W. 2021. Digital Humanism: Epistemological, Ontological and Praxiological Foundations. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 33–47. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.c.

into product, service and market gains.' The term did not refer to humanism as a philosophical tradition.

This is in stark contrast to the intentions of German philosopher and former minister Julian Nida-Rümelin who had used the term for a long time in lectures before he published, together with Nathalie Weidenfeld, a book with the title 'Digitaler Humanismus' (2018), for which the authors received the Bruno Kreisky Prize from the Karl-Renner-Institut, Wien. The German term inspired Hannes Werthner, the then Dean of the Faculty of Informatics at the Vienna University of Technology (TU Wien), to translate it into English when he convened a workshop in April 2019 that ended with a manifesto – the Vienna Manifesto on Digital Humanism.

> This manifesto is a call to deliberate and to act on current and future technological development. We encourage our academic communities, as well as industrial leaders, politicians, policy makers, and professional societies all around the globe, to actively participate in policy formation. Our demands are the result of an emerging process that unites scientists and practitioners across fields and topics, brought together by concerns and hopes for the future. We are aware of our joint responsibility for the current situation and the future – both as professionals and citizens.
>
> …
>
> We must shape technologies in accordance with human values and needs, instead of allowing technologies to shape humans. Our task is not only to rein in the downsides of information and communication technologies, but to encourage human-centered innovation. We call for a Digital Humanism that describes, analyzes, and, most importantly, influences the complex interplay of technology and humankind, for a better society and life, fully respecting universal human rights.

Given these quotations from the manifesto (Vienna Manifesto on Digital Humanism n.d.), Digital Humanism, meaning an update of humanism – of the image of man – in the age of digitalisation, promises to become a label for an answer to the questions raised above in a direction worth supporting, a direction not technology-driven but aiming at promoting a humane digitalisation.

This chapter at hand intends to contribute to philosophical, in particular, philosophy of science aspects such as praxio-onto-epistemology developed from the author elsewhere (Hofkirchner 2013), as sound foundations for such an updated humanism. It aims at clarifying the following problem:

How can a relation between human and machine be established in thinking and acting such that fallacies in theorising are avoided?

There are three ways of framing, modelling and designing the human and the machine, in particular, computer, cyber technology, digitalisation, in relation. One way is conflation – the false assertion of identity of what is different. Another way is the disconnection – the false assertion of a difference of what

Table 3.1: Frames, models and designs in the perspective of conflations, disconnections and combinations.

	Conflations		**Disconnections**			**Combination**
	Anthropo-morphism	Techno-morphism	Anthropo-centrism	Techno-centrism	Man-machine-hybridity	TechnoSocial Systemism
Frames	Cross-disciplinary		Mono- and multi-/inter-disciplinary			Transdisci-plinary
	Sociological colonisation	Technologi-cal takeover	Sociologism	Techno-logism	Methods mix	Systemic complements
Models	Monistic		Dualistic			Dialectical
	Anima	Mechanism	Pride of creation	Post-human	Man-machine-hybrids	Systems of systems
Design	Assimilative		Segregative			Integrative
	techno sapiens	homo deus	Supremacy	Singulari-tarianism	Man-machine-hybridisation	TA and design loop

is identical. And the last but not least way is the combination – the exercise to find out what is identical (what do both sides have in common though they might differ in some respects) and what is different (though they might have something in common). This is the only way with the prospect of transgressing falsehood.

The next three sections discuss these three ways in more detail. Frames, models and designs are dealt with. They refer to epistemological, ontological and praxiological issues respectively, (see Table 3.1).

Conflations

It is conflation if what is widely known as anthropomorphism is the case – the assertion of a human property in a realm where it is not an essential property. But there is also a second kind of conflation – the assertion of a machine property in a realm where it is not an essential property, which might, in analogy to the term anthropomorphism, be labelled technomorphism. Both kinds of conflation should not be conflated. They belong to different ways of thinking and acting and yield different results. Anthropomorphism is based upon a projection, while technomorphism is based upon a reduction. A projection projects higher complexity onto lower complexity so as to simulate higher complexity, while a reduction reduces higher complexity to lower complexity so as to simulate lower complexity. In the first case, you have an upgrading of complexity, whereas, in the second case, you have a downgrading.

Let's now turn to the discussion of how the anthropomorphic and technomorphic conflations work when framing, modelling and designing the relation of human and machine, one by one.

Cross-disciplinary Frames

Both anthropomorphism and technomorphism claim to use a common epistemology, a general frame of investigation for both human and machine.

But in the case of anthropomorphism, that frame is different from the technomorphic frame. Anthropomorphism extends the frame normally used in social sciences and humanities to information technology. It does so on the underlying assumption that those frames that are apt for social phenomena are also apt to investigate phenomena that are technical. That is, it looks upon technical phenomena as if they were social ones and in doing so it carries over to them expectations that they would show what social phenomena are showing. Thus, anthropomorphism is open to apply the term intelligence when speaking of artificial phenomena that shall be compared with human intelligence. Attempts to establish electronic personhoods for AI applications are examples of our inclination to anthropomorphising.

In the case of technomorphism, the situation is reversed. Methodologies that are usually built for technological research cover social phenomena. Thus, they convey expectations of technicality when applied in inquiries into social phenomena. Social phenomena are deemed engineerable. Human intelligence can be researched as if a phenomenon of an artefact. The human brain project of the EU pertains to this kind of fallacy.

In any case, the respective frame cuts across social as well as technological phenomena. The different disciplines of science are conflated – either to a social science take of technical phenomena or a technological take of social phenomena.

The current dominant approaches in social, human and arts research, on the one hand, and in natural science and technology, on the other hand, are still suffering from the divide between the two cultures as batptised by C. P. Snow (1998) in the last century. The first culture has been laying the emphasis on a qualitative methodology, while the second culture has been fixing a quantitative methodology as a must. Of course, there have been transgressions of the boundaries; ecology, pharmaceutics, or parts of physics have partly become friends with anthropomorphisations – one step towards esotericism; psychology, economics, or empirical social research are accustomed to performing as if belonging to natural sciences – one step forward to their computerisation and technisation as might be the case of computational social science.

Though the intent to find a general methodology for research in humans and machines is commendable, neither attempt to let methodology stretch across its own boundaries is a solution, as long as they are not taken up with a third culture.

By applying a method of generating knowledge you will not get findings other than those that are due to the method applied. The method applied is the necessary condition on which a particular model is based.

Monistic Models

Both anthropomorphism and technomorphism come up with a monistic ontology. Being a human and being a machine are assumed to be identical. However, the identity is constituted on the basis of their different framing ways.

Anthropomorphism is prone to stating that any machine resembles essentially a human. Technomorphism is in favour of saying that any human resembles essentially a machine. Anthropomorphism projects essential human features – like disposing of intelligence – onto machines. Technomorphism reduces essential human features – like disposing of intelligence – to features of machines.

Projection and reduction follow a stepwise order of mediation.

The anthropomorphic projection runs through the following steps:

- In a first step, the essential features of sociality of humans, namely, that they live in society governed by social relations, are projected onto the individual actor, thereby making her a social being.
- In a next step, the essential features of this individual actor as social being are projected onto the human body of the individual as a living being, by which she is viewed as a bio-social being.
- In a further step, the essential features of this bio-social being are projected onto the physical substrate of the bio-social being so as to yield a physico-bio-social being.
- In a final step, the essential features of this physico-bio-social being are projected onto any mechanistic compartment of the physico-bio-social being, so as to blur the distinction between the human and the machine.

Human(like)ness is conferred from human intelligence via mechanisms that work in the human body and might be part of human intelligence to the mechanics of artefacts. So, AI can be imagined as being humanly animated. Anthropomorphism is hence close to ideas that conceive our planet as a living organism, or the universe as ensouled or as a big natural computer.

The technomorphic reduction is carried out by a concatenation of the following steps:

- First, the essential features of the society of humans are reduced to those of the individual actor. This is an individualistic fallacy.
- Second, the essential features of the individual social actor are reduced to those of the human body. This is a fallacy of biologism, since the social features of the individual are narrowed down to biotic features.
- Third, the essential features of the human body are reduced to those of its physical substrate. This is a fallacy of physicalism, since the biotic features of the body are narrowed down to physical features.

- Fourth, the essential features of the physical substrate are reduced to those of mechanisms. This is a fallacy of mechanicism, since the physical features of the substrate are narrowed down to mechanical features. The term mechanical denotes here having the property of strict determinism. The physical world is not full of mechanisms only.

According to technomorphism, human intelligence boils down to a mere mechanical capacity that artefacts can be made capable of.

Monistic models that conflate human and machine form necessary conditions for particular design practices.

Assimilative Designs

Both anthropomorphism and technomorphism recommend an indiscriminative strategy when it comes to praxiology. Praxiology is a term that comprises those parts of philosophy that, apart from epistemology and ontology, deal with issues that are suitable for the general guidance of human practice such as values and norms; ethics, aesthetics or axiology belong to this class of philosophical disciplines. Praxeology is the name of a certain school of praxiology.

According to the conflationist suggestions, human and machine shall be treated in one and the same way. But they have different beliefs of how the activity shall be guided.

Anthropomorphism renders the humans colonised by machines, if it declares, in account with its projective ontology and epistemology, that machines shall be treated like humans. By adding to machines a value that is improper, humans become assimilated to them. The design of machines aims at producing 'techno sapiens' (Wagner 2016) – autonomous beings endowed with AI that delimits the generic autonomy of humans and ignores the fact that the evidence of intelligence that is based on the observation of behaviour only is no robust evidence at all (think of the Turing test that, actually, proves how easily human comprehension can be fooled).

The technomorphic credo runs the other way around: not machines shall be treated like humans but humans shall be treated like machines. This is at the same time the motto of transhumanism. The design aims at 'homo deus' (Harari 2016) by perfecting the species with artificial means, including the enhancement of their intelligence. Humans shall be engineered to be optimised. In that humans shall become machines themselves, humans are assimilated to machines, again.

Disconnections

Disconnections are the opposites of conflations. They come up as results of disjunctive ways of thinking and acting. The human and the machine are disjoined

and separated so much that they don't seem to have anything in common. Disconnections come in three variants – one comes as focus on the human with disregard for the machine, another as focus on the machine with disregard for the human, and a last one as focus on an interaction of disjoint humans and machines. The first disconnection is anthropocentric, the second technocentric, and the third hybrid, that is, human–machine-interactive. As to complexity, all variants presume self-contained degrees of complexity independent of any other complexity.

Let's again discuss the frames, models and designs of the three variants.

Disciplinary Frames

In epistemology, all variants agree that data of the human or data of the machine need each a frame of their own. In contrast to the cross-disciplinarity of the conflationist frames, they represent different supporters of disciplinarity. Anthropo- and technocentrism form a group of adherents of mono-disciplinarity and hybrid human-machine-interactivism follows multi- or inter-disciplinarity.

Mono-disciplinarity means intra-disciplinary research, it goes inside one discipline. Anthropocentrism claims social science and humanities methods for social and human data, technocentrism claims technological methods for technical data. Since in the first case the role of the lead science is attributed in that context often to sociology, the anthropocentric frame can thus run under the label sociologism. The technocentric frame might be called – analogically – technologism. Sociologism gives technological issues no attention. Thus, it does not care about artificial intelligence. Technologism is another methodological choice that is found at departments of computer science and others throughout the world. It is nourished by the condition of competitive excellence in one's own discipline and AI is one of the important fields and it has been diversifying into related fields like Autonomous Systems, Deep Learning etc. Both sociologism and technologism add to the existence of two cultures instead of trying to overcome them.

Multi-disciplinarity 'includes several separate disciplines, e.g., when researchers from different disciplines work together on a common problem, but from their own disciplinary perspectives' (Burgin and Hofkirchner 2017, 2). Multi-disciplinarity is a rather undeveloped state of working together. Inter-disciplinarity 'involves interaction and coordination between several disciplines aimed at the development of knowledge in these disciplines, e.g., when researchers collaborate transferring knowledge from one discipline to another and/or transforming knowledge of one discipline under the influence of another discipline' (Burgin and Hofkirchner 2017, 3). But despite cursory exchanges at points of intersection, disciplines keep themselves reciprocally exclusive without significant change – think of Science-Technology-Society, of

Informatik und Gesellschaft in German-speaking countries and else. Hybrid human-machine-interactivism tries a mix of particular frames. As long as a third culture will not be under consideration, a mixed frame will not transform the encounter of human intelligence and AI into a consistent approach.

Those deficient epistemological frames are a shaky premise for ontologies.

Dualistic Models

As to ontologies, anthropo-, technocentric and interactivist models are used to dualism instead of monism as in the case of anthropo- and technomorphism. Human and machine are assumed to be disjunct and to belong to different classes of the real world.

The main point of anthropocentrism is that the human is incommensurable with a machine. Humans and society are modelled as something completely different from a machine. Man is not a machine. Man is unique. Idealistic and spiritualistic positions would share such an approach. Humans are regarded as sentient, robots as corpses. Human intelligence is not mechanical.

What the anthropocentric ontology holds for the human, technocentrism holds for the machine. The machine is modelled as something that avoids human error. This makes machines unique. Technophilia as in trans- and posthumanism are examples of such a position. Machine intelligence is not human.

While the anthropocentric and the technocentric models hypostatise the uniqueness of either the human and social or the machine, the hybrid, interactivist model focuses on the interaction of both sides that enter the interaction as independent entities. But since the different degrees of complexity of both sides are not taken into consideration, a plural network is hypostatised that obscures the effective working of the interaction. This is the result of using the frames of multi- and inter-disciplinarity. Examples are the flat ontologies in Bruno Latour's Actor-Network Theory (ANT) (Latour 2006), which conceives humans and machines as 'actants', as well as sociomaterialism (Barad 2012, Suchman 2007), which conceives of generic 'intra-action' of agents with their ecologies.

Dualistic models that cannot avoid the disconnection of human and machine are the proper basis for designs that segregate.

Segregative Designs

Anthropocentric, technocentric and interactivistic designs follow the pattern of segregation. The human and the machine shall be treated in discriminative ways.

Anthropocentrism holds that the human shall be treated better than the machine. Man is the pride of creation, as theocratic beliefs have been formulating.

The human shall be perfected without resorting to technology. Social processes are placed over and against technological ones – technology is treated as trumpery, engineering might even be dangerous. AI is not needed or might devaluate the position of human intelligence.

The technocentric position is the opposite of anthropocentrism: The machine shall be treated better than the human. The machine is to be perfected to be devoid of human error. If a machine is liable to failure, then it is because of errors of the operators, that is, humans, because of programming errors that are the fault of humans, or because of material defects that are, in the end, due to faults of humans, again. Machines can, in principle, and they do so in reality, outperform humans. Intelligence of machines will render the intelligence of humans obsolescent. That is the credo of posthumanism and singularitarianism – a kind of Nietzsche's Übermensch but ex machina, that is, from the machines, robots, autonomous systems, AI.

The interactivistic position does not prioritise either side: The human and the machine shall be treated on an equal footing. However, doing so falls back into conflationist positions as to the interplay of social and technological practices. Anyway, in hybrid networks, design levels up machines or levels down humans. According to the famous saying of Latour that it is not me who shoots with the pistol but it is the pistol which (maybe better: who?) shoots with me, it is not humans who make decisions but intelligent devices whose decisions we just adapt to or execute (e.g., in the case of so-called expert systems in health care).

Combination

In contradistinction to conflations that frame, model and design human and machine on the sole basis of supposed identity of their degrees of complexity as well as in contradistinction to disconnections that do the same on the sole basis of a supposed difference of their degrees of complexity, a third way of thinking and acting orients towards the acceptance of identity and difference of their degrees of complexity at the same time – an enterprise of integration of human and machine. Integration is a combination that does justice to both what is universal and what is particular to human and machine.

The term that is chosen here to characterise the combinations with regard to the epistemological, ontological and praxiological aspects is techno-social systemism.

Transdisciplinary Frames

Techno-social systemism transgresses cross-disciplinarity and disciplinarity, in particular, it needs more than multi- or inter-disciplinarity – it needs trans-disciplinarity. Transdiciplinarity 'encompasses problems from different

disciplines but goes on a higher level than each of these discipline goes. In other words, transdisciplinarity treats problems that are at once between the disciplines, across the different disciplines, and beyond any of the individual disciplines involved. It is aimed at understanding of broad spheres of the world directed at the unity of knowledge' (Burgin and Hofkirchner 2017, 3).

A transdisciplinary frame needs systemism in the methods, that is, the assumption that different disciplines are to be interrelated in a systemic framework that provides what they have in common and grants, at the same time, relative autonomy to each discipline according to their place in the overall framework. Both social science and technology need to complement each other in order to constitute the big picture. Social data, technical data and data of the techno-social interaction are needed in unison.

Systemism has the potential to combine those data by combining the disciplinary approaches in question. It gives the whole edifice of sciences a new shape, from philosophy over the formal, real-world and applied sciences further on to disciplines on sub- and sub-sub-levels. It turns the formal sciences into a systems methodology, the real-world sciences into systems sciences and the applied sciences into sciences of artificial design of those systems. In such a way, the foundation of a science of techno-social systems is laid. Social science and engineering construe a common understanding of the systemic relationship of society and technology such that social systems science informs 'engineering systems science by providing facts about social functions in the social system that might be supported with technological means'; engineering systems science provides 'technological options that fit the social functions in the envisaged techno-social system'; and social systems science investigates, in turn, 'the social impact of the applied technological option in the techno-social system and provide[s] facts about the working of technology' (Hofkirchner 2017, 7). Hence, the epistemology of techno-social systems research paves the way for an ontology of human and machine, and for a praxiology of an integrated cycle of technology assessment and technology design.

Thus, techno-social systemism claims for a single frame for social and technical data that are comprised on a systemic meta-level.

The way is open to an unfettered scientific understanding of human intelligence, artificial intelligence and their relationship.

Dialectic Models

A techno-social systems ontology cannot resort to monism nor to dualism. It requires dialectic. A dialectical relationship goes beyond duality in that sides or parts are not completely separate. And neither are they brought together by operations on the surface. They hang together intrinsically, but asymmetrically, over steps of emergence. They are evolutionary products, they give rise to evolutionary products, they are nested one in another in line with their complexity.

Techno-social systems are social systems. They emerge from social systems when technologies of any kind are inserted into the social systems so as to improve the functioning of the social systems to reach a certain goal through the mediation of these technologies. These technologies transform those very systems into techno-social ones. These technologies are devised and developed to functionalise a certain cause-effect-relationship of the real world as artificial mechanisms in which the effect becomes the goal and the cause becomes the leverage. In order to serve effectively and efficiently the attainment of the desired or needed goal, artificial mechanisms are prepared to function as strictly deterministic as possible. In this respect, artificial mechanisms resemble natural mechanisms – the latter work according to strict determinism too. An artificially prepared mechanism is what is usually known under the term machine.

Thus, techno-social systems integrate humans and machines. Humans are products of evolution, machines are products of humans. Techno-social systems integrate them in line with their ontic features according to their evolutionary history. Humans and machines share, or have distinct, physical, biotic and social features (Hofkirchner 2020).

Let's first discuss their physical features:

Humans and machines share the fact that they are entities and embrace processes that belong to the physical realm. However, they differ essentially with regards to the specifics of their being physical and behaving physically. Making use of a distinction of Rafael Capurro (2012), humans and society can be interpreted as an agens – that is something that displays agency by itself – whereas a machine can be interpreted as a patiens – that is something that does not display agency and is passive. This is indicated by the following:

- Humans and society are able to organise themselves, that is, to build up order by using free energy and dissipating used-up energy, whereas machines cannot self-organise.
- Humans and society are made up of elements that produce organisational relations that constrain and enable synergy effects and they can constitute superordinate systemic entities, whereas machines are made up of modules that are connected in a mechanical way.
- Humans and society function on the basis of less-than-strict determinacy, which yields emergence and contingency, whereas machines are strictly deterministic and cannot behave in an emergent or contingent manner.

Second, let's turn to the discussion of biotic features:

Humans and society are physical entities and activate processes that belong to the biotic realm. Machines may, but do not need to, have parts that belong to the biotic realm. Even in cases where they do so, they differ essentially in quality. Humans and society are agents that are autonomous in the true sense of the

word (Collier n.d.), whereas machines are heteronomous mechanisms that can thus not show any degree of autonomy, as follows:

- As with any living system, humans and society are able to maintain their organisational relations by the active provision of free energy, whereas machines cannot maintain themselves.
- As any living system, humans and society are able to make choices according to their embodiment, their embedding in a natural environment and the network of conspecifics, whereas machines cannot choose.
- As any living system, humans and society are able to control other systems by catching up with the complexity of the challenges they are faced with by the other systems, whereas machines cannot catch up with complexity and are under control by organisms.

And, third, let's discuss one last category of features – the social one:

Humans and society are not only physical and biotic, they are the only physical and biotic systems on Earth that belong to a specific, the social realm, too. They are, essentially, social agents, that is, actors. Machines are social products, artefacts, that are made by actors, but they do not possess the agency of actors. This is implied by the following:

- Humans in society constitute – by action, interaction and co-action with other actors – social agency that reproduces and transforms the structure of the social system (social relations), that, in turn, enables and constrains the social agency, whereas machines do not partake in the constitution of society but support the action, interaction and co-action of actors.
- Humans in society provide the commons as effects of social synergy, whereas machines support the provision of commons and pertain themselves to the commons.
- Humans in society are the driving force of social evolution, including the evolution of culture, polity, economy, ecology and technology, whereas machines are driven by social evolution. However, they can even play a supportive role in changing the quality of the social system.
- Humans in society reflect upon the social structure, whereas machines do not deliberate but support the thought functions of actors.
- Humans in society set off the transition into actuality of a societal option of choice out of the field of possibilities, whereas machines do not directly trigger emergence.

As to the role of AI in the context of techno-social systems, we can conclude that artificial intelligence is and will be a mediation of the collective intelligence actors are capable of but is not and will never be (a property of) an actor itself. What is labelled AI, is nothing that can become independent and achieve a life

of its own. However, it promotes the intelligence of the social system. In this vein, Francis Heylighen (2015, 2016) rejects the idea of a singularity by which a single supra-human artificial intelligence seems purportedly possible, since intelligence is and will be distributed over social actors that cyber technology merely connects, which means that the emergence of a 'global brain' remains rooted in humans. From this dialectical point of view, what is *in statu nascendi* is a social suprasystem that would be global, notwithstanding the technological infrastructure of a global brain.

Dialectic models are the proper contributions to a paradigm shift towards the third culture.

Integrative Designs

Techno-social systemism demands an integrative way of thinking and acting. It demands responsibility in two different respects: first, the responsibility for the functionality of what shall be designed – does the mechanism effectively and efficiently serve the purpose for which it shall be designed? This is a matter of fact. However, since the question of how functional technology is can be answered in a decontextualised manner from a mere technical point of view, a second respect is required: The responsibility for the meaningfulness, for the social usefulness of what shall be designed – does the purpose for which technology shall be designed also make sense, that is, does it promote a social value, does it conform with a social norm? The whole picture of praxiology can be seen only when in the context of the social. The default value of meaningful technology is to serve the vision of a good society, of individuals living a good life and of cultivating the common good. Such an alter-humanism instead of an old-fashioned humanism or post-humanism is compatible with the third culture – alter-humanism harnesses tools for conviviality (Illich 1973). This means that techno-social systems integrate humans and machines according to their appropriate treating. The check of that necessitates an integrative technology assessment and technology design.

Conclusion

A review of possible ways to establish a relation between human and machine clarifies the shortcomings, if not the stubbornness of old-fashioned humanism, on the one hand, and anti-humanism in a modern disguise, on the other, when an identity of human and machine is affirmed at the cost of their difference that is negated – done so by conflations – and when the difference between human and machine is affirmed at the cost of their unity that is negated – done so by disconnections. The way out is the establishment of a relation through affirming both the identity of, and the difference between, the two sides – as done by

combinations. Combinations provide the proper basis for a humanism that is up to the challenges of digitalisation – Digital Humanism.

References

Barad, K. 2007. *Meeting the Universe Half-Way – Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Barad, K. 2012. *Agentieller Realismus – Über die Bedeutung materiell-diskursiver Praktiken*. Berlin: Suhrkamp.

Burgin, M. and Hofkirchner, W. 2017. Introduction: Omnipresence of Information as the Incentive for Transdisciplinarity. In M. Burgin and W. Hofkirchner (Eds.), *Information Studies and the Quest for Transdisciplinarity – Unity through Diversity*, pp. 1–7. Singapore: World Scientific.

Capurro, R. 2012. Toward a Comparative Theory of Agents. *AI & Society*, 27(4), 479–488.

Collier, J. n.d. What is autonomy? Last accessed 24 February 2020, http://cogprints.org/2289/3/autonomy.pdf.

Gartner Research. 2015, 22 April. *Digital Business: Digital Humanism Makes People Better, Not Technology Better*. Last accessed 22 July 2020, https://www.gartner.com/en/documents/3035017/digital-humanism-makes-people-better-not-technology-better.

Harari, Y.N. 2016. *Homo Deus – A Brief History of Tomorrow*. London: Harvill Secker.

Heylighen, F. 2015. Return to Eden? Promises and perils on the road to a global superintelligence. In B. Goertzel and T. Goertzel (Eds.), *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity*, pp. 243–306. Los Angeles, CA: Humanity+ press.

Heylighen, F. 2016. A Brain in a Vat Cannot Break Out – Why the Singularity Must be Extended, Embedded and Embodied. In U. Avret (Ed.), *The Singularity – Could Artificial Intelligence Really Out-Think Us (and Would We Want it To)?,* vol. 19, pp. 126-142. Luton: Andrews UK Ltd.

Hofkirchner, W. 2013. *Emergent Information – A Unified Theory of Information Framework*. Singapore: World Scientific.

Hofkirchner, W. 2017. Transdisciplinarity Needs Systemism. *Systems*, 5(1), 1–11. DOI: https://doi.org/10.3390/systems5010015.

Hofkirchner, W. 2020. Blurring of the Human and the Artificial – A Conceptual Clarification. *Proceedings*, 47(1), 1–3, DOI: https://doi.org/10.3390/proceedings47010007.

Hofkirchner, W. and Kreowski, H.-J. (Eds.). 2020. *Transhumanism – The Proper Guide to a Posthuman Condition or a Dangerous Idea?* Cham: Springer.

Illich, I. 1973. *Tools for Conviviality*. New York: Harper and Row.

Latour, B. 2006. *Reassembling the Social – An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Nida-Rümelin, J. and Weidenfeld, N. 2018. *Digitaler Humanismus – Eine Ethik für das Zeitalter der Künstlichen Intelligenz*. München: Piper.

Snow, C.P. 1998. *The Two Cultures – A Second Look*. Cambridge: Cambridge University Press.

Suchman, L. 2007. *Human-Machine Reconfigurations – Plans and Situated Actions*. Cambridge: Cambridge University Press.

Vienna Manifesto on Digital Humanism. n.d. Retrieved from: https://www.informatik.tuwien.ac.at/dighum/index.php.

Wagner, P. 2016. Techno Sapiens: Die Zukunft der Spezies Mensch. Documentation, broadcast on 16 November 2016, 3sat. Retrieved from https://www.3sat.de/wissen/wissenschaftsdoku/techno-sapiens-die-zukunft-der-spezies-mensch-100.html.

CHAPTER 4

An Alternative Rationalisation of Creative AI by De-Familiarising Creativity: Towards an Intelligibility of Its Own Terms

Jenna Ng

'There, look!' we could say. 'Look at this art! How dare you claim these children are anything less than fully human?'

– Kazuo Ishiguro (2005, 238)

Introduction

This chapter formulates an alternative understanding of creative Artificial Intelligence (AI) by examining how the computational terms of AI may be rationalised in a framework *intelligible* to humans. The level of algorithmic processing today presents two tensions which hinder a full comprehension of creative AI. The first is the still formidable lack of transparency of AI's workings, as noted by many scholars *à la* the algorithmic 'black box' (Pasquale 2015; Diakopoulos 2016; Brill 2015; Ananny and Crawford 2018). The second is the increasing lack of human intervention in the algorithm's processing not only through the seemingly unfathomable operation of its 'black box', but also through algorithms learning from other algorithms, such as by way of a Generative Adversarial Network (GAN). The result is to consider anew how computers may be

How to cite this book chapter:
Ng, J. 2021. An Alternative Rationalisation of Creative AI by De-Familiarising Creativity: Towards an Intelligibility of Its Own Terms. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 49–66. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.d.

considered to be autonomous creators – to be genuinely creative in and of itself, or creators 'in their own right' (Veale and Cardoso 2019, 2). Or, as Hofstadter (2000) writes: 'It [the mechanical substrate of creativity] may not constitute creativity, but *when programs cease to be transparent to their creators*, then the approach to creativity has begun' (669; emphasis added).

Recent innovations in creative AI bear out both tensions, where the algorithm generates creative decisions, say on note, word or paint placement, out of its *own* processing of the dataset it receives, and in ways not entirely understandable to humans. This level of processing may be contrasted with how computers had previously produced creative work, or in what is known as automated creativity. As early as the 1950s, computers have produced creative outputs, such as music, by running codes of basic material and stylistic parameters which enabled the generating of 'raw materials'. These musical 'materials' were then modified and assembled by human composers into recognisable pieces of music (Alpern 1995). In these cases, the computer was *specifically programmed to produce the creative output*, following instructions on where and how to place notes or dabs of paint, even if those instructions may be rules of randomness.[1] The research field of Computational Creativity recognises such programming as 'pastiche', where the computer's creativity is a 'mere appearance', and only 'due to some specifiable slice of the programmer's own creativity having been imprinted onto the algorithmic workings of the system'. (Veale and Cardoso 2019, 4).

Conversely, current creative AI operates as neural networks which discern specific patterns out of processing large datasets of relevant outputs, thus 'learning' across complex nodal networks of 'reward' and 'punishment' the placements of note, paint and words for producing the creative output in keeping with the patterns they have 'learned'. A couple of examples to illustrate this: in 2016, the Project Magenta team at Google unveiled a 90-second melody produced by a computer to which they fed 'some 4,500 popular tunes' and 'seeded' with four musical notes. By processing the large database of tunes, the network 'composed' its melody by discerning patterns of musical rules and constraints *in ways never specifically programmed into it*. The algorithm may also learn from other algorithms, such as the algorithmically-produced portrait of Edmond Belamy in 2018. In this case, a discriminator algorithm was first trained on a database of 15,000 portraits (painted by human artists) that had been uploaded to it (by human computer scientists). The discriminator algorithm was then used to 'train' a separate generator algorithm which 'learnt' paint placements and so on based on 'reward' and 'punishment' feedback from the discriminator algorithm, both doing so through processing enormous amounts of data. The Edmond Belamy painting later made history as 'the first portrait generated by an algorithm to come up for auction', and eventually sold by Christie's for a not insubstantial sum of US$432,500 (Cohn 2018). The key issue in these cases is that the algorithms have not been specifically programmed to place notes and

paint; instead, they 'learnt' to do so through processing enormous amounts of data and being given signals on what placements were 'correct' or 'wrong'. On that feedback, they then generated their respective outputs.

While not quite the spectre of a Terminator machine out to annihilate the human race, creative AI on these terms is disturbing in how our lack of understanding of its creativity and creative process reinterrogates our notions of humanness, where creativity has always been its indisputable hallmark (Zausner 2007). It is 'part of what makes us human' (Sawyer 2006, 3), and affirms our humanity (Csikszentmihalyi 1990); it colours the domains in which humans work, think, play, produce and perform (Kaufman and Baer 2005). Per the opening quotation of the Introduction, the clones in the speculative society of Kazuo Ishiguro's (2005) acclaimed novel, *Never Let Me Go*, made art as a concerted attempt to evidence their humanity. As their teacher explained to them: 'we thought it would reveal your souls. Or to put it more finely … to prove you had souls at all' (Ishiguro 2005, 238). The clones' creative work were sought to demonstrate humanity, for 'the creativity code', as Marcus du Sautoy (2019) puts it, 'is a code that we believe depends on being human' (2–3). How, then, may we understand so critical a touchstone of humanness in AI when creativity is seemingly manifest on such opaque and unintelligible terms?

This chapter thus proposes *a framework of de-familiarisation* for the paradoxical task of rendering the computer's creativity, seemingly so entrenched on its own terms of computational data and processing, intelligible in human terms. Its aim is to propose an approach with which to rationalise the processes of the computational algorithm, if anything to render the clarity of the imbrications between the human and the computational that colour so much of our algorithmic world today. First, as a brief literature review, I present a few salient tenets of existing rationalisations of AI. In particular, I critique how their approaches, by and large, extract comparative analyses between human functions and computational processes. To formulate an alternative approach, I then draw from rationalisations of media out of media theory, specifically theorisations of the marionette by Heinrich von Kleist (1810) and of the camera via Russian filmmaker Dziga Vertov's (1923) writings on cinema, to present a methodology of *de-familiarisation* as an approach to rationalising technology *on its own terms*. In the third section, I apply that perspective to re-think creative AI via the case study of AlphaGo, an algorithm programmed by Google DeepMind to play the game of Go and which made AI history in 2015 by becoming the first computer programme ever to beat a human professional player at the game. While AlphaGo does not produce artistic work *per se*, it serves as an apt case study as its moves were deemed to be of exceptional novelty – indeed, described as 'creative genius' (Sautoy 2019, 34) – and in various ways considered to have re-defined the frontier of AI. The last section will conclude. The uniqueness of this argument thus lies in how it aims to shift the conversation from an us-and-them framework, where computing is often conceived on the singular

oppositional dimension of humans versus machines (such as comparing computers directly against humans). This alternative approach to understanding algorithms thus suggests a different dimension to that understanding – not one made on human terms, but as a paradoxically impossible approach of the algorithmic being humanly intelligible on its own terms.

Current Rationalisations of Creative AI

Current consideration of creative AI lies in extensive scholarship, not least because much of it sits within a vastly wider enquiry: can computers be human? In the face of this question, current discourse inevitably turns to a comparative methodology, whereby the computer's processes are compared against multiple manifestations of human cognitive function, including creativity (Dreyfus 1972; Dreyfus 1979; Bailey 1996; Moravec 1998; Boden 2004). Various conclusions are then reached by matching one against the other, and working out how each measures up.

In other words, the rationalisation of AI is laid out in *comparative* terms, so that AI becomes intelligible only as *against human capacities*, or against what AI can or cannot do as compared to humans. The multiplicities which reflect this rationalisation across philosophy, computer science, cognitive psychology, cybernetics, neuroscience and myriad other disciplines are myriad and intricate, and far beyond the scope of this chapter to cover comprehensively. A few highlights will hopefully suffice to demonstrate its contours. We might, for instance, think about Vannevar Bush's famous imagining of a memory machine he named the 'memex' (Bush 1945), influential to the present day as a basis for the World Wide Web (Davies 2011). Notably, Bush presented the memex as a technology directly against human memory, specifically referencing the former's mechanised processes of speedy and flexible consultation and storage against the latter's corresponding weaknesses, leading to impermanence (forgetting) and lack of clarity (confusion). Conversely, Bush also noted the strength and speed of association of human memory, concluding that 'man cannot hope fully to duplicate this mental process [of association] artificially, but he certainly ought to be able to learn from it' (Bush 1945, n.p.). Both points illustrate Bush's rationalisation of technology as a counterbalance to human capability, whereby one variously contrasts against, supplements, and demonstrates differences against the other. The technology is thus made intelligible as *against the human*, specifically in terms of what it can augment and surpass, and what it cannot.

As AI – itself the field of computers which simulate human cognitive capacities – increases in operative sophistication to resemble human intelligence, this contrast becomes ever more explicit. The Turing test (Turing 1950), even more famous than Bush's memex machine, reconciled computer cognition in terms of whether it was distinguishable – or not – from human behaviour.

Indeed, John Searle, among many other computer scientists, distinguished between 'strong AI' and 'weak AI' in his now classic 1980 paper, 'Minds, Brains, and Programs' (Searle 1980), later developed into his book, *Minds, Brains and Science* (Searle 1984), precisely on such rationalisations of the computer against characteristics of human cogitation. In his paper, Searle argued that AI, in its state of development then, could only be 'weak', whereby 'the principal value of the computer in the study of the mind is that it gives us a very powerful tool'. Conversely, it was not 'strong', whereby 'the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states … the programs are themselves the explanations' (417; emphasis in original). Searle clinched his argument against 'strong AI' by arguing its lack of free will and other mental states which, according to Searle, characterise human cognition, thus demonstrating the limitations to 'computer simulations of human cognitive capacities' (417). Again, these arguments rationalise AI against the human in a comparative mode. They render AI intelligible by referencing its technological Otherness against constructed definitions of natural human responses.

As a counter-stroke, we might also think about the extensive work in computer and cognitive sciences which rationalise *the human being* in computational terms. However, this shifted understanding of the human as a computer only expands the commensurability between AI and humans, this time not by *difference* (humans against computers), but by *equivalence* – humans are computers. In turn, this intelligibility of AI via counterpointing the human – in terms of underscoring AI's logical and mechanised processes as against the biological and the organic – expands to not only the rationalisation of the human, but the world itself. This, then, is the core of *computationalism*, defined by Golumbia (2009) in its 'received', or 'classical', form, as

> … the view that not just human minds are computers but that *mind itself* must be a computer – that *our notion of intellect is, at bottom, identical with abstract computation*, and that in discovering the principles of algorithmic computation via the Turing Machine human beings have, in fact, discovered the essence not just of human thought in practice but all thought in principle (emphasis added). (7)

The idea has flexed and flagged in multiple ways and forms, but its central concept remains the conceptualisation, understanding and identification of human cognition and mind in computational terms. In this respect, we can also think about, for instance, Allen Newell and Herbert Simon's work across the decades from the 1950s which specifically argued for the model of all human reasoning to be representable as symbolic 'information processing systems' (Newell and Simon 1972). Giants in their respective fields, both awardees of the Turing Award and Simon as well a Nobel laureate, their thinking converged with others at the Dartmouth summer conference of 1956,[2] whose specific mission, notably,

is 'to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it' (McCarthy et al. 1955, n.p.). As the field (and the term) of 'Artificial Intelligence' emerged out of the Dartmouth conference, its ideas folded into another new field developed in the 1970s – namely, cognitive science, which studies human thinking, learning and perception as coloured by cybernetics, neuroscience, linguistics and psychology, but also dominated by AI, mathematics and computation (Gardner 1985). The field of cybernetics, first emerging out of the Macy Conferences on Cybernetics from 1943 to 1954, also forged new paradigms out of information theory, neural functioning and computer processing, among others, to become 'a new way of looking at human beings. Henceforth, humans were to be seen primarily as information-processing entities who are *essentially* similar to intelligent machines' (emphasis in original; Hayles 1999, 7).

This approach of rationalising the human in computational terms infuses much of current thinking about creative AI (Boden 1996; Boden 2004; Miller 2019; Sautoy 2019; Kaufman and Baer 2005). The definitional knottiness of the term 'creative' aside – over 60 definitions of 'creativity' appear in psychological literature alone (Boden 1996, 268) – the broad rationalisation of creative AI continues along comparisons against human creativity as couched *in computational terms*. Hence, for instance, Douglas Hofstadter suggests the 'mechanisation of creativity': while 'creativity is the essence of that which is not mechanical', '[y]et *every creative act is mechanical* – it has its explanation no less than a case of the hiccups does' (emphasis added; Hofstadter 2000, 669). Similarly, Herbert Simon, as already seen, justifies human cognition as informational processing systems, and thus posits that 'creativity involves nothing more than normal problem-solving processes' (as quoted in Csikszentmihalyi 1988, 19). More recently, Miller (1992; 2000; 2019) rationalises human creativity as 'a model for network thinking' in terms of 'many lines of thought taking place at once in parallel, coming together from time to time to enrich each other' (Miller 2019, 36). The model thus demonstrates how 'new ideas do not just pop up out of nowhere, even though they may seem to', thereby visualising human creativity as a mappable process that is also reproducible on a computer (Miller 2019, 29). Having said all that, the idea of mechanised creativity stretches back to the ancient Greeks, where Burkholder, Grout and Palisca (2014), for instance, argues that Pythagoras and his followers held that 'numbers were the key to the universe', and thus thought music as 'inseparable from numbers' (13). Creating music, then, was really the theoretical application of numbers and various mathematical properties in logical and calculable steps, not unlike in algorithmic fashion.

Across these formidable rivers of thought and expansive arguments, if sketched in generous outlines and overlooking many more, we may thus identify how rationalisation of human creativity along mechanical computational

processes dovetails into the persistent thinking of AI as made intelligible against the human. As such, the intelligibility of AI continues to be tracked against shifting interpretations of human capability and, in that respect, facile notions of what constitutes humanness. The arguments which directly oppose creativity as defined in terms of logic and mechanisation just as easily deploy a notion of creativity that appeals to other touchstones of humanness deemed still unachievable by the computer, such as consciousness, self-understanding and awareness: a truly creative computer, after all, 'cannot be a dumb savant that naively flings outputs at an audience' (Veale and Cardoso, 2019, 4). Or that creativity entails unique human experiences, such as 'the need for experience and suffering' (Miller 2019, 16) or self-actualisation, where creativity is 'about humans asserting that they are not machines', and 'to expose what it means to be a conscious, emotional human being' (Sautoy 2019, 283). Hofstadter (2000) refers to 'the depth of the human spirit' for a meaningful notion of creativity:

> A 'program' which could produce music as they [Chopin or Bach] did would have to wander around the world on its own, fighting its way through the maze of life and feeling every moment of it. It would have to understand the joy and loneliness of a chilly night wind, the longing for a cherished hand, the inaccessibility of a distant town, the heartbreak and regeneration after a human death. It would have to have known resignation and worldweariness, grief and despair, determination and victory, piety and awe. (672–673)

The point here is not to argue for any particular definition of or position about the computer's creativity. Rather, it is to underscore the various lenses deployed through which AI creativity is rendered intelligible, namely, *in relation to the human* in terms of comparison, contrast and analogy. In some ways, this is wholly intuitive – as mentioned, technology forms a counter-distinction to humanness; it mirrors the age-old binary of artifice against nature. It thus makes sense to employ humans as the referential framework in understanding the technological Other. Yet, the approach is also flawed. Understanding AI on these terms becomes subject to changing constructions and perspectives of humanness, so that it relies on a precarious balance of what AI *is* against what it is *not*. Conditional on being defined in relational terms, it fails to have its own definitional footing. An understanding of AI on these foundations cannot be a thorough one.

Moreover, this approach of comparison confines our understanding of AI to being *within the intelligibility of human terms*, rather than made intelligible *on its own terms* as a logical, mechanical and computational entity. This is important because at the heart of an intelligibility on human terms is an incommensurability that is never truly addressed: computers are simply not humans. A comparison that renders one intelligible on the terms of another will always

lose something in the translation. The next section, drawing on alternative perspectives from theorisations of media, will suggest a different approach.

De-familiarising Creativity

As a starting point, much of media technology is similarly rationalised as comparisons against human capabilities, as this broad scattering of examples will hopefully suffice: Nicholas Carr (2010), for instance, in resonant echoes of Marshall McLuhan ([1964]; 2013) rationalises technology as an expansion of 'our power and control over our circumstances' (44), such as the map and the clock which 'extend and support' the 'mental powers' of humans in formulating, producing and sharing knowledge. Jonathan Safran Foer (2016) writes of communication technologies as 'substitutes' for real-time face-to-face human interaction: 'We couldn't always see one another face to face, so the telephone made it possible to keep in touch at a distance. One is not always home, so the answering machine made a message possible without the person being near their phone.' (n.p.). Edward Branigan (2006) suggests anthropomorphism as an 'analytic category' to measure 'the degree to which a camera is being used to simulate some feature of human embodiment', whose analysis then relates the qualities of the camera to 'a typical way of human viewing, or moving (or thinking and feeling), and to what degree' (37). William Brown (2009) argues the pole converse in his thinking about 'posthumanist cinema', demonstrating how the digital 'posthuman camera' *omits* human embodiment entirely in its humanly impossible shots.

This comparative approach is also an old rationalisation. In his 1880 essay, Jean-Marie Guyau analogizes the human brain to the phonograph, drawing connections between recorded sound as grooves of vibrations engraved onto the phonograph's metal plate and the 'invisible lines [that] are incessantly carved into the brain cells, which provide a channel for nerve streams' (31). Guyau further analogizes the speed and strength of vibrations of our brain cells in terms of images conjured by our minds to the speed of the phonograph's vibrations and the tones of its sounds. In 1950, George Wald, a professor of biology at Harvard, noted resemblances between the camera and the human eye – 'of all the instruments made by man, none resembles a part of his body more than a camera does the eye' (32). Wald (1950) further detailed similarities between human vision and photography: 'the more we have come to know about the mechanism of vision, the more pointed and fruitful has become its comparison with photography' (32). He described how the chemical changes of exposed photographic film, particularly 'dark reaction' of the 'latent image' in the darkroom, mirrors processes of vision in the eye's exposure of rhodopsin[3] to light (40). Along these broad contours, the rationalisation of media technology thus echoes that of AI – as matched against human capabilities, reflecting similarities and differences; as situated *in human terms*.

However, large swathes of critical theory have grown to critique the human as a referential framework. Specifically, this work decentres the human by shifting the critical lens to those that are not human, such as 'understood variously in terms of animals, affectivity, bodies, organic and geophysical systems, materiality, or technologies' (Grusin 2015, vii). Understanding the Other and accommodating *on their terms* their complex involvement in the consideration of our world thus stands as a long-established enquiry through the humanities. This work spans across multiple areas, such as the social imaginaries of inanimate objects (Appadurai 2014); the posthuman (Braidotti and Hlavajova 2018); the nonhuman (Grusin 2015); post-anthropocentrism (Parikka 2015); the intelligibility of cinema as both subject and object of vision (Sobchack 1992) – to cite again just a few sprinklings as illustration. There are many others.

There is, of course, an inherent contradiction to this approach, which is of intelligibility having to be made intelligible in alien terms, or in terms of an Other-ness that, by definition, we do not and are unable to possess. How might we render something intelligible in its own terms if it is, by definition, outside the intelligibility of our own terms? How do we accommodate our understanding around something that we are not, let alone fathom its terms? It is certainly a valid conceptual difficulty. The key is to understand the enquiry not as a literal one which seeks literal answers. Rather, it is one which involves speculation and imagination in envisioning the perspective of the Other as part of its methodology. It requires the acceptance of the philosophy of things being in themselves, beyond and independent of our experience (Moffat 2019). It entails being open to indeterminacy and contingency, and of acknowledging the nature of things as based on, while not pure fantasy, nonetheless an inexact science.

The approach proposed here, then, for shifting one's critical perspective in relation to an intelligibility of the technological Other is to draw on media theory's alternative rationalisation of technology, namely, through *de-familiarisation* – to disturb or disorder the terms in which we think of an entity so as to re-learn it on different terms, specifically those of its own. I underscore two examples, each of a different media through different rationalisations, to more fully illustrate this approach. The first is Kleist's 1810 essay on marionettes, in which he recounts a conversation with his friend, 'Herr C.', who expressed admiration for the gracefulness of the puppets. This position is counterintuitive: puppets, controlled by their puppet-master, are mechanical and lifeless; it is senseless to consider puppets as graceful as, if not more graceful than, human dancers. The key in 'Herr C.'s' reasoning lies in how he *re-reads* the puppet's mechanical movements not as cold actions with no consciousness or with a surrendered volition, but as precisely the nonconscious and mechanical movements *that only puppets are capable of*, through which beauty and grace *re-emerges*. The puppet's artificial properties are thus read *on their own terms* – not against the human dancer's consciousness of its movements which renders their artistry and beauty. Rather, 'Herr C.' *de-familiarises* what and how

we think about grace, and relearns it in the non- or unconsciousness of the puppet's mechanical operations. We thus come to a different understanding of the puppet by emerging on the other side of its paradox (i.e., of grace from the controlled and the automatic) to arrive at an alternative intelligibility of it and its movements. *We understand the puppet on its own terms.*

The second example is drawn from Russian filmmaker Dziga Vertov's theorisation of cinema. Articulated in the 1920s through various pamphlets, articles, manifestos and public addresses, and fresh from radical societal change in the wake of the Russian Revolution, Vertov sought from cinema and its camera the newness of humanity and society. He read the camera through a new intelligibility – not against the human or in comparison to the human eye, but on its own terms as what he calls a 'kino-eye' in *how it sees a different world*: 'I am a mechanical eye. I, a machine, show you the world *as only I can see it*' (emphasis added; Vertov 1923, 17). Of course, the human hand and eye still control the film camera; it has not literally come alive. But the argument here is not a literal one. Rather, it is a theoretical shift involving imagination and inventiveness to acknowledge the new-ness of the camera's vision and its alien visuality. Like Kleist with the puppet, Vertov came to understand the camera *on its own terms* as he sought a visualisation of the new society birthed from revolution out of its camera eye: its alien-ness as an un-human consciousness is precisely why the kinoeye is capable of ideology to present the real in a way the human eye cannot. He could thus acknowledge the camera's *different-ness* – as he writes, 'it is the realization by kinochestvo[4] of that which cannot be realized in life' (Vertov 1922, 9). Hence, through re-reading the camera on its own terms, Vertov *de-familiarised* the world around us, presenting it anew in a radical language borne out of the camera's foreign intelligibility, and shifting images out of the referential framework of human seeing.

The point here is not to agree or disagree with Vertov or Kleist in their respective readings of media. It is to illustrate how a frame of reference in understanding can shift with a different rationalisation, and in so doing recognise a different intelligibility. The task, then, is to apply this approach to understanding creative AI on their own terms, to which the next section will turn via the case study of AlphaGo.

Rationalising the Creativity of AlphaGo

AlphaGo, an algorithm programmed to play the game of Go, achieved global fame in March 2016 by defeating Lee Sedol, a highly ranked South Korean professional Go player and 18-time world champion, 4 games to 1 in a 5-game tournament. Its victory sent ripples through the AI community and the wider public because 'teaching computers to master Go has long been considered a holy grail for artificial intelligence scientists' (Yan 2017, n.p.). The difficulty of

this 'holy grail' lies in the game's high level of abstraction. Played by two players each placing, in turns, stones of their respective representative colour (black or white) on intersection points between horizontal and vertical lines marked on a board, there are essentially just two rules of play: one on how to 'capture' an intersection; the other on how that intersection is considered 'occupied'. The goal is simple: to have, at the end of the game when all intersections have been 'occupied', stones of your colour 'occupy' more intersections (or territory) than your opponent's.

Like all good Zen koans, its minimalism is also its complexity. Compared to Go, chess, as a fellow strategy game, is clearer in various ways: fewer moves can be made to start a chess game, and thus relatively fewer possibilities branch out from each opening move. Pieces also have set values (the pawn, for instance, has the lowest; the queen the highest) which makes an unfinished chess position relatively easy to calculate and analyse as to which player is winning based on how many and which pieces are left, plus any positional advantages. In comparison, because all there is to Go are stones on line intersections, the result is many more possible board configurations, each one lending themselves to even more possible positions if calculated further down the line. As a result, there is quantitatively *and* qualitatively more ambiguity in Go, with ensuing greater difficulty in analysing who is winning from an unfinished game position. Hence the significance of AlphaGo's victory: due to its multiple positional possibilities – as has been oft-quoted, there are 'more possible configurations of the board than there are atoms in the universe' (Yan 2017, n.p.) – until AlphaGo's triumph, the game was considered unconquerable by computers over human players simply because its level of complexity needed it to be played with human abilities of intuition and grasping of visual structure, rather than the computer's powers of searching and calculation of variations.

Pertinently for our purposes here, AlphaGo was vaunted for not only its tournament victory, but also the creativity of its moves. One move – Move 37 of Game 2 – in particular was so wholly unexpected that commentators described it, if a tad gushingly, as 'a truly creative act' (Sautoy 2019, 37); or 'one of the most creative [moves] in Go history' (Tegmark 2017, 89). AlphaGo's Move 37 was to place a stone on an intersection on the board's fifth line, a move very seldom played at that stage of the game because it was considered too 'high' on the board, giving the opponent room to play on the fourth line down and thereby gain too much solid territory. The media quickly attributed the move to the algorithm's *own* creativity, lauding it, with embarrassing hyperbole, as 'the move no human could understand' (Metz 2016, n.p.). Or, as widely quoted from Fan Hui, the European Go champion who was the first professional Go player to play and lose against AlphaGo: 'it's not a human move. I've never seen a human play this move.' (as quoted in Metz 2016, n.p.).[5] What validated the unusualness of the move – and thus rendering it 'creative' rather than 'insane' or 'nonsensical' – was that, some fifty moves later, that fifth-line stone became

an unexpected linchpin to a battle for territory which started in a different part of the board. In due course, the battle joined up with the Move 37 stone, giving AlphaGo the advantage and eventually the win.

As expected, the rationalisation of AlphaGo's Move 37 lay in the conventional framework of human terms via comparison and contrast, this time by placing the algorithm in its own intelligibility as one outside human sense. Yet that does not achieve much for understanding AI in its conceptual sense – it merely blankets the algorithm with mystique of the technological and a cryptic referencing of its Other-ness on the basis of some kind of mysterious agency. For instance, much was made of AlphaGo's independent learning from its neural network to generate its moves. Like the painting of Edmond Belamy whose generator network 'learnt' paint placement from the discriminator network, AlphaGo, as the generator network, 'learnt' the best moves in Go by playing multiple games against another neural network. While the discriminator network would have been 'trained' to play Go by being fed (by human computer scientists) millions of games (played by human players) as downloaded from the internet, it is the processing of the millions upon millions of games between AlphaGo and its discriminator network that makes up AlphaGo's main 'training', namely, the calibration of the values and weightage for its nodes across its various networks which ultimately generates AlphaGo's moves.

The implication, then, is that the algorithm's creativity in coming up with unusual moves is its own, generated on its own steam and out of its own learning, an idea its Google DeepMind creators were keen to perpetuate. For instance, in a video interview with CNBC, Demis Hassabis, CEO of Google DeepMind, implies the same generative creativity, explaining how algorithms such as AlphaGo 'learn from scratch, learn from *their own mistakes* … they learn from themselves, directly from data or from experience, *rather than being told what to do by human programmers*' (emphasis added; as linked in Yan 2017, n.p.). DeepMind have since developed AlphaGo's successive algorithms, AlphaGo Master and AlphaGo Zero, along similar lines, namely, to 'learn' Go rules without any human guidance, but simply through processing millions and millions of games against another neural network, whose 'reward' and 'punishment' outcomes would thus train the algorithm on the rules and optimisation of gameplay (Silver et al. 2018).

The case here, then, is to re-think AlphaGo's creativity *on its own terms*, as with Vertov and Kleist's up-ending of grace and visuality in relation to marionettes and cinema. Here, we re-orientate the thinking of AlphaGo's creativity from its comparisons against moves by human players to *de-familiarise* its creativity so as to stand on its own terms. Move 37 was not generated on non- or unhuman terms as an alien stroke of creativity; it was *calculated* out of multitudinous values and possibilities arising from that particular position, and then chosen as the one which gave it the highest chance of ending up with more territory and thus a win.

But 'creativity' here, as framed on the algorithm's own terms, is not only its multi-layered[6] levels of calculation of the multitudes of moves from the multitudes of board positions to the multitudes of possible future board positions. Such level and extent of calculation resonate with the earlier discussion on creativity as an account of mechanised thought and logic applicable to human cognitive systems. There is a further nuance here: the algorithm's sense of creativity does not just lie in this manifold expansion of logical thinking (which might indeed be traced back to human interventions in training the discriminator network); it is also about the *speed* of its calculation through the multitudes of board position data. Speed, then, is really about space, or the demolition thereof, *à la* Paul Virilio (1991) who calls speed 'a primal dimension that defies all temporal and physical measurements' (18), and which directly results in 'the crisis of the whole', whereby the substantive, homogeneous and continuous gives way to the fractional, heterogeneous and discontinuous. Without veering too much into Virilio's ideas on speed which include the city, urban architecture and media, we can draw this line of thought on speed and space back to the *de-familiarisation* of creativity, whereby the terms of the algorithm are thus about neither its anthropomorphised independent learning nor even its mechanisation of logic and thought. Rather, they are about the algorithm's *speed* as the fracturing of the space-time of thought and as mirrored by the multiple splitting of its tree searches that is key to its algorithmic operation, and further constantly mapped with the algorithm's training from its datasets. Thus de-familiarised, we can shift our conceptualisation of creativity from cognitive processes in human terms to a different framework of space and spatial dimension across which the algorithm traverses with speed. The more tightly controlled the space is with unambiguous rules and outcomes, the more it is suited for discontinuous and fractional spaces, and the more the algorithm will thrive. The rationalisation of its 'creativity' in generating brilliant moves with large probabilities of achieving game-winning positions is thus not based on the depth of its logical thinking and learning as per the terms of rationalised human creativity. Instead, based on its computational calculatory processes, we can read it as a more de-familiarised conceptualisation of *space*, and begin the course of understanding it on its own terms.

Conclusion

> *The man-machine assemblage varies from case to case, but always with the intention of posing the question of the future.*
>
> – Gilles Deleuze (1985, 263)

Intelligent AI – more specifically, runaway intelligent AI which not only surpasses humans in general intelligence but whose capabilities are no longer under human control – has been identified as an existential risk, capable of

wiping out the human species (The Economist 2020b; Bostrom 2014). Stephen Hawking has pronounced to the BBC that 'the development of full artificial intelligence could spell the end of the human race' (Cellan-Jones 2014, n.p.). AI represents profound fears – culminating in our extinction – but also profound hopes in bettering life for humanity and life on Earth.

For these reasons – to ward off our fears and harness AI for betterment – the need to continually push for deeper understanding of AI is also correspondingly clear. The current rationalisation of AI persists in human terms, as is evident from even the most recent musings on the limitations of AI (The Economist 2020a): comparisons are consistently made with human learning and cognition, such as 'embodied cognition', or references to the 'irredeemably complex' nature of human minds (n.p.). In filling the gap of understanding why AI is still rubbish at doing elementary tasks that humans accomplish without much thinking, such as recognising a stop sign, the current approach appears to be to improve machine learning by developing it to resemble human learning; to write an algorithm that edges ever closer to human cognition, namely, achieve the dream of 'strong' AI.

But perhaps that is neither the question to ask, nor the appropriate task at hand. What is ultimately still not completely explainable is how algorithms think. And while this appears to be a technical question – to open the 'black box' – there are also other ways of arriving at an understanding to answer that question. John Seely Brown's (2017) words come to mind: 'We must also be willing to constantly reframe our understanding of the world. We must regrind our conceptual lenses, and regrind them often.' (n.p.). One of these 're-ground' conceptual lenses, as this chapter has argued, is the issue of intelligibility, insofar as the task of intelligibility is to make the unknown known. In this chapter, I have argued for an approach to an alternative intelligibility via a conceptual approach of de-familiarisation, one that shifts the terms of understanding away from the human to those of an Other. In making this argument, I am aware I have scythed through whole swathes of literature, if only hoping to at least demonstrate the broad contours of the argument. It is also clear that much more work needs to be done to hone this approach into a systematic methodology of a robust conceptual framework of intelligibility applicable to algorithmic systems. But the first step, at least, is taken in attempting to frame an alternative question. For, per this section's opening quotation by Deleuze, the question of our intelligent machines in human society is not about the technology, but always of our future.

Notes

[1] Randomness has also been long associated with creative work by humans, such as aleatory poetry or music by the Surrealists.

[2] More fully known as the Dartmouth Summer Research Project on Artificial Intelligence.

[3] A pigment containing sensory protein which, for many seeing animals, including humans, is located in the retina of the eye that converts light into an electrical signal.

[4] This is a neologism coined by Vertov, referring approximately to 'the quality of the cinema-eye', as noted by the editor and translator of *Kino-eye* (Vertov 1984).

[5] It should be clarified that, contrary to the media's hyperbole, human players have indeed played a fifth line move before and to productive results, as part of the move's strategy would be to emphasise influence and speed to battle for more centre territory in return for giving up peripheral territory, akin to a chess gambit of giving up a pawn piece for centre control. The more precise reading here might be that the *context* for that strategy (for centre territory) was not present in this game, which was what made AlphaGo's move so strange and thus 'creative', rather than to claim that it was an 'inhuman' move.

[6] There are three networks to the algorithm: *the policy network*, which 'come up with what would be the interesting spots to play' to build up 'a tree of variations'; *the value network*, which 'tells how promising is the outcome of [each] particular variation', and finally *the tree search*, which would look through different variations and 'try to figure out what will happen in the future'. *AlphaGo – The Movie*, 47:15–47:50.

References

Alpern, A. 1995. Techniques for Algorithmic Composition of Music. Hampshire College, Fall. Retrieved from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.9364&rep=rep1&type=pdf.

Ananny, M. and K. Crawford. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society*, 20(3), 973–989.

Appadurai, A. 2014. *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge: Cambridge University Press.

Bailey, J. 1996. *After Thought: The Computer Challenge to Human Intelligence*. New York: Basic Books.

Boden, M. A. (Ed.). 1996. *Artificial Intelligence*. San Diego, CA: Academic Press.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. 2nd ed. New York; London: Routledge.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Braidotti, R. and M. Hlavajova. 2018. *Posthuman Glossary*. London; New York: Bloomsbury.

Branigan, E. 2006. *Projecting a Camera: Language-Games in Film Theory*. New York: Routledge.

Brill, J. 2015. Scalable Approaches to Transparency and Accountability in Decisionmaking Algorithms: Remarks at the NYU Conference on Algorithms and Accountability. Federal Trade Commission, 28 February. Retrieved from: https://www.ftc.gov/system/files/documents/public_statements/629681/150228nyualgorithms.pdf.

Brown, J. S. 2017. Sensemaking in our Post AlphaGo World. Stanford University mediaX Keynote, February.

Brown, W. 2009. Man Without a Movie Camera – Movies Without Men: Towards a Posthumanist Cinema? In: W. Buckland (Ed.), *Film Theory and Contemporary Hollywood Movies*, pp. 66–85. New York: Routledge.

Burkholder, J. P., Grout, D. J. and Palisca, C. V. 2014. *A History of Western Music*. (9th ed). W.W. Norton & Company: New York.

Bush, V. 1945. As We May Think. *The Atlantic Monthly*, 101–108.

Carr, N. 2010. *The Shallows: What the Internet is Doing to Our Brains*. New York: WW Norton.

Cellan-Jones, R. 2014. Stephen Hawking Warns Artificial Intelligence Could End Mankind. *BBC News*. 2 December. Retrieved from: https://www.bbc.co.uk/news/technology-30290540.

Cohn, G. 2018. AI Art at Christie's Sells for $432,500. *New York Times*, 25 October. Retrieved from: https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html.

Csikszentmihalyi, M. 1988. Motivation and Creativity: Toward a Synthesis of Structural and Energistic Approaches to Cognition. *New Ideas in Psychology*, 6(2), 159–176.

Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. New York: HarperCollins.

Davies, S. 2011. Still Building the Memex. *Communications of the ACM*, 54(2), 80–88.

Deleuze, G. [1985] 1997. *Cinema 2: The Time-Image*, trans. Hugh Tomlinson and Robert Galeta, Minneapolis, MN: University of Minnesota Press.

Diakopoulos, N. 2016. Accountability in Algorithmic Decision Making. *Communications of the ACM*, 59(2), 56–62.

Dreyfus, H. 1972. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: HarperCollins.

Dreyfus, H. 1979. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Economist, The 2020a. Technology Quarterly: Artificial Intelligence and Its Limits. 13 June. Retrieved from: https://shop.economist.com/products/technology-quarterly-artificial-intelligence-and-its-limits

Economist, The 2020b. What's the Worst That Could Happen? 25 June. Retrieved from: https://www.economist.com/briefing/2020/06/25/the-world-should-think-better-about-catastrophic-and-existential-risks.

Foer, J. S. 2016. Technology Is Diminishing Us. *The Guardian*, 3 December. Retrieved from: https://www.theguardian.com/books/2016/dec/03/jonathan-safran-foer-technology-diminishing-us.

Gardner, H. 1985. *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.

Golumbia, D. 2009. *The Cultural Logic of Computation*. Cambridge, MA: Harvard University Press.

Grusin, R. (Ed.). 2015. *The Nonhuman Turn*. Minneapolis, MN: University of Minnesota Press.

Guyau, J-M. 1880. As Reproduced in Friedrich Kittler. 1999. *Gramophone, Film, Typewriter*, trans. Geoffrey Winthrop Young and Michael Wutz, pp. 30–33. Stanford, CA: Stanford University Press (first publ. in *Revue philosophique de la France et de l'étranger 5* (1880), 319–322).

Hayles, N. K. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago; London: The University of Chicago Press.

Hofstadter, D. 2000. *Gödel, Escher, Bach: An Eternal Golden Braid*, 20th Anniversary Edition. London: Penguin Books.

Ishiguro, K. 2005. *Never Let Me Go*. London: Faber and Faber.

Kaufman, J. C. and Baer, J. (Eds.). 2005. *Creativity Across Domains: Faces of the Muse*. Mahwah, NJ: Lawrence Erlbaum.

Kleist, H. [1810] 1972. On the Marionette Theatre. Trans. Thomas G. Neumiller, *The Drama Review: TDR*, 16(3), 22–26.

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. E. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August. Retrieved from: http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

McLuhan, M. [1964] 2013. *Understanding Media: The Extensions of Man*. Berkeley, CA: Gingko Press.

Metz, C. 2016. How Google's AI Viewed the Move No Human Could Understand. *Wired.com*, 14 March. Retrieved from: https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand

Miller, A. I. 1992. Scientific Creativity: A Comparative Study of Henri Poincaré and Albert Einstein. *Creativity Research Journal*, 5(4), 385–414.

Miller, A. I. 2000. *Insights of Genius: Imagery and Creativity in Science and Art*. Cambridge, MA: MIT Press.

Miller, A. I. 2019. *The Artist in the Machine: The World of AI-Powered Creativity*. Cambridge, MA: MIT Press.

Moffat, L. 2019. Putting Speculation and New Materialisms in Dialogue. *Palgrave Commun*, 5(11), n.p. Retrieved from: https://www.nature.com/articles/s41599-019-0219-8.

Moravec, H. 1998. When Will Computer Hardware Match the Human Brain?, *Journal of Evolution and Technology*, 1(1), n.p. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.136.7883.

Newell, A. and Simon, H. A. 1972. *Human Problem Solving*. London: Echo.
Parikka, J. 2015. *The Anthrobscene*. Minneapolis, MI: University of Minnesota Press.
Pasquale, F. 2015. The *Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
Sautoy, M. 2019. *The Creativity Code: How AI Is Learning to Write, Paint and Think*. Cambridge, MA: Harvard University Press.
Sawyer, R. K. 2006. *Explaining Creativity: The Science of Human Innovation*. Oxford: Oxford University Press.
Searle, J. R. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–457.
Searle, J. R. 1984. *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.
Silver, D. et al. 2018. A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-play. *Science*, 362, 1140–1144.
Sobchack, V. 1992. *The Address of the Eye: A Phenomenology of Film Experience*. Princeton, NJ: Princeton University Press.
Tegmark, M. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf.
Turing, A. 1950. Computing Machinery and Intelligence. *Mind*, LIX(236), 433–460.
Veale, T. and Cardoso, F. A. (Eds.). 2019. *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*. Cham: Springer.
Vertov, D. [1922], [1923] 1984. *Kino-Eye: The Writings of Dziga Vertov*. In Annette Michelson (Ed.), trans. Kevin O'Brien. Berkeley, CA: University of California Press.
Virilio, P. 1991. *The Lost Dimension*, trans. Daniel Moshenberg. New York: Semiotext(e).
Wald, G. 1950. Eye and Camera. *Scientific American*, 183(2), 32–41.
Yan, S. 2017. Google's AlphaGo A.I. beats world's number one in ancient game of Go. *CNBC*, 23 May. Retrieved from: https://www.cnbc.com/2017/05/23/googles-alphago-a-i-beats-worlds-number-one-in-ancient-game-of-go.html.
Zausner, T. 2007. *Artist and Audience: Everyday Creativity and Visual Art*. In R. Richards (Ed.), *Everyday Creativity and New Views of Human Nature: Psychological, Social, and Spiritual Perspectives*, pp. 75–89. Washington, DC: American Psychological Association.

Filmographic References

AlphaGo – The Movie. Video file, 1:30:27. YouTube. Posted by DeepMind, 13 March 2020. https://www.youtube.com/watch?v=WXuK6gekU1Y

CHAPTER 5

Post-Humanism, Mutual Aid

Dan McQuillan

Introduction

There is a growing awareness of the pitfalls of applying AI and algorithms to important social problems. Machine learning can only learn from past data and it's pretty clear that means a perpetuation of existing biases. This collision of AI with civil rights has led to corrective efforts at both technical and ethical levels (Feldman et al. 2015), (High-Level Expert Group on AI 2019). Meanwhile, other observers have pointed out the ways that AI adds its own asymmetries to an already skewed social landscape (Eubanks 2018). There's more data about the poor and marginalised because they are already most surveilled, and they are most surveilled because our social systems already categorise them as troublesome. As a result, any unfairness that algorithms add to the mix will fall more heavily on those who are already struggling the most. However, it's not only or even mainly data that shapes the politics of AI.

Langdon Winner wrote about the way particular technologies appear to have an inherent compatibility with particular socio-political systems (Winner 2020), so it's fair to ask what feedback loops connect AI and the societies into which it has emerged. This attentiveness may help to bring neglected features to the fore, to remind us of framings that are so pervasive they are usually ignored or to highlight new dynamics that are going to change more than just

How to cite this book chapter:
McQuillan, D. 2021. Post-Humanism, Mutual Aid. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 67–83. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.e. License: CC-BY-NC-ND 4.0

our means of technical ordering. For the purposes of this chapter it is important to ask these questions not only to provide a bigger picture of the problems of AI, but to predict the problematic nature of the likely reaction to it. The contention is that a reactive if understandable response to the harms caused by AI will itself risk feeding into wider crises instead of escaping them. The first step in unpacking this is to be more concrete about AI and about the resonances that are set up between its mathematical logic and its action in the world.

The Logic of AI

Actual AI is a form of machine learning; that is, an approach to computational problem-solving that iterates to a solution using data. It's different to more traditional forms of computational modelling: instead of trying to simulate the inner workings of a system, it's a transferable method of number crunching that simply requires sufficiently large amounts of training data. It's also different to traditional statistics, although it branches off from that family tree – where statistics tries to assess very precisely the relationships between variables and the robustness and possible error in the parameters, machine learning really doesn't care – its only goal is to make repeatable predictions. Whereas statistics is realist (in trying to model an underlying truth), machine learning is instrumentalist.

These may seem like nerdy distinctions but they have major consequences when it comes to social impacts, not least because of the inherited aura of infallibility that machine learning inherits from its associations with science and statistics. Like them, it's an approach that elevates quantitative analysis over any other form of insight. But machine learning is all about prediction and not about explanation. For machine learning, all that matters is generalisability; does the pattern learned from training data perform well on test data, in which case it can be let loose on the world.

When people talk about practical AI they mean machine learning as number crunching, and not any of the symbolic attempts to seriously emulate human reasoning that used to be called 'strong AI'. Even the term 'learning' has, at different times, meant a more profound attempt to understand the way we learn as embodied beings with life experience (Marcus 2018). But these approaches struggled to produce practical results, whereas the form of machine learning that simply means improving with 'experience' (i.e., with data) has succeeded spectacularly at previously impossible tasks. If current machine learning has a psychological analogue it is Skinner's behaviourism, where observable behaviours supersede introspection or any understanding of motivation in terms of meanings.

The form of machine learning which most accelerated the current 'AI revolution' is the artificial neural network, which symbolises all these important tendencies more vividly than any other. To begin with, a neural network sounds

like something to do with the brain, and while it's true that biological neurons were the original inspiration for the computational neurons that are densely interconnected in the layers of so-called deep learning, they are nowadays understood as arrangements constructed purely for predictive efficacy and not because of any residual correspondence to organic brains. Each neuron sums the weighted inputs from those in the previous layer that are connected to it, then applies an 'activation function' to the signal it passes to neurons in the subsequent layer (Nielsen 2015). Deep learning depends on orders of magnitude more training data than other methods of machine learning and its number crunching is on a previously inconceivable scale. The weights at each neuron are varied by the optimisation algorithm, and the optimal set of weights are become the 'model' that has been learned. The inner operations of a neural network are more opaque than other machine learning methods, making it extremely difficult to unravel their workings and understand in detail how they reached their conclusions (Feng et al. 2018).

The mathematical logic sets out to 'learn' by minimising a loss function; loosely, the difference or distance between its current predictions and the training data. This requires a well-defined objective to optimise on, which is typically the target classification of interest. The formalism of machine learning expresses any worldly context in terms of a fixed set of possible outcomes which are functions of the input features of each instance of data. Taking the world to be at least functionally composed of entities and their attributes is a philosophical commitment to an ontology of separate individuals and objects, while the very idea of optimisation is itself value-laden; AI promotes a market-friendly and mathematised utilitarianism.

The mathematical iterations of machine learning are implacable in their pursuit of the assigned outcome, so harnessing them to a messy social goal inevitably sets the stage for shockwaves of unintended consequences. Given the requirement for the context of interest to be expressed as purely numerical (and measurable) abstractions, it is also inevitable that the outcome being optimised on is itself a distant proxy for the desired policy goal (Malik 2020). For machine learning, the external environment is a fixed set of givens; woe betide those who rely on it when the underlying distribution shifts (even though it is the nature of the world to constantly change). AI is haunted by the underexamined constructs it assumes in order to make the world amenable to its methods; above all, that the world is essentially a mechanism which can be manipulated by the adjustment of its parameters (Wu 2019).

AI is undoubtedly successful at tackling data sets which were previously offlimits in terms of scale and complexity. Ignorance is strength; by bypassing the need to explain and moving straight to prediction, AI provides a ready-to-hand tool for intervention. Never mind that correlation is not causation; that explainable is not the same as explanatory (even if I can tell which combination of parameters is most significant in determining the final classification, it doesn't provide me with a causal understanding). The predictions of AI are a

dilution of science because they are not the expression of a hypothesis but simply an extrapolation from a set of data points. And yet AI is performative; by intervening in the world and reorganising the phenomena it claims to describe, it brings our experience into closer alignment with its constructs (Mackenzie 2008). This invalidates the methods as a form of statistical insight but makes it very effective as a mode of subjection.

Automated Segregation

As the logic of AI migrates from the abstract mathematical space of tensors to the space of real social tensions, it comes to bear in specific ways. First and foremost of these is automated segregation.

There is nothing personal about the predictions of AI – at root, they are always some form of labelling in terms of 'people/objects like you'. As an offshoot of the statistics family tree, machine learning's classifications are governed by the heavy hand of the central tendency (Malik 2020); that is, the principle that there exists a central or typical value for a probability distribution. Predictions about you are centred on some recomposition of the past behaviours of those with similar attributes. Such a prediction may be useful for an institution dealing with large numbers of people at a distance, but it is not about you at all. The subjects of AI are represented as entities with attributes inasmuch as they are present to the algorithm as vectors of values.

Understanding how this plays out in terms of the distribution of benefits and harms means reflecting on resonances between the intrinsic logics of AI and our social institutions. When these algorithms execute mathematical operations of classification, ordering and ranking that carry over into our lived experience, they offer support to certain ways of doing and limit the likelihood of certain others. The significance of these resonances will vary with context. It doesn't seem problematic to classify and rank the likely failure modes of an engineering infrastructure but it becomes far more delicate the closer we apply the same approach to other people. Questions of class and classification, the assumption of certain orders as normative, and ideas about rank and hierarchy are so deeply embedded in our psyches and societies that calculative methods with the same logic act as an amplifying stimulus.

When we deal with social classification we can't escape questions of power. The distribution of power in society may be complex and multivalent but it is also highly asymmetric. This not a problem created by machine learning, of course, but machine learning was produced within these structures of power and it is acting back on them. While the mathematics of AI may be expressed as matrices, it is a human activity that is inescapably immersed in history and culture. AI acts as an activation function for specific social tendencies. As an idea, or ideology, AI seeks to escape association with these worldly concerns by

identifying with a pure abstraction and a neoplatonic purity of forms (McQuillan 2017), but this is only a plausible cover story to those who occupy an already privileged standpoint. The optimisation functions themselves will in general also be defined from these positions of social privilege. When AI talks in terms of 'models' it means the learned weights in a neural network, not the classic idea of a model that describes inner workings; as the goal is prediction not explanation, it is not supplying insights that could be used for causal interventions.

For AI's impacts on the ground, the operative concerns are discrimination and segregation. AI is a racist technology, in the sense that AI operates so as to segregate, and racism itself can be understood as a technology of segregation (Lentin 2018). This is easy to see when it comes to facial recognition, one of the most egregious applications that AI has so far gifted to society. It is not just that facial recognition seems to perform less well on people of colour, it is that it carries out what Simone Browne calls 'digital epidemermalisation': 'the exercise of power cast by the disembodied gaze of certain surveillance technologies … that can be employed to do the work of alienating the subject by producing a "truth" about the body and one's identity (or identities) despite the subject's claim' (Browne 2015). In other words, AI's operations of facial classification are actually reconstructing the category of race for subsequent intervention (Stark 2019). Facial recognition itself forces race onto a face. Clearly, any alternative approach to AI must be at the very least decolonial. When applied to people, AI's operations with entities and attributes distil us down to innate differences. It excludes perspectives from critical race studies which might question the construction of identity gradients, nor does it acknowledge any sociological understanding of why people might be trapped in particular social patterns.

On this basis, we can confidently say that the overall impact of AI in the world will be gendered and skewed with respect to social class, not only because of biased data but because engines of classification are inseparable from systems of power. Writers like Virginia Eubanks highlight some of the ways this comes to pass for social class as well as race; how it seems to always be the poorest and most marginalised who bear the brunt of collateral damage from algorithmic systems even when the bureaucrats involved are making sincere efforts to be fair (which they often aren't) (Eubanks 2018). The data demands of AI mean that the pattern of having to trade private personal information for services will become even more invasive. The optimisations of AI act as an inverse intersectionality, applying additional downward pressure on existing fissures in the social fabric. Like Eubanks, we should be asking what specific forms these fractures will take, and how to recognise them. One marker will be the emergence of machinic moralism. The more that AI is seen as a solution to austerity, the more its classifications and rankings will be enrolled in the rationing of goods and the assigning of sanctions. AI will be put in the position of deciding between the deserving and the undeserving. The most advanced forms of computation seem destined to re-enact a Victorian morality.

We can expect a lot more 'production of truth … despite the subject's claims', as Browne puts it. Not only digital epidermalisation but epistemic injustice, a concept developed by philosopher Miranda Fricker through her analysis of the ways women have historically been degraded both in terms of the credibility of their testimony and in their very capacity as a knower (Fricker 2009). The operations of AI at scale will produce 'continuous partial states of exception' (McQuillan 2015) where the access to shared social, political and economic rights will be adaptively suspended for particular cohorts with the justification (if one is ever given) that this is required for an overall optimisation of the system. AI itself is indifferent to unintended consequences and collateral damage, as its version of statistical methods dispenses with robust estimates of error. The effects on the people 'in the loop', that is, operating within these systems, will be the production of carelessness because they won't be in a position to challenge the opaque but apparently empirical judgements of the system. The wider output will be a scaling of callousness, as the specific effects on the most vulnerable and least able to speak up will be occluded by the abstractions that drive the algorithms.

Clearly, both the logics and the impacts of AI are rooted in ways of approaching the world that go deeper than a recent acceleration in computational capacity or the availability of plentiful training data. It's important to try to characterise this overall stance towards the world, not only to challenge it but to be wary of challenging it in ways that simply make the problems worse.

Machine Learning Modernism

Machine learning can't simply be summed up as the implementation of a particular philosophy. It is an active and performative force in the world, which has the potential to change the conditions of thought itself. However, it can be useful to point out how much machine learning inherits from modernism.

To start with, AI is a form of computation and its algorithms are expressions of computational thinking, that is, decomposition, pattern recognition and abstraction (Wing 2008). While the first two are the most apparent – a world decomposed into data and acted on by statistical pattern finding – it's abstraction that most defines the character of AI's impact. AI is above all a mode of abstraction that allows any issue to be treated as a mathematical optimisation problem. Any aspect of the world deemed relevant must be quantified and normalised for inclusion in this operation; in the innermost workings of deep learning, all data is rendered as vectors of numbers between zero and one. AI is deeply reductive, asserting in effect that it can predict the unfolding of a system in terms of those elements which can be reduced to data, and the only attributes of the world that count are those that can literally be counted. Unlike science, which at least seeks a careful explanation of how a layer of reality arises from the interactions of simpler elements, AI is epistemologically careless; it's

not concerned about whether its reductions represent anything more fundamental as long they produces repeatable predictions.

Applied to the social realm, AI takes on the kind of reductiveness that was elucidated by Heidegger; that is, a reduction of being itself. The being of entities subject to AI's 'enframing' is reduced to a calculative order which 'drives out every other possibility of revealing' (Heidegger 2013). Making the social world into data for the benefit of AI is to convert ourselves into a standing reserve for optimisation and prediction. This echoes the way that dualistic metaphysics combined with the capitalism system reduce the natural world into raw material and resource. AI models the world, but only to get something out of it.

The discourse of AI uses terms like 'model' and 'representation' pretty interchangeably. They are used as shorthand for the nexus of feature set and algorithmic architecture which are being applied to get a result. The layers of a deep learning model apply successive transformations to the input feature space that will eventually allow it to be distilled down to required target labels. Each layer contains a different representation of the original data, by way of the weights at each node in the layer. A prominent practitioner likens it to trying to uncrumple a paper ball; each hand movement is like the geometric transformation carried out by one layer, and 'deep learning models are mathematical machines for uncrumpling complicated manifolds of high-dimensional data' (Chollet 2017).

What gets easily overlooked in the intriguing detail is the form worlding that is taking place; what resonances that are set up by dealing with the world this way. Prioritising representations over presence may be necessary for modelling, but it is a move that has political as well as philosophical implications. A fundamental operation of social power at any level is the power to represent someone's reality back to them in a way that is asserted as being more real. AI is at the stage of having representations that are opaque and barely understandable even to those who produce them, while these representations are increasingly relied on to make robust social interventions.

Looking at the data as it is transformed through the layers evokes another essential aspect of AI; the imposition of equivalence. By representing attributes as numbers in a vector they are made commensurable whether they happen to represent 'likes' on Facebook or the oscillations of an ICU heart-rate monitor. The values of each feature are traded against each other in the iterations of the optimising algorithm as it seeks to descend to the minima of its loss function. AI effects the same operation of equivalence as money; rendering objective diversity in terms of values that can be traded against each other.

AI is marked by the same modes of abstraction, representation and equivalence as the rest of modernity. It applies an instrumental rationality that subsumes social relationality and material under a single regime of equivalence which discards the incommensurable (Horkheimer and Adorno 2002). The single optic of optimisation admits no outside; the methods of machine learning are more than generalising, they are universalising. AI carries on the

tradition of modern thought that Whitehead criticised as 'explaining away' – by taking its abstractions as something concrete, everything that does not fit into the schema is denied the status of proper existence (Whitehead 1997). AI operates as automated segregation, in the same key as racism, patriarchy and the class system, applying an inevitable hierarchy of humanness to it subjects. The reaction to the evils of colonialism from liberation thinkers like Fanon was to demand colonial subjects' rightful membership in the category of human (Fanon 2005). Given the promotion of such profoundly alienating and dehumanising processes it is only natural that the obvious callousness of AI will be opposed by calls for return to human values and to a valuing of the human. Like the writers and activists of postcolonialism, the unhappy subjects of algorithmic governance will come to demand their full membership of the category of humanity. It's in this reaction, though, that further perils lie in wait.

Reactive Humanism

The call for a post-algorithmic humanity that leaves no-one behind needs to find a way to escape the legacy of humanism. Historical definitions of humanness tend to carry with them the assumption of human exceptionalism; that is, the uniqueness of the human in relation to the animal and material worlds. Whereas this originally had religious roots, the secular version born out of the Enlightenment centred on consciousness, morality and particular notions of reason. Over time, these notions have deeply shaped the psycho-political landscape we still inhabit.

There's a bifurcation in our way of understanding the world that still acts to separate us from the world. Whatever the success of the scientific approach, we still seem to distinguish ourselves as having an agency and a will that is different from the deterministic conception of nature that science implies. The foundational distinction between observer and observed remains, despite the challenge of quantum mechanics, and still cascades into operations as mundane as those of applied machine learning.

The bounded individualism that comes with humanism is not merely a metaphysical curiosity but an active factor in our political economy. Along with the rational consciousness of individual actors, the very concept of the separated individual underpins neoliberalism and classical economics. Humanism, as the species-separateness of humanity, also provides the logic for treating the rest of nature as a resource, as an externality to be plundered at will for productivity. And this 'nature' includes, of course, those people who are in whatever way seen as less than fully human.

Humanism is a vector for some of the same problems that plague a modernist AI. The historically constructed idea of the human was that it was endowed with the ability to make moral choices. It is exactly this aspect that led Nietzsche to question the idea of the 'I' as the illusion of continuity that enables morality;

that is, the identity that is the cause of the actions and so is deserving of reward or punishment. For him, the moral concept of the 'I' is projected onto events in the world (Nietzsche 1998). As we have seen, AI is already projecting algorithmic forms of moral attribution into its predictions, and in line with Nietzsche's original critique this moralism acts in the interests of some rather than all. A reactive humanism would only modify this mode of moralising rather than replacing it.

Similarly for actions that contribute to climate change; modernist AI is part of a wider system where extraction follows closely on the heels of abstraction, where everything of the world is seen as a utilitarian resource, not a fragile component of a self-regulating ecosystem. AI is making its own contribution to global warming via the exponential increase in computing required for the latest deep learning models and the consequent carbon emissions (Strubell, Ganesh and McCallum 2019). But humanism itself, as the vision of the human as separate and subject to special rules, is the precursor of worldviews that have created the possibility of the Anthropocene.

Perhaps the most immediately dangerous aspect of human exceptionalism is the one linked directly to the definition of AI; the question of consciousness and superior intelligence. Humanism sees the spark of rational intelligence as a marker of uniqueness. The field of AI meanwhile, while its current best practice is the steamhammer of statistical prediction, still holds on to the idea that this narrow form of computational 'intelligence' is a foothill on the way to artificial general intelligence; that is, machines that can think like us (Hodson 2016). Both humanism and AI understand intelligence as something hierarchical, that can be ranked from lower to higher. But ranking on the basis of intelligence is the backbone of race science, the pseudo-empirical justification for colonialism and white supremacy (Golumbia 2019). Not only that, but the implicit ranking of human worth by IQ has been a historic justification for eugenics through programmes such as forced sterilisation, and is re-emerging at the time of writing in terms of criteria for COVID-19 triage that downgrade those seen as in some way disabled (NoBodyIsDisposable 2020).

In short, by reacting to the dehumanising effects of automated segregation by reaching for a ready-to-hand humanism, we are not escaping challenges like climate change and the politics of racial supremacy, or the underlying assumptions of human exceptionalism and subjectivity that are based on a dualistic metaphysics.

New Materialist AI

We are seeking an alternative AI that avoids the dehumanisation induced by automated segregation. Where AI is an engine of injustice it is because it intensifies the reductiveness, representationalism and universalism that privileges an existing social hegemony. At the same time, we recognise the dangers of a

reactive humanism; of fetishising human uniqueness in a way that perversely ensures some humans don't make the cut, and whose bordering off of the rest of the material world reduces it to an exploitable resource. We're looking for the possibility of post-AI that is at the same time postcolonial and posthuman (Mitchell 2015).

However, this is not an exercise in reconciling theories. The aim is to sketch out a practice, or a praxis; an approach that can alter the current performativity of AI, not just critique it. So whatever we draw from the field of new materialism (Sanzo 2018), from its fluidity of being and its immersive relationality, it needs to retain a clear possibility of agency. The aim is not simply to overcome dualism and reattach ourselves to a reality we have misunderstood, but to act politically against the amplification of injustice.

The idea of a new materialist AI is important to explore because of the way it opens questions about the boundaries and hierarchies constructed between beings, and concerns itself with what these structures obscure and erase. The starting point here is the materiality of the world, but without any assumptions about meanings. The focus is on the way the material world and social meanings are part of a process of co-construction that is at the same time marked by relations of power. In other words, there's a non-dualistic politics acting at the point of intersection between subjectivity and matter.

AI takes sides here through its promotion of a worldview whose rigid categories of meaning have real material consequences. The AI we know acts on and through individualised and itemised entities, and carries forward the political payload in terms of a world of atomised individuals and externalised nature. It reinforces particular boundaries in terms of what exists and how it gets distributed. Seeking an alternative AI suggests it's worth exploring a more posthuman approach, focusing on the interactions from which the familiar phenomena are emergent. Instead of an AI that takes the position of an outside observer, we can start with the idea of being as immersive and embedded, undermining the gaze of objectivity from which a single optimised truth can be affirmed.

An immersive and emergent perspective on the world also suggests the idea of agency is no longer confined to the human but is distributed across the sociomaterial landscape. However, there are some drawbacks to distributed agency if we are attempting to construct a political project, especially the kind of distributed agency that falls under the umbrella of Actor-Network Theory. If the starting point is to describe complex networks of material-semiotic actors in a way that makes intentionality a secondary phenomenon, we open up the space for pacification; describing the becoming of what is, rather than striving for what should be. We are not simply seeking to reconnect to a non-dualistic reality, but to change it.

One way to overcome dualism is by starting from the intertwining of phenomena that were previously classified as distinct. Our approach is to follow Karen Barad by identifying the way both the material world and subjects of

knowledge emerge through the actions of what she calls material-discursive apparatuses (Barad 2007). Her ideas of 'agential realism' bring together perspectives from Foucault and from the quantum philosophy of Niels Bohr.

The findings of quantum mechanics led Bohr to reject the assumption that the world is made of determinate objects with well-defined properties independent of specific experimental arrangements. Instead, phenomena are determined by the wholeness of the measurement event. The particular way this is put together produces a particular division between the object and the observation, which has the consequence of materialising some properties while excluding others. This is an irrefutable experimental result at a quantum level but, as Barad spotted, parallels the way social constructivism analyses the formation of subjectivity through the operations of power. So she also draws on Foucault's notion of a heterogeneous apparatus ('dispositif') of physical, administrative and knowledge structures that produces both us as social subjects and the societies we inhabit.

Barad uses the term 'intra-action' to talk about the mutual constitution of objects and subjects. Phenomena are produced by the intra-actions of apparatuses, which are active not passive; they are not just measuring instruments but boundary-drawing practices. These apparatuses are neither the laboratory instruments of Bohr or the semiotic institutions of Foucault but both at the same time; they are 'material-discursive' apparatuses.

Humans, according to Barad are part of the ongoing reconfiguration of the world produced by these apparatuses. 'Humans (like other parts of nature) are of the world, not in the world, and surely not outside of it looking in. Humans are intra-actively (re)constituted as part of the world's becoming' (206). We have agency through our participation in the iterative production of reality and the space that exists within that for new possibilities. Human practices are 'agentive participants' in the way phenomena are 'sedimented out' of this ongoing process. The idea of sedimentation makes us pay attention to the fact that the world we experience, and our experiencing of it, are not the starting point for analysis but are already the product of active processes. Material and meaning are not separate but the depositions from a dynamic that is not dualistic. Seen in this light, the systems of AI are aspects of a material-discursive apparatus that are themselves sedimented out from other material-discursive systems, all of which are open to participatory reworking.

AI as we currently know it is an instance of representation gone wrong, built on a foundation of the same mistake spread across the philosophical and political landscape. AI is not simply a layer of representation imposed on a solid ontology but part of a stack of practices that splits subject and object all the way down. Its automated segregations are boundary-drawing practices that act in the world. We're not going to find a line of flight by means of a better mapping, a more accurate metaphysical correspondence. By throwing our hat in with agential realism we're also trying to switch allegiance to a process philosophy,

where the emphasis is not on being but on becoming. Instead of mirroring reality, it's about the making of realities; the important thing is to make meanings that matter. This approach to an alternative AI is pragmatic, based on the principle that knowing the world is inseparable from agency within it. It's a situated metaphysics that is committed to making a difference by not only overcoming dualisms but by overcoming the division between knowing and caring.

Care exists in the shadow of the kind of detachment and abstraction that is valorised by AI. Care starts from a concern with exclusions and boundaries, and with the asymmetry of consequences for the most vulnerable (Bellacasa 2017). Dematerialising the divides between observer and observed, subject and object, humanity and nature is an opening to a kind of caring cosmopolitics; the term Isabelle Stengers uses for being attentive and responsive to the multiples of being with which we are entangled and co-constituted (Mitchell 2015). This acceptance of heterogeneity without fixed boundaries and an interdependence that is also an intra-dependence gives us the basis for an approach that is both posthuman and postcolonial. A situated caring means starting from the experiences of those at the margins, from ways of knowing that can challenge the erasure of lived experience by the ideology of efficiency, in order to counter the algorithmic extension of carelessness. The question that remains is how we might go about applying ideas of agential realism and care to produce an alternative form of AI.

Post-AI Politics

Moving beyond the injustices powered by institutional machine learning means moving beyond representation to social recomposition. The material-discursive apparatus of AI acts to reinforce the wrong answer to the question of how to be together. It co-produces ineffective concepts of fairness and skewed distributions of social goods that reinforce each other and the status quo. As apparatuses that make meanings, the current instances of AI optimise Mark Fisher's invocation of Frederic Jameson, that it is easier to imagine the end of the world than an end to capitalism (Fisher 2009) – it actively contributes to the former and erases the possibility of the latter. And yet we are agentive participants in this wider system whose intra-actions are open to reworking. We can be part of altering these boundary-drawing practices. The question of recomposition is the question of whether agential realism can be composed as collective action.

AI is already earmarked as a solution to austerity through calculative rationing and optimised extensions of precarity and scarcity. At the time of writing, in the midst of the COVID-19 pandemic, the prospect of heightened post-pandemic austerity in an even more datafied and surveilled environment only boosts the likelihood of algorithmic optimisation being substituted for social justice. For many of the issues where AI is being applied to single out those deserving

intervention, that is, the most 'risky', the situation would be made fairer overall if resources were redistributed to lower the overall risks. For example, rather than sinking resources into deep learning models that try to predict which parents will abuse their children, why not acknowledge that poverty and drug abuse are highly correlated with child abuse and put resources into reducing poverty and providing more accessible drug treatment services (Keddell 2015). Instead of seeing the problem as one of identifying the most risky 'entity', it's about starting from the inseparability of all entities and a recognition that they are co-constituting; not just an ethics of relationality (Birhane and Cummins 2019) but an ethico-ontology of relationality.

This approach doesn't depend on a top-down restructuring that moves from metaphysics to science to social policy. It simply requires an openness to the speculative starting point of agential participation motivated by care. It means acting 'as if' the intra-actions of a material-discursive apparatus could be determined by caring about the consequential meanings that are produced. In more familiar political terms, it means acting as if solidarity were not only a stance but a core facet of being, as if mutual aid was not simply a choice made after social reality was sedimented out but a driving element in the iterative reproduction of the world. What we are currently experiencing instead is not an established order but the entropic disorder established by apparatuses that utterly lack the balance necessary to sustain us or our world.

The current and ongoing sedimentation of reality has its own pyroclastic momentum. What we can hope for at this time is to both slow it down and to diffract it through the introduction of difference. The pragmatic approach proposed here is to introduce structures that 'slow the universalizing process by unsettling existing assumptions, boundaries and patterns of political action' (Mitchell 2015). For this role we propose people's councils for AI (McQuillan 2018). People's councils are bottom-up, federated structures that act as direct democratic assemblies. The mutual encounters and consensus-making of people's councils are themselves transformative in terms of creating different relationalities. The purpose of people's councils is to become a mode of 'presencing', of forcing the consideration of the unconsidered, or more fundamentally of reordering the idea of AI such that its production of pairings of concepts and material effects iterates towards an actually different society.

People's councils, based on solidarity and mutual aid, are an attempt to inoculate our meaning-making structures against fascism. The operations of fascism past and present show the ability to embrace technology as technique while replacing modernism with a cult of authoritarian traditionalism, a disturbing tendency already visible in the 'dark enlightenment' narratives of neoreaction circulating in Silicon Valley (Haider 2017). AI as we know it forms a harmonic with neoreaction's 'near-sociopathic lack of emotional attachment' and 'pure incentive-based functionalism' (MacDougald 2015). People's councils are a diffraction of AI, introducing the difference of care as a mode of interference and superposition.

People's councils are not a form of collectivised humanism or an attempt to re-centre the human as the only actor that matters, but a situated intervention in the ongoing reiteration of wider conditions. They are directed at the creation of new possibilities. We still have the possibility of reforming the structures, such as AI, that are increasingly becoming part of co-constituting us and our material world. The proposition is that these can be modes of differencing rather than of machinic modernism. The danger is that the mounting collateral damage caused by pervasive AI will drive a more atavistic response, whose boundary-drawing practices will increasingly be determined by notions such as racial superiority or eugenic justifications that some should be left to die in order to preserve the economy and/or the planet, narratives that we can already see emerging as a neoliberal and fascist reaction to the COVID-19 pandemic and as Malthusian responses to climate change.

As a rule of thumb, we should examine every situation where AI is being offered as a solution and ask how on-the-ground collective action might enable a radical commoning of both risks and resources. Instead of a technocratic solution to precarious labour, for example, that imposes some spurious metric of fairness on a structure that embodies injustice, we look to a complete socialisation of the relations and materialities involved. This happens, for example, when workers react to layoffs by occupying their workplaces and transforming material production in collaboration with the local community (Pazos 2018). The only material-discursive politics consistent with a cosmopolitical care is a radical commoning.

As Donna Haraway reminds us, our intra-actions and interdependencies stretch across vast fields of biota and abiota. Nevertheless 'the doings of situated, actual human beings matter. It matters with which ways of living and dying we cast our lot rather than others' (Haraway 2016). Changes start with grassroots collectives who are prepared to take on the necessary activities of repair and resistance. The modelling which needs to take priority is not that delivered from on high by vast structures of computation but the modelling to each other of different ways of living and caring through mutual aid. Reclaiming political agency from engines of abstraction without the need for the rigid boundaries of humanism means taking solidarity as the starting point for our becoming.

References

Barad, K. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press Books.

Bellacasa, M. 2017. *Matters of Care*. (3rd edn.). Minneapolis, MN: University of Minnesota Press.

Birhane, A. and Cummins, F. 2019. Algorithmic Injustices: Towards a Relational Ethics. *ArXiv:1912.07376[Cs]*, December. Retreived from http://arxiv.org/abs/1912.07376.

Browne, S. 2015. *Dark Matters: On the Surveillance of Blackness*. Durham, NC: Duke University Press Books.

Chollet, F. 2017. *Deep Learning with Python*. (1st ed.). Greenwich, CT: Manning Publications Co.

Crofts, P. and van Rijswijk, H. 2020. Negotiating 'Evil': Google, Project Maven and the Corporate Form. *Law, Technology & Humans*, 2(1), 75–90.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Fanon, F. 2005. *The Wretched of the Earth: Grove Atlantic*. Grove Paperback. https://groveatlantic.com/book/the-wretched-of-the-earth.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. KDD 2015. New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/2783258.2783311.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P. and JBoyd-Graber, J. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3719–3728. Brussels, Belgium: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D18-1407.

Fisher, M. 2009. *Capitalist Realism: Is There No Alternative?* Winchester: John Hunt Publishing.

Fricker, M. 2009. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Golumbia, D. 2019. The Great White Robot God. *Medium*, 1 February 2019. https://medium.com/@davidgolumbia/the-great-white-robot-god-bea8e23943da.

Haider, S. 2017. The Darkness at the End of the Tunnel: Artificial Intelligence and Neoreaction. *Viewpoint Magazine*, 28 March 2017. https://www.viewpointmag.com/2017/03/28/the-darkness-at-the-end-of-the-tunnel-artificial-intelligence-and-neoreaction/.

Haraway, D. 2016. Tentacular Thinking: Anthropocene, Capitalocene, Chthulucene. *E-Flux*, September. https://www.e-flux.com/journal/75/67125/tentacular-thinking-anthropocene-capitalocene-chthulucene

Heidegger, M. 2013. *The Question Concerning Technology, and Other Essays*. Reissue edition. New York; London: Harper Perennial Modern Classics.

High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. Text. Shaping Europe's Digital Future – European Commission. 8 April 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Hodson, H. 2016. Revealed: Google AI Has Access to Huge Haul of NHS Patient Data. *New Scientist*, 29 April 2016. https://www.newscientist.com

/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/.

Horkheimer, M. and Adorno, T. W. 2002. *Dialectic of Enlightenment*. In Noeri, G. S. Translated by E. Jephcott. (1st edn.). Stanford, CA: Stanford University Press.

Keddell, E. 2015. The Ethics of Predictive Risk Modelling in the Aotearoa/New Zealand Child Welfare Context: Child Abuse Prevention or Neo-Liberal Tool? *Critical Social Policy* 35 (1): 69–88. DOI: https://doi.org/10.1177/0261018314543224.

Lentin, A. 2018. The Future Is Here – Revealing Algorithmic Racism. *Alana Lentin.Net* (blog). 22 October 2018. http://www.alanalentin.net/2018/10/22/the-future-is-here-revealing-algorithmic-racism

MacDougald, P. 2015. The Darkness Before the Right. *The Awl*, 28 September 2015. https://www.theawl.com/2015/09/the-darkness-before-the-right/.

Mackenzie, D. 2008. *An Engine, Not a Camera: How Financial Models Shape Markets*. (1st edn.). Cambridge, MA: The MIT Press.

Malik, M. M. 2020. A Hierarchy of Limitations in Machine Learning. *ArXiv:2002.05193 [Cs, Econ, Math, Stat]*, February. http://arxiv.org/abs/2002.05193.

Marcus, G. 2018. Deep Learning: A Critical Appraisal. *ArXiv:1801.00631 [Cs, Stat]*, January. http://arxiv.org/abs/1801.00631.

McQuillan, D. 2015. Algorithmic States of Exception. *European Journal of Cultural Studies* 18 (4–5), 564–576. DOI: https://doi.org/10.1177/1367549415577389.

McQuillan, D. 2017. Data Science as Machinic Neoplatonism. *Philosophy & Technology*, August, 1–20. DOI: https://doi.org/10.1007/s13347-017-0273-3.

McQuillan, D. 2018. People's Councils for Ethical Machine Learning. *Social Media + Society* 4 (2). DOI: https://doi.org/10.1177/2056305118768303.

Mitchell, A. 2015. Posthumanist Post-Colonialism? *Worldly* (blog). 26 February 2015. https://worldlyir.wordpress.com/2015/02/26/posthumanist-postcolonialism

Nielsen, M. A. 2015. Neural Networks and Deep Learning. http://neuralnetworksanddeeplearning.com.

Nietzsche, F. 1998. *Twilight of the Idols*. Reissue edition. Oxford: Oxford University Press.

NoBodyIsDisposable. 2020. Open Letter to Care Providers and Hospitals. March 2020. https://nobodyisdisposable.org/open-letter

Pazos, A. 2018. Ours to Master and to Own – We Visit Viome, Greece's Only Worker-Managed Factory. *Jacobin*, 10 June 2018. https://jacobinmag.com/2018/10/viome-self-management-factory-takeover-greece.

Sanzo, K. 2018. New Materialism(s). *Critical Posthumanism Network* (blog). 25 April 2018. http://criticalposthumanism.net/new-materialisms

Stark, L. 2019. Facial Recognition Is the Plutonium of AI. Association for Computing Machinery. DOI: https://doi.org/10.1145/3313129.

Strubell, E., Ganesh, A. and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. Florence, Italy: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P19-1355.

Whitehead, A. N. 1997. *Science and the Modern World*. New York: Simon and Schuster.

Wing, J. M. 2008. Computational Thinking and Thinking about Computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366 (1881): 3717–3725. https://doi.org/10.1098/rsta.2008.0118.

Winner, L. 2020. *The Whale and the Reactor: A Search for Limits in an Age of High Technology,* Second Edn. Chicago, IL: University of Chicago Press.

Wu, J. 2019. Optimize What? *Commune*, 15 March 2019. https://communemag.com/optimize-what

PART 2

Discourses and Myths About AI

CHAPTER 6

The Language Labyrinth: Constructive Critique on the Terminology Used in the AI Discourse

Rainer Rehak

Introduction

In the seventies of the last century, the British physicist and science fiction writer Arthur C. Clarke coined the phrase of any sufficiently advanced technology being indistinguishable from magic – understood here as mystical forces not accessible to reason or science. In his stories Clarke often described technical artefacts such as anti-gravity engines, 'flowing' roads or tiny atom-constructing machinery. In some of his stories, nobody knows exactly how those technical objects work or how they have been constructed, they just use them and are happy doing so.

In today's specialised society with a division of labour, most people also do not understand most of the technology they use. However, this is not a serious problem, since for each technology there are specialists who understand, analyse and improve the products in their field of work – unlike in Clarke's worlds. But since they are experts in few areas and human lifetime is limited, they are, of course, laypersons or maybe hobbyists in all other areas of technology.

How to cite this book chapter:
Rehak, R. 2021. The Language Labyrinth: Constructive Critique on the Terminology Used in the AI Discourse. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 87–102. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.f.

After the first operational universal programmable digital computer – the Z3 – had been invented and built in 1941 in Berlin by Konrad Zuse, the rise of the digital computer towards today's omnipresence started. In the 1960s, banks, insurances and large administrations began to use computers, police and intelligence agencies followed in the 1970s. Personal computers appeared and around that time newspapers wrote about the upcoming 'electronic revolution' in publishing. In the 1980s professional text work started to become digital and in the 1990s the internet was opened to the general public and to commercialisation. The phone system became digital, mobile internet became available and in the mid-2000s smartphones started to spread across the globe (Passig and Scholz 2015).

During the advent of computers, they were solely operated by experts and used for specialised tasks such as batch calculations and book-keeping at large scale. Becoming smaller, cheaper, easier to use and more powerful over time, more and more use cases emerged up to the present situation of computer ubiquity. More applications, however, also meant more impact on personal lives, commercial activities and even societal change (Coy 1992). The broader and deeper the effects of widespread use of networked digital computers became, the more pressing political decisions about their development and regulation became as well.

The situation today is characterised by non-experts constantly using computers, sometimes not even aware of it, and non-experts making decisions about computer use in business, society and politics – from schools to solar power, from cryptography to cars. The only way to discuss highly complex computer systems and their implications is by analogies, simplifications and metaphors. However, condensing complex topics into understandable, discussable and then decidable bits is difficult in at least two ways. First, one has to deeply understand the subject and second, one has to understand its role and context in the discussion to focus on the relevant aspects. The first difficulty is to do with knowledge and lies in the classical technical expertise of specialists. But the second difficulty concerns what exactly should be explained in what way. Depending on the context of the discussion, certain aspects of the matter have to be explicated using explanations, metaphors and analogies highlighting the relevant technical characteristics and implications. Seen in this light, this problem of metaphors for technology is not only philosophically highly interesting but also politically very relevant. Information technology systems are not used because of their actual technical properties, but because of their assumed functionality, whereas the discussion about the functionality is usually part of the political discourse itself (Morozov 2013).

Given the complexity of current technology, only experts can understand such systems, yet only a small number of them actively and publicly take part in corrective political exchanges about technology. Especially in the field of artificial intelligence (AI) a wild jungle of problematic terms is in use. However, as long as discussions take place among AI specialists those terms function just

as domain-specific technical vocabulary and no harm is done. But domain-specific language often diffuses into other fields and then easily loses its context, its specificity and its limitations. In this process terms which might have started as pragmatic 'weak' metaphors within the technical field, then develop into proper technical terms eventually starting to be seen as 'proper' metaphors outside their original professional context. In addition to the effect of specific terms, those metaphors can also unfold effects beyond concrete technologies but also fuel or inhibit larger narratives around them or digital technology in general. Hence, powerful metaphors push the myths of unlimited potential of (computer) technology, the superiority of computation over human reasoning (Weizenbaum 1976) or the leading role of the 'digital sublime' in transforming society (Mosco 2004). On the other hand, less colourful and less visionary metaphors keep such myths at bay and narratives grounded. Of course, it would be short-sighted to interpret the choice, development and dissemination of technical metaphors and specifically AI terminologies purely as a somehow chaotic process of misunderstandings and unclear technical usage. Those discourses, as all discourses, are a playing field of interests and power where actors brawl over the 'proper' narration either because they find sincere truth in it (e.g., transhumanist zealots of the singularity) or because it plainly benefits them politically or financially (e.g., companies selling AI), or both. Practically speaking, if relevant decision-makers are convinced that AI can develop a real 'understanding' or properly 'interpret' issues, its regular use for sensitive tasks like deciding about social benefits, guiding education, measuring behavioural compliance or judging court cases problematically looms; and corresponding companies will then eagerly come forward to sell such systems to them.

All the above dynamics motivate this work to scrutinise the AI discourse regarding its language and specifically its metaphors. The paper analyses central notions of the AI debate, highlights their problematic consequences and contributes to the debate by proposing more fitting terminology and hereby enabling more fruitful debates.

Conceptual Domains and Everyday Language

Unlike the abstract field of mathematics, where most technical terms are easily spotted as such, AI makes heavy use of anthropomorphisms. Considering AI-terms such as 'recognition', 'learning', 'acting', 'deciding', 'remembering', 'understanding' or even 'intelligence' itself, problems clearly loom across all possible conversations. Of course, many other sciences also use scientific terms that are derived from everyday language. In this case, these terms then have clearly defined meanings or at least linked discourses reflecting upon them. Examples are the terms 'fear' in psychology, 'impulse' in physics, 'will' in philosophy or 'rejection' in geology and 'ideology' mathematics. Often the same words have completely different meanings in different domains, sometimes

even contradictory meanings, as the examples of 'work' in physics and economic theory (energy transfer via application of a force while moving an object vs. planned and purposeful activity of a person to produce goods or services) or 'transparency' in computer science and political science (invisibility vs. visibility) illustrate.

Hence, problems arise when these scientific terms are transferred carelessly into other domains or back into everyday language used in political or public debates. This can occur through unprofessional science journalism, deliberate inaccuracy for PR purposes, exaggerations for raising third-party funding, or generally due to a lack of sensitivity to the various levels and contexts of metaphors.

The Case of Artificial Intelligence

For some years now, technical solutions utilising artificial intelligence are widely seen as means to tackle many fundamental problems of mankind. From fighting the climate crisis, tackling the problems of ageing societies, reducing global poverty, stopping terror, detecting copyright infringements or curing cancer to improving evidence-based politics, improving predictive police work, local transportation, self-driving cars and even waste removal.

Definitions

The first step towards a meaningful discussion about AI would be to define what exactly one means when talking about AI. Historically there have been two major understandings of AI: strong AI or Artificial General Intelligence (AGI) and weak AI or Artificial Narrow Intelligence (ANI). The goal of AGI is the creation of an artificial human like intelligence, so an AI system with true human-like intelligence including perception, agency, consciousness, intentions and maybe even emotions (see Turing 1950 or more popular Kurzweil 2005). ANI, on the other hand, refers to very domain-specific AI systems being able to accomplish very specific tasks in very narrowly defined contexts only. Questions of agency or consciousness do not arise with ANI systems, they are merely tools, although potentially very powerful tools.

So far and tellingly, AGI can only be found in manifold media products within the fantasy or science fiction genre. Famous examples are *Samantha* in 'HER', *Data* in 'Star Trek', *HAL 9000* in '2001' (based on a novel of the aforementioned writer Arthur C. Clarke), *Bishop* in 'Aliens', the *Terminator* in the movie series of the same name or even the *Maschinenmensch* in 'Metropolis' (Hermann 2020).

In contrast, ANI systems are the ones calculating the moves in advanced chess games, the ones enhancing smartphone pictures, the ones doing pattern

recognition concerning speech (e.g., natural language processing) or images (e.g., computer vision) or even the ones optimising online search results. Furthermore, within the ANI discourse mainly two more specific definitions should be mentioned. The first one focuses on the technical process of how ANI works and goes along the lines of AI being computer algorithms that improve automatically through experience (cf. Mitchell 1997 about machine learning). The second one focuses more on the phenomenon of ANI by defining AI broadly as computer systems that are able to perform tasks normally requiring human intelligence (Gevarter 1985).

Technically there are a multitude of approaches to actually build AI systems. Those approaches are usually referred to as the field of machine learning (ML) and comprise the so-called symbolic approaches with explicit data representations of relevant information like simple decision trees or formal logic-based ones like knowledge databases with inference algorithms. These approaches are comparatively limited due to the necessity of explicit data representation. Then again there are the more recent sub-symbolic approaches of ML which do not use explicit data representations of relevant information but mathematical (e.g., statistical) methods for processing all kinds of data. Artificial neural networks (ANN) or evolutionary computation are examples of sub-symbolic approaches in ML. Interestingly so far, none of the actual methods available seem to promise a path to AGI.

Yet, despite having at least some general definitions at hand, the common discussion usually ignores those and therefore the range of AI-assigned functionality reaches from applying traditional statistics to using machine learning (ML) techniques up to solely movie inspired ideas or even generally to 'highly complex information systems', as in the official 'Social Principles of Human-centric AI' of Japan (Council for Social Principles of Human-centric AI 2019).

In the following, we will concentrate on artificial neural networks to illustrate the fallacies and pitfalls of questionably used language. The focus on ANN in this text is in line with the current debate of AI, where AI is predominantly used synonymously with machine learning using artificial neural networks (Eberl 2018). Nevertheless, the problems mentioned here also apply to debates concerning other forms of AI, when a similar terminology is being used.

Key drivers for the current AI renaissance are the successes of applying artificial neural networks to huge amounts of data now being available and using new powerful hardware. Although the theoretical foundations of the concepts used were conceived as early as the 1980s, the performance of such a system has improved to such an extent over the last years, that they can now be put to practical use in many new use cases, sometimes even in real-time applications such as image or speech recognition. Especially if huge data sets for training are available, depending on the task results can be much better than traditional symbolic approaches where information is written into databases for explicit knowledge representation.

Before we analyse the language being used to describe the functionality, we should have a look at the inner workings of artificial neural networks to have a base for scrutinising terminologies.

Basic Structure of Artificial Neural Networks

Artificial neural networks are an approach of computer science to solve complex problems that are hard to explicitly formulate, or more concretely: to program. Those networks are inspired by the function of the human brain and its network of neurons; however, the model of a neuron being used is very simplistic. Many details of biological neuronal networks, such as myelination or ageing (Hartline 2009), are left out, as well as new mechanisms, such as backpropagation (Crick 1989), are introduced. Trying to follow the original model, each artificial neuron, the smallest unit of such systems, has several inputs and one output. In each artificial neuron, the inputs are weighted according to its configuration and then summed up. If the result exceeds a certain defined threshold the neuron is triggered, and a signal is passed on to the output. These neurons are usually formed into 'layers', where each layer's outputs are the next layer's inputs. The resulting artificial neural network thus has as its input the individual inputs of the first layer, and as its output the individual outputs of the last layer. The layers in between are usually called 'hidden' layers and with many hidden layers an artificial neural network is usually called 'deep'.

In the practical example of image recognition, the input would consist of the colour values of all distinct pixels in a given image and the output would be the probability distribution among the predefined set of objects to recognise.

Configuring the Networks

From a computer science point of view ANNs are very simple algorithms, since the signal paths through the connections of the network can easily be calculated by mathematical equations. After all, it is the variables of this equation (weights, thresholds, etc.) that accord the powerful functionality to ANNs. So ANNs are basically simple programs with a very complex configuration file and there are various ways of configuring artificial neural networks, which will now briefly be described. Building such a network involves certain degrees of freedom and hence decisions, such as the number of artificial neurons, the number of layers, the number and weights of connections between artificial neurons and the specific function determining the trigger behaviour of each artificial neuron. To properly recognise certain patterns in the given data – objects, clusters etc. – all those parameters need to be adjusted to a use case. Usually there are best practices how to initially set it up; then the artificial network has to be further improved step by step. During this process the weights of the connections

will be adjusted slightly in each step, until the desired outcome is created, may it be the satisfactory detection of cats in pictures or the clustering of vast data in a useful way. Those training cycles are often done with a lot of labelled data and then repeated until the weights do not change any more. Now it is a configured artificial neural network for the given task in the given domain.

Speaking about the Networks

Now we will take a closer look at how computer scientists speak about this technology in papers and in public, and how those utterances are carried into journalism and furthermore into politics. As mentioned above, the description of ANNs as being inspired by the human brain already implies an analogy which must be critically reflected upon. Commonly used ANNs are usually comparatively simple, both in terms of how the biochemical properties of neurons are modelled but also in the complexity of the networks themselves. A comparison: the human brain consists of some 100 billion neurons while each is connected to 7,000 other neurons on average. ANNs on the other hand are in the magnitude of hundreds or thousands of neurons while each is connected to tens or hundreds of other neurons. This difference in orders of magnitude entails a huge difference in functionality, let alone understanding them as models of the human brain. Even if to this point the difference might only be a matter of scale and complexity, not principle, we have no indication of that changing anytime soon. Thus, using the notion of 'human cognition' to describe ANN is not only radically oversimplifying, it also opens up the metaphor space to other neighbouring yet misleading concepts. For example, scientists usually do not speak of networks being configured but being 'trained' or doing '(deep) learning'. Along those lines are notions like 'recognition', 'acting', 'discrimination', 'communication', 'memory', 'understanding' and, of course, 'intelligence'.

Considering Human Related Concepts

When we usually speak of 'learning', it is being used as a cognitive and social concept describing humans (or, to be inclusive, intelligent species in general) gaining knowledge as individual learner or as a group, involving other peers, motivations, intentions, teachers or coaches and a cultural background (Bieri 2017). This concept includes the context and a whole range of learning processes being researched, tested and applied in the academic and practical fields of psychology of learning, pedagogy, educational science (Piaget 1944) and neuroscience (Kandel and Hawkins 1995). This is a substantial difference to the manual or automated configuration of an ANN using test sets of data. Seen in this light, the common notion of 'self-learning systems' sounds even more misplaced. This difference in understanding has great implications, since, for

example, an ANN would never get bored with its training data and therefore decide to learn something else or simply refuse to cooperate (Weizenbaum 1976); metaphors matter, no Terminator from the movies in sight.

'Recognition' or 'memory' are also very complex concepts in the human realm. Recognising objects or faces requires attention, focus, context and – depending on one's school of thought – even consciousness or emotions. Human recognition is therefore completely different from automatically finding differences of brightness in pictures to determine the shape and class of an expected object (Goodman 1976). Furthermore, consciously remembering something is a highly complex process for humans which is more comparable to living through imagined events again and by that even changing what is being remembered. Human memory is therefore a very lively and dynamic process, and not at all comparable to retrieving accurate copies of stored data bits (Kandel and Hawkins 1995).

Especially the notions of 'action' or even 'agency' are highly problematic when being applied to computers or robots. The move of a computer-controlled robotic arm in a factory should not be called a robot's 'action', just because it would be an 'action' if the arm belonged to a human being. Concerning human actions, very broad and long-lasting discussions at least in philosophy and the social sciences already exist, note the difference between 'behaviour' and 'action' (Searle 1992). The former only focuses on observable movement, whereas the latter also includes questions of intention, meaning, consciousness, teleology, world modelling, emotions, context, culture and much more (Weizenbaum 1976). While a robot or a robotic arm can be described in terms of behavioural observations, its movements should not easily be called actions (Fischer and Ravizza 1998).

Similarly complex is the notion of 'communication' in a human context, since communication surely differs from simply uttering sounds or writing shapes. 'Communication' requires a communication partner, who knows that the symbols used have been chosen explicitly with the understanding that they will be interpreted as deliberate utterances (von Savigny 1983). Communication therefore needs at least the common acknowledgement of the communicational process by the involved parties, in other words an understanding of each other as communicating (Watzlawick 1964). A 'successful' communication is then the result of both parties agreeing that it was successful and therefore the creation of a common understanding. Hence, the sound of a loudspeaker or the text on a screen does not constitute a process of communication in the human sense, even if their consequences are the production of information within the receiving human being. If there is no reflection of the communication partner, no deliberation, no freedom of which symbols to choose and what to communicate one should not easily apply such complex notions as 'communication' outside its scope without explanation.

Furthermore, the concept of 'autonomy' – as opposed to heteronomy or being externally controlled – is widely used nowadays when dealing with

artificial intelligence, may it be concerning 'intelligent' cars or 'advanced' weapon systems. Although starting in the last century even human autonomy has been largely criticised within the social sciences (some even say completely deconstructed, Krähnke 2006) since individuals are largely influenced by culture, societal norms and the like, the concept of autonomy seems to gain new traction in the context of computer science. Yet, it is a very simplistic understanding of the original concept (Gerhardt 2002). Systems claimed to be 'autonomous' heavily depend on many factors, e.g., a stable, calculable environment, but also on programming, tuning, training, repairing, refuelling and debugging, which are still traditionally done by humans, often with the help of other technical systems. In effect those systems act according to inputs and surroundings, but they do not 'decide' on something (Kreowski 2018), certainly not as humans do (Bieri 2001). Here again, the system can in principle not contemplate its actions and finally reach the conclusion to stop operating or change its programmed objectives autonomously. Hence, artificial intelligence systems – with or without ANN – might be highly complex systems, but they are neither autonomous nor should responsibility or accountability be attributed to them (Fischer and Ravizza 1998). Here we see one concrete instance of the importance of differentiating between the domain-specific ANI and universal AGI (Rispens 2005). This clarification is not meant to diminish the technical work of all engineers involved in such 'autonomous' systems, it is purely a critique about how to adequately contextualise and talk about such systems and its capabilities in non-expert contexts.

Instances and Consequences

After having briefly touched upon some areas of wrongly used concepts, we can take a look at concrete examples, where such language use specifically matters.

A very interesting and at that time widely discussed example was Google's 'Deep Dream' image recognition and classification software from 2015, codename 'Inception'. As described above, ANNs do not contain any kind of explicit models; they implicitly have the 'trained' properties distributed within their structures. Some of those structures can be visualised by inserting random data – called 'noise' – instead of actual pictures. In this noise, the ANN then detects patterns exposing its own inner structure. What is interesting are not the results – predominantly psychedelic imagery – but the terminology being used in Google's descriptions and journalists' reports. The name 'Deep Dream' alone is already significant, but also the descriptive phrases 'Inside an artificial brain' and 'Inceptionism' (Mordvintsev and Tyka 2015). Both phrases (deliberately) give free rein to one's imagination. In additional texts provided by Google, wordings such as 'the network makes decisions' accumulate. Further claims are that it 'searches' for the right qualities in pictures, it 'learns like a child' or it even 'interprets' pictures. Using this misleading vocabulary to describe ANNs

and similar technical artefacts, one can easily start to hope that they will be able to learn something about the fundamentals of human thinking. Presumably those texts and descriptions have been written for the primary purpose of marketing or public relations, since they explain little but signify the abilities and knowledge of the makers, yet that does not diminish the effect of the language used. For many journalists and executive summary writers or even the interested public those texts are the main source of information, not the hopefully neutral scientific papers. In effect, many of those misleading terms where widely used, expanded on and by that spread right into politician's daily briefings, think-tank working papers and dozens of management magazines, where the readers are usually not aware of the initial meanings. This distorted 'knowledge' then becomes the basis for impactful political, societal and managerial decisions.

Other instances where using wrong concepts and wordings mattered greatly are in car crashes involving automated vehicles, e.g., from companies like Uber, Google or Tesla. For example, in 2018 a Tesla vehicle drove into a parked police car in California, because the driver had activated the 'autopilot' feature and did not pay attention to the road. This crash severely exposed the misnomer. The driver could have read the detailed 'autopilot' manual before invoking such a potentially dangerous feature, yet, if this mode of driving had been called 'assisted driving' instead of 'autopilot', very few people would have expected the car to autonomously drive 'by itself'. So, thinking about a car having an autopilot is quite different from thinking about a car having a functionality its makers call 'autopilot'. Actually reading into Tesla's manuals, different levels of driving assistance are being worked on, e.g., 'Enhanced Autopilot' or 'Full Self-Driving', whereas the latter has not been implemented so far. Further dissecting the existing 'autopilot' feature one finds it comprises different sub-functionalities such as Lane Assist, Collision Avoidance Assist, Speed Assist, Auto High Beam, Traffic Aware Cruise Control or Assisted Lane Changes. This collection of assistance technologies sounds very helpful, yet it does not seem to add up to the proclaimed new level of autonomous driving systems with an autopilot being able to 'independently' drive by itself.

Those examples clearly show how a distinct reality is created by talking about technology in certain terms, yet avoiding others. Choosing the right terms, is not always a matter of life and death, but they certainly pre-structure social and societal negotiations regarding the use of technology.

Malicious Metaphors and Transhumanism

Suddenly we arrive in a situation where metaphors are not only better or worse for explaining specifics of technology, but where specific metaphors are deliberately being used to push certain agendas; in Tesla's case to push a commercial and futurist agenda. Commercial because of using 'autonomy' as a unique

selling point for cars and futuristic, as it implies that 'autonomy' is a necessary and objective improvement for everyone's life and the society as a whole. Generally, most innovative products involving 'artificial intelligence' and 'next generation technology' are being communicated as making 'the world a better place', 'humans more empowered' or 'societies more free' by the PR departments of the offering companies and spread even further by willing believers and reporting journalists. The long-standing effects of metaphors let loose can also be seen vividly in the discourse about transhumanism, where humans themselves, even humankind as a whole, should be enhanced and improved using (information) technology, predominantly by using AI. Here again the proponents either really believe in or profit from those narratives, or both (Kurzweil 2005).

In this discourse all mistakes of the AI terminology can be observed fully developed with many consequences, since when we pose the transhumanist question regarding how information technology can help human beings the answer is usually 'enhancement'. Yet the notion of 'enhancement' is being used in a very technical way, ignoring its fundamental multiplicity of meanings. With information technology, so the argument from the classic flavour of transhumanism goes, we will soon be able to fix and update the human operating system: merging with intelligent technical systems will make our brain remember more faces, forget less details, think faster, jump higher, live longer, see more sharply, be awake longer, be stronger, hear more frequencies and even create new senses – exactly how a technologist would imagine what new technology could deliver for humanity (Kurzweil 2005). More recent concepts see humans and AI systems in a cooperative even symbiotic relationship. Those concepts exemplify the direction of imagination once we assume there are truly 'intelligent' systems with 'agency' who can 'decide' and 'act'.

However interestingly the underlying and implicit assumption is a very specific – to be precise: technical – understanding of what is considered 'good' or 'desirable'. But does every human or even the majority primarily want to remember more, forget less, live longer or run faster? Are those aspects even the most pressing issues we want technology to solve? In addition, not only do those fantasies happily follow along the lines of the neo-liberal logic of applying quantification, competition, performance and efficiency into all aspects of life, they also unconsciously mix in masculinist – even militarist – fantasies of power, control, strength and subjugation of the natural or finally correcting the assumed defective (Schmitz and Schinzel 2004).

As valid as those opinions concerning optimisations are, still it is important that views like that imply absolute values and are incompatible with views which put social negotiation, non-mechanistic cultural dynamics or in general pluralistic approaches in their centre. To structure the discourse, I call those conflicting groups of views *regimes of enhancement*. Clearly it is not possible to 'enhance' a human being with technically actualised immortality, if this person does not want to live forever or does not find it particularly relevant. Many

other conflicting views can be thought of. However, the mere acceptance of the concept of *regimes* already breaks any claim for absoluteness and opens the door for discussing different understandings of 'enhancements'. Accepting this already makes any positions somehow compatible and allows for individual or even societal endeavours of creative re-interpretations of the concept of transhumanism itself (Haraway 1991).

The transhumanist discourse outlines the consequences of not reflecting on core notions like 'enhancement' in the same way as it is consequential not to reflect on 'intelligence' in the AI discourse (Bonsiepen 1994). Broken visions and faulty applications are to be expected. Furthermore, this kind of language also shapes and attracts a certain kind of mindset where the above mentioned reductionist metaphors are not even used as metaphors anymore, but as accurate descriptions of the world (Coy 1993).

Constructive Wording

So, next time decision-makers and journalists will be asked about possibilities of technology they will surely remember having heard and read about computers winning Chess and Go, driving cars, recognising speech, translating text, managing traffic and generally finding optimal solutions to given problems (Dreyfus 1972). But using deficient anthropomorphisms like 'self-learning', 'autonomous' or 'intelligent' to describe the technical options of solving problems will lead to malicious decisions (The Royal Society 2018, 7–8).

Surely the best solution for this problem would be to completely change the terminology, but since large parts of the above mentioned are fixed scientific terms, a clean slate approach seems unrealistic. Therefore, at least in interdisciplinary work, science journalism activities or political hearings, a focus should be put on choosing the appropriate wording by scientists and (science) journalists. Only then policy and decision-makers have a chance to meaningfully grasp the consequences of their actions. In addition, interdisciplinary research could also get a more solid (communication) ground. Of course, this change of terms will not be the end of discipline-limited jargon in AI, but it would surely increase the efficiency of exchange between the different fields.

For concretely deciding which terms to use and which words to change it would be generally preferable to have some kind of criteria. Following the above descriptions, the used terminology should be as close as possible to the technical actualities while at the same time avoiding:

- technical terms that have a connotation in common language reaching far beyond the actual technical function, e.g., recognition, agent, communication, language, memory, training, senses, etc. since they will be understood as metaphors

- anthropomorphisms which are not technical terms but usually used as metaphors to describe technical details, e.g., thinking, (re)acting, deciding, remembering, etc.
- concepts widely used in popular science, media and science fiction implying a completely different meaning e.g., intelligent machine, android, self-improving, autopilot, etc.

Certainly those words can be replaced by more fitting vocabulary. Depending on context 'remembering' could be paraphrased by 'implicitly stored in configuration', 'learning' by 'changing/improving configuration', 'recognition' by 'detection', 'intelligent' by 'automated' (cf. Butollo 2018), 'action' by 'movement' or 'response', 'decision' or 'judgement' by 'calculation' (cf. Weizenbaum 1976), and 'communication' by 'indication' or 'signalling'. However, terms like 'agency' and 'autonomy' should be discarded entirely, since they are neither accurate or necessary nor helpful; they are just completely misleading.

Being aware that this change might also bear consequences for scientific grant proposals which usually have to sound societally important, scientifically innovative and relevant, it is imperative here too as part of science ethics to reflect on the wider consequences of the language used to communicate. Admittedly it should be noted that those suggestions won't be applied by speakers who are deeply convinced of such metaphors fitting the subject matter, yet, they would then be clearly visible as such.

Closing Remarks

Technology is used and politically decided upon perceived functionality, not upon the actually implemented functionality. However, communicating functionality is much more driven by interests than creating the actual technology. Therefore, attribution ascription is a very delicate and consequential issue that paints a differentiated picture of the consequences of careless use of terms. If relevant decision-makers in politics and society are (really) convinced at some point that these 'new' artificial neural networks can develop an understanding of things or properly interpret facts, nothing would stand in the way of their use for socially or politically sensitive tasks like deciding about social benefits, teaching children or judging court cases. Here the difference between 'judging' and judging, 'acting' and acting play out. If one acts in the social science meaning of the word, one has to take responsibility for one's actions, if a computer only 'acts', used as a metaphor, responsibility is blurred.

Hence, especially computer professionals but also scientific journalists should follow the professional responsibility to be more sensitive about the criticised misleading metaphors and in effect to change them to more fitting ones. The danger here does not lie in incompletely understanding computers or AI but in not understanding them while thinking that they have been understood. A

possible way out of this tricky situation is certainly more disciplinary openness towards interdisciplinary research and communication. Especially the discipline of computer science could embrace this kind of exchange much more, from student curricula to research projects. This would maybe not so much change their disciplinary core work but it would contextualise this work, create better accessibility for other less technical fields and produce overall more useful results. Naturally, this would require all parties involved to speak to each other but also to listen and teach each other ways of looking at the world. Of course, and not least those ventures must be encouraged and facilitated by field leaders, research grant givers and research politics alike.

So, if technological discussions and societal reflections on the use of technology are to be fruitful, scientists and (science) journalists alike have to stop joining the buzzword-driven language game of commercial actors and AI believers alike, which does neither help with solutions nor advances science. It merely entertains our wishful thinking of how magical technology should shape the future. Finally, we record that a chess computer will never get up and change its profession, exponential growth in computing power does so far not entail more than linear growth of cognitive-like functionality, and the fear of computers eliminating all human jobs is a myth capable of inciting fear since at least 1972.

But maybe, indeed, any sufficiently advanced technology is indistinguishable from magic – to the layperson – but we also have to conclude that this 'magic' is being constructed and used by certain expert 'magicians' to advance their own interests and agendas, or that of their masters (Hermann 2020). So not even such magical interpretation would spare us the necessity to pay attention to power, details and debate (Kitchin 2017). This chapter tries to constructively be a part of this interdisciplinary project.

References

Bieri, P. 2001. *Das Handwerk der Freiheit*. Munich: Hanser.

Bieri, P. 2017. *Wie wäre es, gebildet zu sein?* Munich: Komplett Media GmbH.

Bonsiepen, L. 1994. Folgen des Marginalen. Zur Technikfolgenabschätzung der KI. In: G. Cyranek and W. Coy (Eds.), *Die maschinelle Kunst des Denkens. Theorie der Informatik*. Braunschweig/Wiesbaden: Vieweg.

Butollo, F. 2018. Automatisierungsdividende und gesellschaftliche Teilhabe. *Regierungsforschung.de*, NRW School of Governance. Retrieved from https://regierungsforschung.de/wp-content/uploads/2018/05/23052018_regierungsforschung.de_Butollo_Automatisierungsdividende.pdf

Council for Social Principles of Human-centric AI. 2019. *Social Principles of Human-Centric AI*. Council for Social Principles of Human-centric AI: Japan.

Coy, W. 1992. Für eine Theorie der Informatik! In: W. Coy et al. (Eds.), *Sichtweisen der Informatik. Theorie der Informatik*, pp. 17–32. Wiesbaden: Vieweg +Teubner Verlag.

Coy, W. 1993. Reduziertes Denken. Informatik in der Tradition des formalistischen Forschungsprogramms. *Informatik und Philosophie* 22, 31–52.
Crick, F. 1989. The Recent Excitement About Neural Networks. *Nature* 337 (6203), 129–132.
Dreyfus, H. L. 1972. *What Computers Can't Do*. New York: Harper & Row.
Eberl, U. 2018. Was ist Künstliche Intelligenz – was kann sie leisten? *Aus Politik und Zeitgeschichte,* 6–8 (2018), 8–14.
Fischer, J. M. and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
Gerhardt, V. 2002: *Freiheit als Selbstbestimmung*. In: Wobus A. N. et al, (Eds.) *Nova Acta Leopoldina*, Issue 324, Volume 86, Deutsche Akademie der Naturforscher, Leopoldina, Halle (Saale).
Gevarter, W. B. 1985. *Intelligent Machines: Introductory Perspective on Artificial Intelligence and Robotics*. Englewood Cliffs, NJ: Prentice Hall.
Goodman, N. 1976. *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis, MA: Hackett Publishing.
Haraway, D. 1991. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. *Simians, Cyborgs and Women: The Reinvention of Nature*, pp. 149–181. New York: Routledge.
Hartline, D. K. 2009. *What is Myelin?* Cambridge: Cambridge University Press.
Hermann, I. 2020. Künstliche Intelligenz in der Science-Fiction: Mehr Magie als Technik. *Von Menschen und Maschinen: Interdisziplinäre Perspektiven auf das Verhältnis von Gesellschaft und Technik in Vergangenheit, Gegenwart und Zukunft. Proceedings der 3. Tagung des Nachwuchsnetzwerks 'INSIST', 5–7 October 2018, Karlsruhe (INSIST-Proceedings 3)*.
Kandel, E. R. and Hawkins, R. D. 1995. Neuronal Plasticity and Learning. In: R. D. Broadwell (Ed.), *Neuroscience, Memory, and Language. Decade of the Brain, Vol. 1*, S. 45–58. Library of Congress: Washington, DC.
Kitchin, R. 2017. Thinking Critically About and Researching Algorithms. *Information, Communication & Society* 20 (1), 14–29.
Krähnke, U. 2006. Selbstbestimmung. *Zur gesellschaftlichen Konstruktion einer normativen Leitidee*. Weilerswist: Velbrück Verlag.
Kreowski, H.-J. 2018. Autonomie in Technischen Systemen. *Leibniz Online* 32.
Kurzweil, R. 2005. *The Singularity Is Near: When Humans Transcend Biology*. London: Penguin.
Mitchell, T. 1997. *Machine Learning*. New York: McGraw Hill.
Mordvintsev, A. and Tyka, M. 2015. Inceptionism: Going Deeper into Neural Networks. *Google AI Blog*. 17 June 2015. Retrieved from: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.
Morozov, E. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs.
Mosco, V. 2004. *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge, MA: MIT Press.
Passig, K. and Scholz, A. 2015. *Schlamm und Brei und Bits – Warum es die Digitalisierung nicht gibt*. Stuttgart: Klett-Cotta Verlag.

Piaget, J. 1944. *Die geistige Entwicklung des Kindes*. Zurich: M.S. Metz.
Rispens, S. I. 2005. *Machine Reason: A History of Clocks, Computers and Consciousness*. Doctoral Thesis, University of Groningen.
Royal Society, The. 2018. *Portrayals and Perceptions of AI and Why They Matter*. London. Retrieved from: http://lcfi.ac.uk/media/uploads/files/AI_Narratives_Report.pdf.
von Savigny, E. 1983. Zum Begriff der Sprache – Konvention, Bedeutung, Zeichen, Stuttgart: Reclam.
Schmitz, S. and Schinzel, B. 2004. *Grenzgänge: Genderforschung in Informatik und Naturwissenschaften*. Ulrike Helmer Verlag.
Searle, J. R. 1992. *The Rediscovery of Mind*, Cambridge, MA: MIT Press.
Turing, A. 1950. Computing Machinery and Intelligence. *Mind* LIX (236), 433–460. DOI: https://doi.org/10.1093/mind/LIX.236.433.
Watzlawick, P. 1964. *An Anthology of Human Communication*. Palo Alto, CA: Science and Behavior Books.
Weizenbaum, J. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W. H. Freeman.

CHAPTER 7

AI Ethics Needs Good Data

Angela Daly, S. Kate Devitt and Monique Mann

Introduction

Artificial Intelligence (AI) is increasingly a part of our societies and economies, principally paid for and benefiting private organisations and governments. AI applications are offered free to consumers and end-users in products such as Google Maps in exchange for access to vast amounts of data (Zuboff 2019). While AI has the potential to be used for many socially beneficial purposes, there is concern about dangerous and problematic uses of the technology, which has prompted a global conversation on the normative principles to which AI ought adhere, under the banner of 'AI ethics'. Governments, corporations and NGOs throughout the world have generated their own sets of AI ethics principles. Questions and critiques arise about the content of these ethics principles, whether they are actually implemented, and their (legal) enforceability (Wagner 2018). Broader issues emerge about the power and privilege of the organisations, governments and individuals which are creating and implementing AI and accompanying ethical principles. For example, Google has recently announced an ethics service (Simonite 2020), yet has been mired in ethics controversies from violating privacy law (Finley 2019), working on controversial military projects (Crofts and van Rijswijk 2020) and dissolving its Ethics Board merely a week after its establishment (Statt 2019). The

How to cite this book chapter:
Daly, A., Devitt, S. K. and Mann, M. 2021. AI Ethics Needs Good Data. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 103–121. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.g.

creation of, and compliance with, ethical frameworks can be expensive in time and resources, making it easier for wealthy organisations and nation-states to comply with ethical governance and even profit from it while maintaining fundamentally inequitable, unjust and self-protecting practices.

In this chapter, we argue that we need to focus more on the broader question of power and privilege than merely the aforementioned and depoliticised language of 'AI ethics'. AI ethics principles and frameworks tend to centre around the same values (fairness, accountability, transparency, privacy, etc.) and are insufficient to address the *justice* challenges presented by AI in society. Indeed, Amoore (2020, 81) contends that 'a cloud ethics must be capable of asking questions and making political claims that are not already recognised on the existing terrain of rights to privacy and freedoms of association and assembly'. This can be connected to arguments made by Hoffman (2019, 907) that a focus on a 'narrow set of goods, namely rights, opportunities, and resources' is limiting in that it 'cannot account for justice issues related to the design of social, economic, and physical institutions that structure decision-making power and shape normative standards of identity and behaviour'. Hoffman (2019, 908) also contends that an 'outsized focus on these goods obscures dimensions of justice not easily reconciled with a rubric of rights, opportunities and wealth' and that 'emphasis on distributions of these goods fails to appropriately attend to the legitimating, discursive, or dignitary dimensions of data and information in its social and political context'.

In light of these nascent critiques, we present a politically progressive approach to AI governance based on 'good data' which seeks to empower communities and progress the priorities of marginalised and disenfranchised groups worldwide. Our approach also moves the conversation beyond anthropocentrism by incorporating AI's environmental impact into normative discussions (see Foth et al. 2020). Data is the fuel for AI, providing value and power. AI capabilities are typically designed, funded, developed, deployed and regulated (if indeed at all) by the wealthy progressing the values of profit, power and dominance. AI is constructed in a way that typically reinforces and cements the status quo and existing power relationships. AI will continue to be unethical without political consciousness regarding the actors and scenarios into which it is being conceptualised, designed and implemented and the actors and scenarios that are currently excluded from consideration. Our Good Data approach instead seeks to bring these actors, issues and scenarios clearly into the spotlight and thereby into the normative conversation on AI and digital technology more generally.

Accordingly, the chapter will offer an overview and critique of AI ethics, before presenting a conceptual analysis of Good Data in the context of AI. We advance Good Data as an alternative framing for ethical (in the broad sense) questions involving digital data and conclude with some directions on how Good Data can be implemented in practice vis-a-vis AI. However, for a 'Best Data' scenario for AI to be achieved, greater change contesting and replacing neoliberal capitalism may be necessary (Daly 2016; Zuboff 2019), given the

political economy roots of many contemporary Bad Data practices by governments and large corporations throughout the world, which are also being implemented via AI applications. Thereby any 'quick fixes' offered by AI ethics principles may be illusory and indeed longer term more comprehensive approach/es to 'goodness' in AI, data and society overall are needed.

AI Ethics and Governance

In the last few years, a global debate and discussion has emerged about governing AI and, in particular, whether and to which norms AI should adhere. This debate acknowledges the possibility and actuality of AI being used for normatively problematic purposes, including in physically dangerous and other harmful ways, as well as what ethical approaches humans should take towards potentially autonomous AI that may mimic our own characteristics (see e.g., Bennett and Daly 2020; Donath 2020; Dörfler 2020). A variety of stakeholders, from nation-states throughout the world to regional blocs like the European Union to large technology companies (both US- and China-based) to religious institutions have participated in this debate by issuing their own iterations of 'ethics' principles to which AI ought adhere (Daly et al. 2019). There is now also a corresponding blossoming of academic literature on AI ethics from a number of disciplines from computer science to law to philosophy to engineering examining, collating, comparing and critiquing these ethics statements and proposals (see e.g., Fjeld et al. 2020; Larsson 2020).

There have been two prominent critiques of this 'turn to ethics': one related to the form of these ethics initiatives and one relating to the substance.

One prominent critique of this 'turn to ethics' has been from Wagner (2018), who has expressed concerns about these initiatives constituting 'ethics washing' or 'ethics shopping' – that 'ethics' may 'provide an easy alternative to government regulation', in the context of strong regulation (especially containing fundamental rights protections) having encountered resistance from industry players. The majority of the ethics statements to date, even those from nation-states and other public actors, do not have legally binding force. This raises the question of how sincere and effective such principles may be, particularly in the context of a world in which large technology companies are very powerful, have engaged in problematic conduct in the past and at present, and are not always well regulated by governments (Daly 2016). It is these pre-existing 'infrastructures' and scenarios into which AI is being developed and deployed, yet these aspects are often divorced from the AI ethics discourses (Veale 2020). In addition, we see the involvement of industry players in defining these ethical principles, with the EU High-Level Expert Group a notable and controversial example as the presence of industry lobbying seemingly had an impact on the final document (see e.g., Rességuier and Rodrigues 2020). Law is one way of enforcing ethical principles but not the only one, and Hagendorff (2020) points

to broader issues with AI ethics 'usually lack[ing] mechanisms to reinforce its own normative claims', mechanisms that include law but also cover technical implementations and business process operationalisations of 'ethics'.

It is also important to note the 'weaponisation' of 'ethics' in the AI debates, to refer to these ethics initiatives issued mainly by governments and corporations and the 'ethics-washing' critiques which can descend into 'ethics bashing', distracting from the more general and broad meaning of ethics as it relates to morality and virtue (Bietti 2020). In other words, 'ethics' has been used in a specific way in the AI debates both to promote (usually non-binding) lists of norms and as a target for critique that these norms are insufficient, formulated by the wrong actors and not backed up with enforcement or implementation. However, 'ethics' in the more general term can make an important contribution to considerations of morality in a broad sense when it comes to AI (Bietti 2020): the ethics baby should not be thrown out with the 'AI ethics initiatives' bathwater.

The critique of ethics initiatives' substance relates to what is included and excluded as normative principles. Hagendorff (2020) identifies recurring norms in ethics declarations he analyses and compares, notably 'accountability, privacy or fairness'. Along with 'robustness or safety' Hagendorff (2020) considers these frequently occurring norms as those that 'are most easily operationalised mathematically and thus tend to be implemented in terms of technical solutions'. Hagendorff (2020) also points to the 'omissions' from many AI ethics frameworks comprising 'red lines' on uses of AI which should be prohibited, political abuse of AI systems and the '"hidden" social and ecological costs' of AI systems. Furthermore, ethical AI principles such as fairness, accountability, transparency judge only the information systems within which AI is installed, without stepping back to analyse the socio-economic and political realities of the organisations which own and use the data – as per Bigo et al.'s (2019) *data politics* – and AI, and the people who willingly or ignorantly provide the fuel to power AI.

The critique of AI ethics' substance also extends to scenarios where AI ethics are backed by legal enforceability. While a latecomer to state-led AI ethics, the United States has made up for lost time since 2019 starting with the Trump Administration Executive Order on Maintaining American Leadership in Artificial Intelligence (US White House 2019). What is notable about this Executive Order is its legal enforceability albeit in the form of a direction to other US government agencies rather than, for example, a piece of general legislation containing a set of legally binding normative principles (Daly et al. 2020). The Executive Order does contain five high level principles, including the US driving the development of appropriate technical standards and protecting civil liberties and privacy (US White House 2019). However, the US's policy among its own government agencies since then has promoted a deregulatory approach to AI whereby agencies should reduce regulatory barriers to AI development and adoption (Daly et al. 2020).

This approach demonstrates the limits of legal enforceability in AI ethics contexts whereby, in the US case, legal enforceability is used to mandate a deregulatory approach (Daly et al. 2020). Accordingly, 'the legal enforceability of AI governance and ethics strategies does not necessarily equate to substantively better outcomes as regards actual AI governance and regulation' (Daly et al. 2020). Instead, we must be alert to legal enforceability as a form of 'law washing', or when the binding force of law does not in itself prevent unethical uses of AI (Daly et al. 2020).

This leads to the need to interrogate and evaluate both the form and substance of AI ethics – what they do and do not contain in substance, and what binding force they may or may not have, whether they are only 'performative' and 'instrumentalist' or whether the language of ethics is only performative. In other words, AI ethics need a Good Data approach.

A Good Data Approach

Ethics as currently utilised in the AI debates is a limited frame through which AI issues can be viewed. While we acknowledge that ethics has a broader and more general sense than its use in AI ethics so far (Bietti 2020), we do not seek to reclaim it as a linguistic device given the term's history and tarnishment in these debates. Instead, we propose 'Good Data' (Daly, Devitt and Mann 2019), as a more expansive concept to elucidate the values, rights and interests at stake when it comes to AI's development and deployment as well as that of other digital technologies. In particular, we argue that discourses, design and deployment on and of AI must engage with power and political economy, perspectives which are largely lacking in AI ethics initiatives to date (see Johnson 2019).

We conceived the notion of 'Good Data' to move beyond critique of the digital (in which we have participated and continue to do so) to the (re)imagining and articulating of a more optimistic vision of the datafied future and, in particular, how digital technologies and data, including but not limited to AI, can be used to further wider social, economic, cultural and political goals (Daly, Devitt and Mann 2019). We draw on work on data justice (Dencik, Hintz and Cable 2016), data activism (Milan and van der Velden 2016) and data politics (Bigo, Isen and Ruppert (2019) as key elements or examples of Good Data, while our concept involves 'broader visions of goodness or ethics or politically progressive data' (Daly, Devitt and Mann, 2019).

AI ethics frameworks to date focus on evaluating the design and impacts of AI without sufficient attention on the socio-political contexts in which AI is developed and employed (as per Amoore 2020). What this means is that AI can be responsible, governable, trusted, equitable, traceable to the persons to whom it applies and reliable within the complex systems it is deployed – i.e., 'ethical' by the measure of many of the AI ethics initiatives – but the governing organisation/s responsible for the AI itself may be unethical in the broader

society and environment within which it is deployed. A Good Data approach interrogates these broader situations and factors which are often absent from AI ethics initiatives.

It is reasonable to consider the scope of 'goodness' and what philosophical commitments we might have to the assertion of 'good' data. For our purposes, goodness can be a property of a thing, a service, a method, an event, a system, a process, a judgement, a sensation, a feeling or a combination of these. To identify 'good', we could suppose that there are moral facts (see Parfit 2011; Scanlon 2014). Agreement on moral facts enables standards, policies, practices and frameworks to improve information systems and communicate expectations. However, given the limits of our knowledge of moral facts (should they exist) and in light of colonial and post-colonial data practices (Arora 2016) we assume a hybrid moral theory – where we allow that some moral facts may be objective (e.g. 'tolerance' or 'openness') and others relative (e.g. Wong 1984). A hybrid theory allows respect for cultural diversity and demands case-by-case determinations of goodness and systematic values and standards. By promoting a hybrid account, we are prepared for disagreement about what is good and assume that the discovery of moral facts (if they do exist) is non-trivial and unresolved. We advocate an ethic of active seeking, openness and tolerance to diverse views on 'the good' particularly, and perhaps stridently, consultation with the underrepresented, marginalised and unheard.

Pillars of Good Data

Good Data contribute to understanding and justifying progressive political action by collectives. Good Data is thus situated in an ethical perspective to progress society, rather than simply satisfying an epistemic goal to inform. Therefore, we connect Good Data with political action and social justice – it means doing something *good* in the world, or equally not doing something *bad,* i.e. forbearance. Good Data also can take place and be relevant to all stages in the data collection process, from the beginning to the end encompassing: when the decision is made for the data to be collected for AI use and by whom; at the point the data is collected by AI; at the point the data is processed/analysed by AI; at the point the data is used by AI; and at the point the data is reused by AI. In order to conceptualise the process and outcome of Good Data, we advance these 'pillars' on which it should rest, rather than principles to which it should apply.

We present four pillars: Community, Rights, Useability, Politics that emerge from the corpus of *Good Data* (Daly, Devitt and Mann 2019) which can guide digital technologies and data development, including for AI. We propose that these pillars can guide an ethical and politically progressive approach to AI development, governance and implementation.

Community

Good Data must be orchestrated and mediated by and for individuals and collectives. Individuals and collectives should have access to, control over and opportunities to use their own data to promote sustainable, communal living, including communal sharing for community decision-making and self-governance and self-determination (see Lovett et al. 2019; Ho and Chuang 2019). Data collection, analysis and use must be orchestrated and mediated by and for data subjects and communities advancing their data and technological sovereignty, rather than determined by those in power (Kuch et al. 2019; Mann et al. 2020). AI, constructed by communities, should be designed to assist community participation in data-related decision-making and governance. This community element is usually absent from AI ethics initiatives, both in their formation – principally by elites – and their content. Commitment to theories and practices of sovereignty (Kurtulus 2004) are critical to systems' architecture design including data permissions, accessibility and privacy.

Rights

While we recognise that the discourse of rights is limiting (as per Amoore 2020; Hoffman 2019), Good Data should still be collected with respect to humans and their rights, and that of the natural world, including animals and plants (Trenham and Steer 2019). The rise of big data and AI makes individual control over all their shared digital personal or community data a possibly insurmountable task. Rights language and power stems from a protection of the individual (especially in western worldviews), and there may be conflicts between the community's values and priorities with community data and the preferences of an individual within those communities. While Kalulé and Joque (2019) criticise the contemporary anthropocentrism of AI and privileging of the western human, we believe that a rights discourse is not completely futile and in principle AI can be developed to improve abidance with human rights and the rights of the environment. However, such a language of rights and especially the rights of the non-human are usually absent from AI ethics initiatives. With Good Data we urge that the environmental cost and impact of AI technologies is an 'externality' which must be 'internalised' in discussions of ethics and politically progressive AI and digital data.

Usability

Good Data is usable and fit for purpose, consensual, fair and transparent (Trenham and Steer 2019). Measures of fairness and other values attributed to data should be explicit (McNamara et al. 2019), and extend beyond

narrowly conceived technical explanations to challenge broader structural/societal unfairness (see Hoffmann 2019 on the limits of 'fairness'). Data driven technologies must respect interpersonal relationships such as appropriate (Flintham et al. 2019), e.g. members of the same household may wish for limits on accessing each other's data – in other words, data is relational. Good Data is dependent on context, and with reasonable exceptions, should be open and published, revisable and form useful social capital where appropriate to do so (Trenham and Steer 2019). AI ethics frameworks on data dovetail with the requirements of usability, though they do fall short on the nuances of respecting interpersonal relationships and community values. There is substantial overlap between usability and community; and usability and dependence. Which is to say, that data must be usable and dependable for the community who must access and control it. If ICT systems leave communities dependent on the expertise of 'outsiders' to maintain them, then they create vulnerabilities for their sovereignty. Usability for communities is vital for all kinds of communities, from families, hospitals, schools, cultural groups, businesses and organisations as well as for the nation(-states).

Politics

Good Data reveals and challenges the existing political and economic order so that data empowered citizens can secure a good polity. Citizen-led data initiatives lead to empowered citizens (see e.g. Valencia and Restrepo 2019). Open data enables citizen activism and empowerment (see Gray and Lämmerhirt 2019). Strong information security, online anonymity and encryption tools are integral to a good polity. Social activism must proceed with 'good enough data' (Gutierrez 2019; Gabrys, Pritchard and Barratt 2016) to promote the use of data by citizens to impose political pressure for social ends. How can AI contribute to the empowerment of citizens, without data, models and algorithms putting them at risk? AI systems need to be understood by citizens so that outputs and recommendations are trusted as working in their favour. To this end, the politics pillar on activism for Good Data and good AI is drawn from the other three: community, rights and usability. AI infrastructure controlled and accessed by communities that progresses their rights and interests is the gold standard of genuinely ethical AI.

Our research into Good Data encourages data optimism beyond minimal ethical checklists and duties – thus our aim is supererogatory (Heyd 1982). We recommend these Good Data pillars to progress political and social justice agendas such as citizen-led data initiatives, accepting 'good enough' data to achieve aims (Gutierrez 2019; Gabrys, Pritchard and Barratt 2016). The aim is to dismantle existing power structures through the empowerment of communities and citizens via data and digital technologies and enhancing technological sovereignty (Mann et al. 2020).

Moving away from the body of critique of pervasive 'bad data' practices by both governments and private actors in the globalised digital economy, we paint an alternative, more optimistic but still pragmatic picture of the datafied future. In order to secure a just and equitable digital economy and society we need to consider community, rights, usability and politics.

AI for Good Data? Good Data for AI?

But how can we implement these pillars in practice? Can AI be fed with or nourished by Good Data? Can AI feed and nourish Good Data? We acknowledge that here too we are not the first people to consider this issue. For instance, Floridi et al. (2020) discuss the emerging 'AI for Social Good' (AI4SG) trend and formulate seven 'essential' but not 'sufficient' sociotechnical principles of their own for AI4SG. These principles are rather technocratic although Floridi et al. (2020) do acknowledge the wider contexts in which AI development and deployment take place and the power imbalances which persist forming the backdrop to these developments and deployments. We instead seek to centre these contexts and power imbalances in proposing Good Data as a frame or concept for AI development.

For AI to be Good Data and Good Data to be AI, AI would need to be built for communities using data available and relevant to them. To achieve good data, communities need to gather, store and process their own data; they need to have access to open and closed data sets of relevance to their interests. Communities need cloud storage, AI classifiers, data scientists and so forth to build the tools communities need to become empowered. It is unclear what socioeconomic structures would enable genuinely ethical AI with Good Data – but certainly not current ones. At the very least it would require massive investment, democratisation and reimagining of ICT infrastructure. The most powerful produce and selectively hide data. The least powerful depend on data gathered, curated and displayed by the empowered, often data about them.

What, then, are the barriers to achieving Good Data? We have mentioned a few in passing above: the smokescreen of ethics to obscure enforceable state-led regulation; the limits of law; the broader political economy of neoliberal capitalism and corporate greed and its impacts of extractive logic on the natural world; and the corresponding power imbalances and inequalities in a world characterised by privilege and division. In addition, creating effective and trusted AI is elusive and expensive and requires access to valuable data sets, knowledge workers as well as access to high quality digital architecture, test and evaluation processes and user testing in the anticipated context of use. Data and software must be incorporated into secure and compliant back-end databases with user-friendly front-end interfaces. While AI is becoming much more 'plug and play', enabling those with less skills to add data to software products and curate algorithms, the end-to-end construction of AI products to

meet a societal need is still a bespoke and expensive business. The complexity of AI in addition to its cost is why AI production is dominated by three kinds of enterprise: (a) technical startups, (b) medium to large sized corporations and (c) governments. There is comparatively very little AI constructed by low-resource community groups, non-profits, aid agencies, advocacy groups for the marginalised and disenfranchised.

What, then, can be done to address these barriers? A multipronged approach is necessary to (start to) dismantle (some of) them, with a recognition of the limits of law, code, markets and social norms (as per Lessig's (1999) modes of regulation) as tools through which Good Data can be achieved. Central though to the problem is the political economy of data in the context of neo-liberal capitalism. Both governments and corporations have a strong, in some cases existential, interest in gathering, analysing and using data. Truly curbing these practices through law or ethics or markets may be extremely difficult to achieve in practice absent major societal change.

However, not to give into defeatism regarding these large challenges, we view the way forward as beginning to create an alternative vision of a datafied society and economy which promotes and achieves social and environmental justice goals, and we view incremental change for now to be the most likely pragmatic path in this direction.

There are some cases of AI for social good, for example, a software engineer developed an AI that could automatically write letters for people who received parking tickets in a way that got them a waiver from having to pay the fine (Dale 2016; DoNotPay 2020). The business aims to connect people to legal advice from parking tickets to divorce (Krause 2017). The people most likely to benefit from such an AI included those in the community who lack the funds to pay the parking ticket and may lack the education, literacy, knowledge or experience required to negotiate written legal documents. AI products for legal aid use AI to improve equity and fairness. There are cases of AI for environmental good, such as poaching deterrence and identification of rare and endangered species. AI can automatically count flocks, track animals, assess perimeter, monitor habitats.[1] Although the use of AI for surveillance – even for ostensibly 'good' reasons – remains controversial. Consequentialist justifications will never satisfy rights-based or duty-based obligations to other humans such as protection from persistent surveillance.

However, AI for social good is ad hoc. That is to say, private individuals generate the concept of AI to alleviate some source of injustice and proceed to develop a technology that may be useful for a specific purpose, but does not have the backing of a significant entity to ensure that AI products are chosen to improve quality of life for the most vulnerable through a process of consultation and oversight.

Non-profits and other organisations dedicated to the alleviation of human suffering and improving justice are traditionally staffed by less technical persons, such as those with legal training rather than technical training in software

engineering and machine learning. It is difficult to build up the technical competence required to create AI for social justice within organisations already struggling to deliver their organisation's missions within tight budgets.

Governments might be good candidates to make AI for the good of all citizens. However, time and time again governments are found to use citizen data for uses that do not align with the values and expectations of marginalised groups within society, such as First Nations peoples (e.g., see Kukutai and Taylor 2016; Lovett et al. 2019), the unemployed or marginally employed (e.g., for an overview of Australia's RoboDebt welfare surveillance program see Mann 2019; 2020).

The quality of AI outputs is based on the data that it is fed and curated with. Organisations lacking access to large data sets will be unable to participate in the AI economy. Conversely, large corporations that focus on data collection as a primary asset collect vast data sets to feed AI algorithms.

Moving Towards 'Better' Data in AI

We view the way forward as beginning to create an alternative vision of a datafied society and economy which promotes and achieves social and environmental justice goals, and we view incremental change for now to be the most likely pragmatic path in this direction.

As a way forward to ensuring Good Data we look to integrate Lessig's (1999) various approaches or modes of regulating technology, namely: law, code/architecture, social norms and markets with philosophical models of information, acknowledging epistemic, ethical and political conceptions of what constitutes 'the good'. In order to ensure technology, such as AI, rests on Good Data pillars and exhibits Good Data values, a multipronged approach involving these different modalities of regulation and conceptual apparatus is necessary.

It is insufficient to rely only on formal law to achieve ethical and politically progressive outcomes – as also recognised by Kalulé and Joque (2019), Kalulé (2019), Hoffman (2019) and Amoore (2020). We do not necessarily accept the determinist view that the law follows technological innovation which the 'regulatory disconnect' suggests (Brownsword and Goodwin 2012). Indeed, we acknowledge that digital technologies have benefitted from emerging in a period when deregulatory, neoliberal ideologies have prevailed which has led to the law playing 'catch-up' (Daly 2016).

However, we do see some potentially promising provisions such as Article 25 of the EU's General Data Protection Regulation on 'data protection by design and default', as an attempt to 'join up' law and code in implementing ethical principles. But, we and others have questions about how this translates – if indeed it is possible to do so – into the design or hardcoding of systems (Koops and Leenes 2014), and what the consequences, including unintended of such an intention to embed principles into technological systems may be

(Ihde 2006). As we also see in the US, the mere fact of AI ethics principles having the binding force of law is insufficient to establish their 'goodness'.

It is also insufficient to focus solely on social norms in the form of unenforceable ethical principles as these can indeed result in ethics washing. We agree with Powles and Nissenbaum (2018) who see a focus on 'solving' issues of fairness, accountability and transparency as foremost among these discussions through code (and to some extent social norms) as problematic, and one which obscures the broader structural problems which are at the root of bias in machine learning. Again, Ihde's critique of the 'designer fallacy' mentioned above also applies here regarding the unintended consequences of attempting to 'design out' bias and other problems in digital technologies. Furthermore, broader existential questions about whether a particular system or technology should even be used in the first place are of key importance but also often obscured in the focus on issues such as bias (Powles and Nissenbaum 2018). A Good Data approach to AI would certainly ask these questions before any such system was implemented and see that these problems are pertaining to broader social, political and economic contexts which will not easily be 'solved' by technology alone.

While it may not eventuate in a Good Data utopia, we view that laws, social norms, code and markets ought to promote and attempt to ensure Good Data practices. While at least one underlying problem of Good Data may be the capitalist political economy (Daly 2016; Benthall 2018), some incremental steps to promote positive change can still be taken (Raicu 2018) – or 'better' data.

Better data for AI can be promoted through a number of ways. While each alone is insufficient, together they may equate to progress. The multiplicity of ways and methods to achieve better data for AI reflect the embeddedness of AI in pre-existing, currently existing and future socio-environmental-economic conditions, and as not something that can easily be 'solved' via a statement of ethics principles.

Pragmatically, to achieve better data even the market can assist through corporate social responsibility initiatives by private sector players through ensuring they act in ethical ways (beyond legal obligations) in their product development, manufacturing, implementation and sales (Grigore, Molesworth and Watkins 2017). Better data for AI can also be advanced by environmentally sustainable corporate social responsibility initiatives (see e.g. Chuang and Huang 2018) and circular economy initiatives, and by corporations ensuring adherence to high labour standards at all stages in the supply chain. The Fairphone is an example of an attempt to produce such an ethical piece of digital technology in the private sector (Akema, Whiteman and Kennedy 2016). Workers in technology companies can also do what they can to resist Bad Data practices, as we have witnessed at Google and Amazon in recent years (Montiero 2017; Salinas and D'Onfro 2018). Data ethical norms and practices need to be inculcated at all levels of society including formal and informal

educational settings; internet and social networking standards, media and communication channels and in the attainment of professional accreditation and qualifications. While law alone is insufficient, it also should not be dispensed with as a tool for moving towards better data for AI.

Moreover, the debate on AI ethics has been dominated by western approaches to this topic. We also look to the Indigenous Data Sovereignty movements developed and led by First Nations peoples as presenting radically different visions of data collection and usage from the hegemonic western norm, and bring to the fore key questions of whether data should be collected and by whom (Kukutai and Taylor 2016; Lovett et al. 2019). Good Data approaches must take account of Indigenous perspectives and worldviews on data and the discrimination and oppression that Indigenous peoples and nations have hitherto experienced through western colonialism and imperialism. We are already seeing promising developments. New Zealand has recently released a draft algorithmic charter that explicitly seeks to 'embed a Te Ao Māori perspective in algorithm development or procurement'.[2]

Conclusion

We have argued here that AI needs Good Data. The four pillars of Good Data: community, rights, usability and politics are at the forefront of a just digital society and economy. Good Data situates genuinely ethical AI within communities and collectives, rather than individuals or large organisations. The well-being of the people and environment must be at the forefront of AI ethical considerations, including considerations not to use AI at all.

We have also argued that AI needs Good Data because the issues that are at the forefront of the digital society and economy go beyond pre-existing discussions of ethics. Like data itself, it is impossible for us to cover everything encompassed by 'Good Data' and accordingly we cannot offer a 'complete', 'comprehensive' or 'perfect' account of Good Data at this stage (if indeed ever). But we can say that Good Data is a more expansive concept which aims to encompass practices beyond 'ethics' and also human rights, environmental and social justice concerns arising around data which may involve extending beyond the focus to date on 'AI ethics' and an emerging focus on 'AI law' to address deficiencies with 'AI ethics'.

Good Data should permeate digital technology development, implementation and use at all stages in the process, and involve different tools, notably law, norms, code and markets, in order to bring about 'better' – or 'good enough' scenarios, even if the broader societal conditions and limitations mean that it is difficult to bring about 'Best Data'. Good Data can also involve the forbearance from generating and using data, either at all, or in some circumstances or by some specific people. This has implications for businesses as

data collection, analysis and use should be orchestrated and mediated by, with and for data subjects, rather than determined by those in power (corporate or otherwise).

We also hope that Good Data can encompass a more global approach, rather than just (re)centring perspectives from the Global North, as already noted – and critiqued – by Arora (2016) and Kalulé and Joque (2019). However, we acknowledge we are also coming from a northern/western perspective ourselves.[3] Already there is emerging discussion from China, in particular, on technology ethics, and legislative activity in many jurisdictions around the world regarding data localisation (Melashchenko 2019). The Indigenous Data Sovereignty movements also display different worldviews and approaches to issues of data situated in Indigenous laws, cultures and traditions, countervailing the practices and uses of data by colonial and imperial forces against Indigenous peoples, and representing more perspectives beyond the western focus on normativity and ethics as regards technologies including AI.

Pragmatically, we view the next steps for all involved in the digital society and economy (which, in fact, is all of us) as trying to engage and empower each other to build Good Data initiatives and communities of change, rather than letting governments and corporations build a Bad Data future for us. Yet it is also important that governments and corporations contribute positively to the Good Data future by taking note and implementing 'good' and more ethical data practices. Only with such a multifaceted approach encompassing will we be able to achieve some semblance of Good Data for AI and for the digital more generally.

Notes

[1] From https://www.dronezon.com/drones-for-good/wildlife-conservation-protection-using-anti-poaching-drones-technology

[2] The Draft NZ Algorithmic Charter is available here: https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

[3] Or perhaps more accurately for two of us, a 'Global North-in-South' perspective – see Mann and Daly (2019).

References

Akema, O., Whiteman, G. and Kennedy, S. 2016. Social Enterprise Emergence from Social Movement Activism: The Fairphone Case. *Journal of Management Studies*, 53(5), 846–877.

Amoore, L. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press.

Arora, P. 2016. The Bottom of the Data Pyramid: Big Data and the Global South. *International Journal of Communication*, 10, 1681–1699.

Bennett, B. and Daly, A. 2020. Recognising Rights for Robots: Can We? Will We? Should We? *Law, Innovation and Technology*, 12(1), 60–80. DOI: https://doi.org/10.1080/17579961.2020.1727063

Benthall, S. 2018. The Politics of AI Ethics is a Seductive Diversion from Fixing our Broken Capitalist System. *Digifesto*.

Bietti, E. 2020. From Ethics Washing to Ethics Bashing A View on Tech Ethics from Within Moral Philosophy. Proceedings of ACM FAT* Conference (FAT* 2020). ACM, New York. DOI: https://doi.org/10.1145/3351095.337286

Bigo, D., Isin, E. and Ruppert, E. 2019. *Data Politics: Worlds, Subjects, Rights*. Abingdon: Taylor & Francis.

Brownsword, R. and Goodwin, M. 2012. *Law and Technologies of the Twenty-First Century*. Cambridge: Cambridge University Press.

Chuang, S.P and Huang, S.J. 2018. The Effect of Environmental Corporate Social Responsibility on Environmental Performance and Business Competitiveness: The Mediation of Green Information Technology Capital. *Journal of Business Ethics*, 50(4), 991-1009.

Crofts, P. and van Rijswijk, H. 2020. Negotiating 'Evil': Google, Project Maven and the Corporate Form. *Law, Technology & Humans*, 2(1), 75–90.

Dale, R. 2016. The Return of the Chatbots. *Natural Language Engineering*, 22(5), 811–817.

Daly, A. 2016. *Private Power, Online Information Flows and EU Law: Mind the Gap*. Oxford: Hart.

Daly, A., Devitt, S.K. and Mann, M. (Eds.). 2019. *Good Data*. Amsterdam: Institute of Network Cultures.

Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B., Wang, W.W. and Witteborn, S. 2019. *Artificial Intelligence, Governance and Ethics: Global Perspectives*. The Chinese University of Hong Kong Faculty of Law Research Paper No. 2019-15. Retrieved from: https://ssrn.com/abstract=3414805. DOI: https://doi.org/10.2139/ssrn.3414805

Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B. and Wang, W.W. 2020. AI, Governance and Ethics: Global Perspectives. In O. Pollicino and G. de Gregorio (Eds.), *Constitutional Challenges in the Algorithmic Society*, forthcoming. Cambridge: Cambridge University Press.

Dencik, L., Hintz, A. and Cable, J. 2016. Towards Data Justice? The Ambiguity of Anti-Surveillance Resistance in Political Activism. *Big Data & Society*, 3(2), 1–12.

Donath, J. 2020. Ethical Issues in Our Relationship with Artificial Entities. In M.D. Dubber, F. Pasquale and S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.

DoNotPay. 2020. Do Not Pay Community. Retrieved from https://donotpay.com/learn

Dörfler, V. 2020. Artificial Intelligence. In M.A. Runco and S.R. Pritzker (Eds.), *Encyclopedia of Creativity* (3rd ed., vol. 1), pp. 57–64. Oxford: Academic Press.

Finley, K. 2019. EU Privacy Law Snares Its First Tech Giant: Google. *Wired*. Retreived from https://www.wired.com/story/eu-privacy-law-snares-first-tech-giant-google

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berkman Klein Center Research Publication No. 2020–1*. Retrieved from: https://ssrn.com/abstract=3518482 DOI: https://doi.org/10.2139/ssrn.3518482

Flintham, M., Goulden, M., Price, D. and Urquhart, L. 2019. Domesticating Data: Socio-Legal Perspectives on Smart Homes and Good Data Design. In A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 344-360. Amsterdam: Institute of Network Cultures.

Floridi, L., Cowls, J., King, T.C. and Taddeo, M. 2020. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, 26, 1771–1796. DOI: https://doi.org/10.1007/s11948-020-00213-5

Foth, M., Mann, M., Bedford, L., Fieuw, W. and Walters, R. 2020. *A Capitalocentric Review of Technology for Sustainable Development: The Case for More-Than-Human Design*. Melville, South Africa: Global Information Society Watch (GISWatch), Association for Progressive Communications (APC).

Gabrys, J., Pritchard, H. and Barratt, B. 2016. Just Good Enough Data: Figuring Data Citizenships through Air Pollution Sensing and Data Stories. *Big Data & Society*. DOI: https://doi.org/10.1177/2053951716679677.

Gray, J. and Lämmerhirt, D. 2019. Making Data Public. The Open Data Index as Participatory Device. In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 174–188. Amsterdam: Institute of Network Cultures.

Grigore, G., Molesworth, M. and Watkins, R. 2017. New Corporate Responsibilities in the Digital Economy. In A. Theofilou, G. Grigore and A. Stancu (Eds.), *Corporate Social Responsibility in the Post-Financial Crisis Era: CSR Conceptualisations and International Practices in Times of Uncertainty*, pp. 41–62. London: Palgrave Macmillan.

Gutierrez, M. 2019. The Good, the Bad and the Beauty of 'Good Enough Data'. In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 54–76. Amsterdam: Institute of Network Cultures.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*, 30, 99–120. DOI: https://doi.org/10.1007/s11023-020-09517-8

Heyd, D. 1982. *Supererogation: Its Status in Ethical Theory*. Cambridge: Cambridge University Press.

Ho, C.H. and Chuang, T.R. 2019. Governance of Communal Data Sharing. In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 202–215. Amsterdam: Institute of Network Cultures.

Hoffman, L. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Anti-Discrimination Discourse. *Information, Communication & Society*, 22(7), 900–915.

Ihde, D. 2006. The Designer Fallacy and Technological Imagination. In: J. Dakers (Ed.), *Defining Technological Literacy: Towards an Epistemological Framework*, pp. 121–131. London: Palgrave Macmillan.

Johnson, K. 2019. AI Ethics is All About Power. *Venture Beat*, 1 November. Retrieved from: https://venturebeat.com/2019/11/11/ai-ethics-is-all-about-power

Kalulé, P. 2019. On the Undecidability of Legal and Technological Regulation. *Law Critique*, 30, 137–158. DOI: https://doi.org/10.1007/s10978-019-09240-z

Kalulé, P. and Joque, J. 2019. Law & Critique: Technology Elsewhere, (yet) Phantasmically Present. *Critical Legal Thinking*, 16 August. Retrieved from: https://criticallegalthinking.com/2019/08/16/law-critique-technology-elsewhere-yet-phantasmically-present

Koops, B. and Leenes, R. 2014. Privacy Regulation Cannot be Hardcoded. A Critical Comment on the 'Privacy by Design' Provisions in Data-Protection Law. *International Review of Law, Computers and Technology*, 28(2), 159–171.

Krause, E. 2017. This Robot Will Handle Your Divorce Free of Charge. *The Wall Street Journal*, 26 October. https://www.wsj.com/articles/this-robot-will-handle-your-divorce-free-of-charge-1522341583

Kuch, D., Stringer, N., Marshall, L., Young, S., Roberts, M., MacGill, I., Bruce, A. and Passey, R. 2019. An Energy Data Manifesto. In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 77–95. Amsterdam: Institute of Network Cultures.

Kukutai, T. and Taylor, J. (Eds.). 2016. *Indigenous Data Sovereignty: Towards an Agenda*. Canberra: ANU Press.

Kurtulus, E. 2004. Theories of Sovereignty: An Interdisciplinary Approach. *Global Society*, 18(4), 347–371.

Larsson, S. 2020. On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society*, 7(3), 437-451. DOI: 10.1017/als.2020.19

Lessig, L. 1999. *Code: And Other Laws of Cyberspace*. New York: Basic Books.

Lovett, R., Lee, V., Kukutai, T., Cormack, D., Carroll Rainie, S. and Walker, J. 2019. Good Data Practices for Indigenous Data Sovereignty and Governance. In A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 26–36. Amsterdam: Institute of Network Cultures.

Mann, M. 2019. *Social (In)Security and Social (In)Justice: Automation in the Australian Welfare System*. Report on Artificial Intelligence: Human Rights, Social Justice and Development. Global Information Society Watch (GISWatch). Association for Progressive Communications (APC), Melville, South Africa.

Mann, M. 2020. Technological Politics of Automated Welfare Surveillance. *Global Perspectives*, 1(1), 1–12.

Mann, M. and Daly, A. 2019. (Big) Data and the North-in-South: Australia's Informational Imperialism and Digital Colonialism. *Television & New Media*, 20(4), 379–395. DOI: https://doi.org/10.1177/1527476418806091

Mann, M., Mitchell, P., Foth, M. and Anastasiu, I. 2020. #BlockSidewalk to Barcelona: Technological Sovereignty and the Social Licence to Operate Smart Cities. *Journal of the Association for Information Science and Technology*, 71(9), 1103–1115.

McNamara, D., Graham, T., Broad, E. and Ong, C.S. 2019. Trade-offs in Algorithmic Risk Assessment: An Australian Domestic Violence Case Study. In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 96–116. Amsterdam: Institute of Network Cultures.

Melashchenko, N. 2019. Data Localization: Policymaking or Gambling? In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 156–173. Amsterdam: Institute of Network Cultures.

Milan, S. and van der Velden, L. 2016. The Alternative Epistemologies of Data Activism. *Digital Culture & Society*, 2(2), 57–74.

Montiero, M. 2017. Ethics Can't be a Side Hustle. *Dear Design Student*, 19 March. Retrieved from: https://deardesignstudent.com/ethics-cant-be-a-side-hustle-b9e78c090aee

Parfit, D. 2011. *On What Matters* (Vol. 2). Oxford: Oxford University Press.

Powles, J. and Nissenbaum, H. 2018. The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. *One Zero*, 7 December. Retrieved from: https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

Raicu, I. 2018. False Dilemmas: Technology Ethics, Law, and Fairness in AI. *Internet Ethics Blog*, 14 December. Retrieved from: https://www.scu.edu/ethics/internet-ethics-blog/false-dilemmas/

Rességuier, A. and Rodrigues, R. 2020. AI Ethics Should not Remain Toothless! A Call to Bring Back the Teeth of Ethics. *Big Data & Society 7*(2). DOI: https://doi.org/10.1177/2053951720942541

Salinas, S. and D'Onfro, J. 2018. Google Employees: We No Longer Believe the Company Places Values Over Profit. *CNBC*, 27 November. Retrieved from: https://www.cnbc.com/2018/11/27/read-google-employees-open-letter-protesting-project-dragonfly.html

Scanlon, T.M. 2014. *Being Realistic about Reasons*. Oxford: Oxford University Press.

Simonite, T. 2020. Google Offers to Help Others with the Tricky Ethics of AI. *Ars Technica*, 29 August. https://arstechnica.com/tech-policy/2020/08/google-offers-to-help-others-with-the-tricky-ethics-of-ai

Statt, N. 2019. Google Dissolves AI Ethics Board Just One Week After Forming It. *The Verge*, 4 April. https://www.theverge.com/2019/4/4/18296113/google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation

Trenham, C. and Steer, A. 2019. The Good Data Manifesto. In: A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 37–53. Amsterdam: Institute of Network Cultures.

US White House. 2019. Executive Order on Maintaining American Leadership in Artificial Intelligence. 11 February. Retrieved from: https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence

Valencia, J.C and Restrepo, P. 2019. Truly Smart Cities: Buen Conocer, Digital Activism and Urban Agroecology in Colombia. In A. Daly, S.K. Devitt and M. Mann (Eds.), *Good Data*, pp. 317-329. Amsterdam: Institute of Network Cultures.

Veale, M. 2020. A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence. *European Journal of Risk Regulation*, 1–10. DOI: https://doi.org/10.1017/err.2019.65

Wagner, B. 2018. Ethics as an Escape from Regulation: From Ethics-washing to Ethics-shopping. In E. Bayramoglu, I. Baraliuc, L. Janssens et al. (Eds.), *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, pp. 84–98. Amsterdam: Amsterdam University Press.

Wong, D.B. 1984. *Moral Relativity*. Berkeley, CA: University of California Press.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

CHAPTER 8

The Social Reconfiguration of Artificial Intelligence: Utility and Feasibility

James Steinhoff

Introduction

This chapter addresses the notion of 'AI for everyone' via the concept of social reconfiguration. This was originally formulated by Jasper Bernes (2013) in his critique of what he calls the 'reconfiguration thesis' or the assumption, held by many Marxists and other critics of capital, that 'all existing means of production must have some use beyond capital, and that all technological innovation must have … a progressive dimension which is recuperable'. In other words, existing technologies which have been produced by capital for the advancement of capitalist industry can and should be appropriated and redirected towards non-capitalist, democratically-determined and socially-beneficial ends – the means of production can and should be seized.

The reconfiguration of AI is a timely topic because, since 2015, almost all the big USA tech companies, such as Google and Microsoft, have announced commitments to the so-called 'democratisation' of AI. Critiques of such programs have already been provided (Garvey 2018; Sudmann 2020; Dyer-Witheford, Kjøsen and Steinhoff 2019, 52–56) and Marxists have long pointed out that capitalism is defined by technology not being democratically controlled, but rather designed and deployed to serve the interests of one small sub-group of the world population, the owners of capital (Braverman 1998).

How to cite this book chapter:
Steinhoff, J. 2021. The Social Reconfiguration of Artificial Intelligence: Utility and Feasibility. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 123–143. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.h.

In this chapter, I schematise the concept of reconfiguration with two dimensions: utility and feasibility. I argue that existing considerations of the reconfiguration of AI have primarily focused on utility and have largely neglected questions of feasibility. Feasibility is considered primarily in relation to the materiality of AI, or its concrete aspects which 'set constraints on and offer affordances for use' (Leonardi and Barley 2008, 171). By attending to the materiality of AI we can see how it differs from traditional, industrial means of production.

The chapter first discusses the contemporary form of AI called machine learning and its increasing importance to the tech industry. Then I discuss several aspects of its materiality. Next, I discuss Marxist theories of technology and existing evaluations of reconfiguring AI, which focus primarily on utility. Then I turn to the question of feasibility. I conclude that the social reconfiguration of AI faces substantial difficulties posed by the lack of visibility and non-modularity of AI, but that some promise is to be derived from the data commons movement. I suggest that further research on socially reconfiguring technology should focus more on feasibility, rather than utility and can begin by looking at concrete ways to resist the impositions of AI capital.

Machine Learning Materiality

Industry

Early approaches to AI attempted to automate high-level logical reasoning represented in formal languages. Such approaches to AI are called 'symbolic' or 'good old-fashioned' AI (Haugeland 1989) and have largely been overshadowed by a different approach to AI known as machine learning. Machine learning is often anthropomorphised, but it is at base the use of statistical methods, called learning algorithms, to find patterns in large datasets. On the basis of these patterns an algorithm called a 'model' is produced which may be used to analyse new data (Alpaydin 2014, 2–3). A model thus represents 'knowledge' of the patterns found by the learning algorithm and can be used to make useful analyses or predictions. Much of the hype around machine learning derives from this automated production of models from data, which Domingos (2015) calls the 'inverse of programming' (6–7).

Machine learning is being applied almost anywhere electronic data is accessible. Brynjolfsson and McAfee (2017) argue that machine learning is a general-purpose technology comparable to the combustion engine. While this remains to be seen, AI has found diverse applications from recommendation engines and targeted advertising to predictive policing software, predictive maintenance, customer resource management and fraud detection. Capital became visibly interested in machine learning around 2015. All the biggest tech companies in the world have since shifted to AI-intensive directions, including

Google, Amazon, Microsoft, Facebook, IBM, Baidu, Alibaba and Tencent. Older industrial capitals like Siemens and General Electric have followed suit. In addition to these huge companies are a variety of middle sized companies and an array of startups. Investment in AI startups increased from $1.3 billion in 2010 to over $40.4 billion in 2018 (Perrault et al. 2019, 6). This money trickles down to some, but not all, workers involved in producing AI. Salaries for machine learning scientists and engineers average $100,000 to $150,000 USD, with lavish benefits (Stanford 2019) while essential data-preparing 'ghost workers' are precariously employed through platforms like Amazon Mechanical Turk and are minimally remunerated (Gray and Suri 2019; Li 2017).

Data

Perhaps the most fundamental aspect of the materiality of machine learning is that it requires a lot of data from which to extract patterns (Alpaydin 2014, 1–4). One can get an idea of the requisite quantities by looking at some popular datasets. The dataset MNIST, a collection of handwritten digits, contains 70,000 images. Compare it to ImageNet, comprising 14,197,122 images labelled with categories and subcategories. The category 'person' has 952,000 images and 2,035 subcategories (ImageNet 2010). ImageNet is dwarfed by the Gmail Smart Reply training set which contains 238,000,000 examples, and the Google Books Ngram set which amounts to 468,000,000,000 examples. Google Translate is said to employ a dataset numbering somewhere in the trillions (Google 2019). Machine learning is no more than the sophisticated recognition of patterns across such large datasets (for a sober walkthrough of this process see Broussard (2018, 87–120)).

The companies that produce machine learning commodities are unsurprisingly concerned with obtaining vast quantities of diverse data. It is no coincidence that the major producers of AI operate a variety of platform business models in which, by acting as intermediaries between users, they can appropriate all kinds of data (Srnicek 2017). However, quality, as well as quantity, of data is important. Data does not come ready-to-use and requires labour intensive formatting, cleaning and labelling (Gitelman 2013; Gray and Suri 2019).

Compute

Producing machine learning systems requires powerful computing hardware. Since few companies can afford to buy such hardware, most advanced machine learning models are trained and deployed through the cloud platforms of the tech giants. Amazon Web Services dominates the market, but Google, Microsoft, IBM and Baidu all have their own cloud platforms. Computational power required for both training and deploying machine learning models is continually increasing. The amount of computing power used in the largest AI

training runs increased 300,000 times from 2012 and 2018, with no end in sight (Amodei and Hernandez 2018).

Such computation is energy intensive. Cloud platforms thus rely on access to energy infrastructures. According to Pearce (2018), the largest data centres consume as much power as a city of a million people, and in total, data centres consume 'more than 2 percent of the world's electricity and emit roughly as much CO_2 as the airline industry'. Google, Microsoft and Baidu derive 15%, 31% and 67% of their energy, respectively, from coal (Cook 2017, 8). Efforts to 'green' the cloud by increasing renewable energy sources are ongoing, but many such campaigns consist of offsetting or buying carbon credits and do not actually mean that clouds are contributing less to CO_2 production.

Distribution

Machine learning requires sources of data, such as social media platforms. Neither can it function without storage for data, the substantial work which goes into preparing data, nor the cloud or energy infrastructures. Contemporary AI is thus not a discrete technological artifact. Even a relatively simple AI product, such as a smart home speaker, draws on a 'vast planetary network' of labour, data, resources and technologies (Crawford and Joler 2018). Machine learning cannot be analytically separated from the globally distributed infrastructure, both technical and human, on which it relies. And it is perhaps on its way to itself becoming another layer of infrastructure. Science and technology studies scholars have demonstrated how as infrastructures mature, they become 'ubiquitous, accessible, reliable, and transparent' (Edwards et al. 2007, i). Although machine learning as yet possesses none of these qualities perfectly, it is integrated into many people's daily lives in ways that increasingly approach them. Advocates of the AI industry are already positioning AI as a utility comparable to electricity or the internet, a kind of 'cognition on tap' (Steinhoff 2019). This effort to cast AI as something immediately available everywhere for users is complemented by a technological effort, dubbed 'democratisation', to make the production of AI available to a wider range of users.

'Democratisation'

The tech giants have, since around 2015, extolled the 'democratization' of AI. According to Microsoft CTO Kevin Scott, this means 'making sure that everyone has access to those platforms so that they can use the techniques of AI to enhance their own creativity, to build their own businesses, to do their jobs' (Agarwal 2019). For Madhusudan Shekar, Head of Digital Innovation at Amazon Internet Services, the democratisation of AI 'is about making the tooling and capabilities of AI/ML available to developers and data scientists at

various levels of competence so that anybody can use AI to increase the velocity of new customer acquisition, reduce costs and look for new business innovations' (quoted in Ananthraj 2019). In general, democratisation efforts take the form of more or less automated, cloud-based tools and libraries (some of which are free or open source) which either assist developers with building machine learning models or help unskilled users incorporate premade models into other media. A much-feted early episode of democratisation occurred in 2015 when Google open-sourced its now widely-used TensorFlow library.

Garvey (2018) correctly points out that claims of a 'democratization' of AI draw on 'explicitly political' language but do not specify whether or how such programs are 'informed by political conceptions of democratic governance' (8079). The idea appears to be that simply distributing the tools constitutes the democratisation of AI. But this neglects consideration of how the AI products which people encounter in their daily lives are not produced or deployed through processes of democratic deliberation. Nor do such formulations address the larger issue of how the capitalist mode of production itself is premised on the exclusion of certain stakeholders from social decision-making processes. Marxists have discussed this via the distinction between the capitalist (who own and control the means of production) and working classes (who own only their ability to labour). A real democratisation of AI would require that not only capital controls its development and deployment.

Capital, Labour and Machines

Marx held that the relation between the capitalist and working classes was necessarily antagonistic. Capital has one primary goal: to increase. Marx (1990) called this valorisation (293). Mechanisms for valorisation vary across the historical permutations of capitalism, but all rely on the exploitation of labour and the capture of surplus-value. While capitalists and functionaries of capital may argue that the valorisation of capital is co-extensive with social flourishing, the COVID-19 pandemic of 2020 laid the antagonism bare, with CEOs like Jeff Bezos and political functionaries like former US President Trump openly willing to sacrifice workers for the generation of surplus-value. It appears increasingly obvious that, as Land (2017) has argued, capital has 'no conceivable meaning beside self-amplification'.

While the interests of labour in the context of work typically take the form of better wages and working conditions, the broader interests of labour are the interests of socially-existing humans as such and are therefore not amenable to *a priori* description. Marx (1993) describes the ultimate interests of labour as the 'absolute working-out' of 'creative potentialities' (488). Class antagonism results from labour's broad horizon of self-development encountering the narrow logic of capital. Labour might flourish in any number of ways not conducive to valorisation, so capital 'systematically selects for human ends compatible

with its end … and systematically represses all human ends that are not' (Smith 2009, 123). One of its most effective means for doing so are machines. Driven by competition and class struggle, capital introduces machines to increase productivity by cheapening labour power, increasing control over, and dispensing with, labourers.

But this is not to say that Marx considered technology inherently opposed to labour. On the contrary, Marx held that human flourishing could only be achieved with the help of machines. For Marx and Engels (1969), communism only becomes possible when 'machinery and other inventions [make] it possible to hold out the prospect of an all-sided development, a happy existence, for all members of society'. However, before communism can be attempted, machinery must be wrested from capital. Workers must seize the means of production, or 'overthrow … the capitalists and the bureaucrats [and] proceed immediately … to replace them in the control over production and distribution, in the work of keeping account of labor and products' (Lenin 1918). Machines thus are neither inherently wedded to capital nor labour, but are rather a medium for their antagonistic relation.

Since the valorisation of capital can and often does run orthogonal to the interests and wellbeing of labour and since most AI research and production today is conducted by capital, one can assume that it is largely conducted in accord with the exigencies of valorisation. In other words, AI predominantly takes a commodity form (i.e., is designed as something which can be sold for profit) or a form which can otherwise augment the valorisation of capital (i.e., harvesting user data for inputs). There is no reason to assume that AI as a means for capital valorisation stands to benefit society beyond capital. Therefore, consideration of 'AI for everyone' needs to consider how control over AI might be taken away from capital and transferred to a democratic public. If AI is to be directed towards democratically determined ends, it will first have to be seized, in the sense that Marxists have talked of seizing the means of production.

Reconfiguration and Artificial Intelligence

Bernes (2013) defines the 'reconfiguration thesis' as the assumption that 'all existing means of production must have some use beyond capital, and that all technological innovation must have … a progressive dimension which is recuperable'. Bernes first raised the notion of reconfiguration in an analysis of capital's logistics networks. In the course of his argument, Bernes interweaves the increasingly logistical nature of capitalism, critical theory, and how it can arise from workers who inhabit logistical sites of struggle. In stark contrast to his wide-ranging discourse, I will focus narrowly on the notion of reconfiguration. We can schematise reconfiguration with two dimensions: utility and feasibility.

Utility

A first step to thinking about the potential utility of a reconfigured technology is to consider how it is useful now, and to whom. Bernes (2013) argues that logistics is 'capital's own project of cognitive mapping' because it allows capital to keep track of its dispersed moving parts. It enables a new emphasis on circulation characterised by practices such as outsourcing, just-in-time production and the global arbitrage of commodities, including labour-power. It allows the segmentation and stratification of labour, and the brutal creation of 'sacrifice zones' free of labour regulations (Hedges and Sacco 2014). The utility of logistics for capital is thus 'exploitation in its rawest form' (Bernes 2013). This is not likely a use-value for a socially reconfigured AI.

Andrejevic (2020) argues that under capital, what he calls 'automated media' (including AI) tend towards the automation of subjectivity itself (129). Andrejevic argues that this is ultimately impossible on psychoanalytic grounds, but the argument that the ultimate end of capitalist AI is the emulation of subjectivity has been advanced by others. Land (2014) holds that capital and artificial intelligence possess a 'teleological identity' and that a perfected capitalism will dispense with human labour for a full-machine economy. Such speculations range afield from this paper, but they reinforce the more immediate utility of AI for capital. AI is an automation technology with diverse applications for reducing and/or eliminating labour costs and implementing new forms of control over labour processes and social relations. It was these use-values for capital that the earliest Marxist analyses of AI reacted to. In the 1980s, AI was first commercialised in the form of 'expert systems' intended to capture and automate the knowledge and reasoning of skilled workers (Feigenbaum, McCorduck and Nii 1989). Most Marxists of this era were not interested in reconfiguring AI. The near consensus was that AI heralded a new wave of deskilling and concomitant automation, aimed at cognitive, as well as manual, forms of labour (Cooley 1981; Athanasiou 1985; Ramtin 1991).

Planning

However, another strand of Marxist thought saw utility in reconfigured technologies of automation like AI and cybernetics. Both the USSR (Peters 2016) and socialist Chile (Medina 2011) attempted to apply cybernetics to solve the 'socialist calculation problem', as the economist Ludwig von Mises described it. Von Mises (1935) contended that the distribution of resources in a planned economy requires an infeasible amount of calculation and that a capitalist market economy achieves this automatically through the market and price system. While the attempts at planned economies by Chile and the USSR failed due to the primitive computers available at the time, some Marxists have continued to pursue the idea of automated economic planning.

Cockshott (1988) argued that heuristic processing techniques 'developed in artificial intelligence can be applied to solve planning problems with economically acceptable computational costs' (1). More recently he has described big data and supercomputers as the 'foundations of Cyber Communism' (Cockshott 2017). Others have pointed out that algorithmic technologies for processing vast quantities of economic data have already been developed by large corporations like Walmart and Amazon (Jameson 2009; Phillips and Rozworski 2019). Beyond the processing of economic data, Dyer-Witheford (2013) has suggested that AI could be used to lessen bureaucratic burdens: democratic processes might be 'partially delegated to a series of communist software agents … running at the pace of high-speed trading algorithms, scuttling through data rich networks, making recommendations to human participants … communicating and cooperating with each other at a variety of levels' (13).

Bernes (2013) argues that such positions assume that 'high-volume and hyper-global distribution' possess 'usefulness … beyond production for profit'. For instance, a society not structured around commodity production would not be driven to implement planned obsolescence, so one can imagine that the overall volume of things that need to be shipped across the world would decrease substantially. In addition, more localised systems of production might obviate much of the need for vast planning techniques. The broader point is that the utility of a given existing technology for socially-determined, non-capitalist ends is not a given if it was built by capitalist firms to advance valorisation. Utility therefore 'needs to be argued for, not assumed as a matter of course' (Bernes 2013).

Full Automation

Some Marxists have also speculated on the use of AI to eliminate work. This line of thought derives from Marx's notion that 'the true realm of freedom' has its 'basic prerequisite' the 'reduction of the working-day' (Marx 1991, 959). Thinkers in the USSR held that automation had a 'crucial role in the creation of the material and technical basis of communist society' (Cooper 1977,152). Since the mid-2010s, a group of Marx-influenced thinkers referred to variously as left accelerationism (Srnicek and Williams 2015), postcapitalism theory (Mason 2016) and fully automated luxury communism (Bastani 2019) have renewed support for such ideals. I refer only to the left accelerationists here, but all of these thinkers are united in calling for full automation.

Left accelerationists argue that under capital, 'the productive forces of technology' are constrained and directed 'towards needlessly narrow ends' (Williams and Srnicek 2014, 355). The technology developed by capital should be seized: 'existing infrastructure is not a capitalist stage to be smashed, but a springboard to launch towards post-capitalism' (Williams and Srnicek 2014, 355). They hold that 'existing technology [can be] repurposed immediately' (Srnicek and

Williams 2015). Alongside decarbonising the economy, developing renewable energy sources, cheap medicine and space travel, they advocate 'building artificial intelligence' (Srnicek and Williams 2015). For left accelerationists, a reconfigured AI is useful primarily in that it could contribute to full automation, which is desirable because 'machines can increasingly produce all necessary goods and services, while also releasing humanity from the effort of producing them' (Srnicek and Williams 2015). Eventually, a 'fully automated economy' could 'liberate humanity from the drudgery of work while simultaneously producing increasing amounts of wealth' (Srnicek and Williams 2015, 109).

The automation of bad work and the administration of a planned economy are certainly useful applications of AI that extend beyond the logic of valorisation. But utility should be considered alongside feasibility.

Feasibility

Even if a given capitalist technology presents useful possibilities, it is not necessarily the case that its social reconfiguration appears feasible. Bernes presents several reasons why a social reconfiguration of logistics is infeasible, two of which derive from its distributed nature, and are likewise applicable to contemporary machine learning. The first of these pertains to visibility.

Visibility

Logistics comprises a vast, heterogeneous network of technologies and institutions which remains invisible as a whole to the workers who populate its variegated zones. The means of logistical production are distributed across this network, but '[o]ne cannot imagine seizing that which one cannot visualise, and inside of which one's place remains uncertain' (Bernes 2013). Logistics is capital's means for knowing itself, but this knowledge is barred from workers. This sense of visibility is not only an issue when considering the initial seizure of a technology, but also for tracking the progress of its social reconfiguration, which is unlikely to occur instantly. To persevere, 'struggles need to recognise themselves in the effects they create, they need to be able to map out those effects … within a political sequence that has both past and future, that opens onto a horizon of possibilities' (Bernes 2013). Contemporary AI presents similar problems of visibility, from several different angles.

AI is temporally and physically distributed across layers of infrastructures. To 'see' AI we need ways to chart this vast network and make it appear as a coherent collection of people and things. Excellent work on visualising AI has been done in visual essays by Crawford and Joler (2018), which reveals the diverse materiality of AI, and by Pasquinelli and Joler (2020) which aims to 'secularize' AI by casting it not as alien intelligence, but something more like an

optical instrument, akin to a microscope. But visual essays can only go so far. In a more fundamental sense, visibility is a problem of knowledge.

'Democratisation' of AI programs aim to make AI accessible to less skilled users, but they do so by abstracting from the underlying code with user-friendly interfaces. Of course, all computing technology today uses layers of abstraction, whether to allow skilled users to achieve complex ends more easily or to allow less-skilled users to do something at all (including the word processor I am using to write this chapter). Not many people write machine code. But as Kittler (1995) pointed out, increasing layers of abstraction from the underlying materiality of the computer mean that the potential ends it might be put to are reduced; layers of software act as a 'secrecy system' blocking access to basic functionality. So-called 'democratised' machine learning does not enable the production of novel applications of the technology beyond pre-determined bounds. At best, it allows more users to apply pre-canned software tools.

Further, while the open sourcing of AI tools and libraries like Google's TensorFlow may seem like a truly democratic move insofar as companies are giving away proprietary software, it also has competitive dimensions motivated by valorisation. Open sourcing can generate a community around the software which entails skilled developers (and potential future employees) for the company who produces the software. It can also create a software ecosystem based on those tools, which a company can retain control over through a variety of mechanisms from mandatory lock-in agreements to closed source variants of programs. Google used (and uses) such strategies to make Android the most popular mobile operating system in the world (Amadeo 2018). While Google's TensorFlow can currently be run on competing clouds, there are indications that the tech giants are aiming towards fully siloed AI ecosystems. Google is not alone in developing proprietary hardware specially designed for AI. Google's Tensor Processing Unit (TPU) provides a 'performance boost' over traditional hardware, but 'only if you use the right kind of machine-learning framework with it … Google's own TensorFlow' (Yegulalp 2017). Open source AI software is thus one tactic of a larger strategy by which AI capitals combat their rivals for a share of surplus-value.

Visibility is also a technical problem. Machine learning has a 'black box' problem because the complex computations that occur within a system cannot be disassembled and examined and thus its outputs remain inexplicable. As one researcher puts it, the: 'problem is that the knowledge gets baked into the network, rather than into us' (quoted in Castelvecchi 2016). Even if some machine learning models could be reconfigured without being rebuilt, their operations would remain inscrutable, presenting problems of accountability (Garigliano and Mich 2019). The delegation of economic planning or bureaucratic decision-making to a black box might be tolerable for some, as long as no mistakes are made, but it seems dubious that such occult mechanisms would represent a substantial improvement for democratic decision-making over delegating social decisions to the so-called logic of the market.

Non-modularity

A second dimension of feasibility also concerns distribution, but from a tactile, rather than visual, standpoint. Bernes (2013) argues that while revolutions are necessarily localised, 'any attempt to seize the means of [logistical] production would require an immediately global seizure'. Without connection to the rest of the logistical network, a reconfigured port facility is of little use. On the other hand, maintaining connection with the rest of the network entails 'trade with capitalist partners, an enchainment to production for profit … the results of which will be nothing less than disastrous' (Bernes 2013). One might reply that taking the whole system over at once is not necessary – one can appropriate it piecemeal. This might be the case, but it needs to be taken into account that infrastructures are built on top of infrastructures and intertwined with them in 'recursive' ways (Larkin 2013, 30). A technology that is part of a larger system may not necessarily be possible to reconfigure by itself.

In a second consideration of the problem of reconfiguration, focused this time on agriculture and energy, Bernes discusses the non-modularity of certain technologies. By this he means technologies that 'fit together into technical ensembles that exhibit a strong degree of path-dependency, meaning historical implementation strongly influences future development, precluding or making difficult many configurations we may find desirable' (Bernes 2018, 334). He singles out energy infrastructure as particularly non-modular and argues that hopes of simply substituting clean energy sources, even if all political opposition were removed, is wishful thinking because the 'technology [we] would inherit works with and only with fossil fuels' (Bernes 2018, 334).

To consider the non-modularity of machine learning, recall its reliance on the highly centralised clouds maintained by the tech giants. Any reconfiguration of AI would require a seizure of the data centres which make up the cloud as well as the energy sources and infrastructures necessary to power them. Such facilities could, certainly, be seized like more traditional means of production, such as factories. But this presents its own host of material problems. One concerns the powerful hardware required for AI and its energy consumption. While some greening of data centres is evidently possible, it is uncertain whether greening efforts can keep pace with the increasing computational demands of machine learning. Developers at OpenAI recently stated that 'it's difficult to be confident that the recent trend of rapid increase in compute usage will stop, and we see many reasons that the trend could continue' (Sastry et al. 2019).

Cutting-edge machine learning is increasingly out of reach for organisations without resources on par with Facebook or Google. OpenAI was founded as a non-profit research lab with substantial donations from the likes of Elon Musk and Peter Thiel, but in 2019, justified a switch to a 'limited profit' model, in partnership with Microsoft, because AI research 'requires a lot of capital for computational power' (Brockman 2019). If contemporary machine learning algorithms are indefinitely scalable, meaning that their performance improves

as long as more data and computational power are made available, then the hardware cost of AI research and development will continue to rise. If the reconfiguration of AI is to occur in a local context, and if it wishes to remain on functional par with capitalist AI, it will have to devote considerable resources to the requisite hardware and figure out how to make their operation more ecologically feasible.

But perhaps seizing data centres is not necessary for the reconfiguration of AI. Some commentators hope to shift the computational load from the centralised cloud onto individual devices in a technique called decentralised or edge computing. While increasing amounts of edge computing seem likely as components continue to decrease in size, data centres will always offer more space and thus more total computing power. The expert consensus seems to be that with existing technology it is 'not possible to move Cloud-levels of compute onto the edge' (Bailey 2019). Another alternative to seizing the existing cloud could be to construct an alternative cloud. Such initiatives exist, such as the CommonsCloud Alliance, which aims to build a cloud based not on centralised data centres, but on computing power and storage space shared amongst users (Sylvester-Bradley 2018). This seems feasible, but unlikely to compare to the capacities of the clouds of the tech giants.

Data itself also raises several questions of feasibility. The first pertains to data collection. Many AI systems are trained on publicly available datasets in early stages of development, but usually, proprietary datasets are necessary to complete a project (Polovets 2015). The preparation and labelling of these is a labour-intensive and time-consuming process (Wu 2018). Creating a dataset also requires a venue for data collection in the first place. Companies such as Amazon and Google harvest reams of data from the interactions of users of their applications, even when they claim not to be, as smart home devices have shown (Fingas 2019). One business analysis of IBM suggests that because the company lacks a data collection venue, it will face difficulties developing its AI endeavours (Kisner, Wishnow and Ivannikov 2017, 19–20). This perhaps indicates why, in 2020, IBM entered into partnership with data-rich enterprise software company Salesforce. AI entails a capitalism built around surveillance, enabling 'data extractivism' (Zuboff 2019). How desirable is pervasive, multimodal surveillance for a socially reconfigured AI?

Machine learning's reliance on data also necessitates a unique form of maintenance. A model which functioned well when it was deployed will no longer do so if the domain it is applied to changes such that the data it was trained on no longer accurately reflects that domain (Schmitz 2017). Imagine a hypothetical model trained to recognise traffic signs. If overnight the red octagons reading STOP were replaced with purple triangles reading HALT, the model would no longer function and would require maintenance. A social reconfiguration of AI will presumably be one component of a larger democratic restructuring of society with substantial changes to the normal routines of social life. A preview of this sort of disruption for AI has been provided by the COVID-19

pandemic, 'models trained on normal human behavior are now finding that normal has changed, and some are no longer working as they should' (Heaven 2020). When substantial shifts in human behaviour occur, models no longer map onto reality. It is reasonable to assume that models trained on data produced by life under capital may not function in a society striving to fundamentally change its basic axioms.

There is, however, at least one reason for optimism concerning data. A promising alternative to mass surveillance and siloed data ecosystems comes from the notion of data commons, in which individuals and institutions share data willingly, with controls over anonymity and a goal to make data valuable not only to tech companies, but also to its producers. The DECODE projects in Barcelona and Amsterdam have piloted aspects of a data commons successfully and are planning to scale up in the future (Bass and Old 2020). An interesting aspect of these projects is their use of other relatively new technologies, such as smart contracts (Alharby and Van Moorsel 2017), to aggregate and analyse sensitive data in ways which preserve privacy and retain user control. These projects provide a concrete demonstration of the feasibility of reconfiguring some aspects of data ecosystems. That they draw on novel smart contracts should remind us that assessments of feasibility are necessarily contextual; they are constrained by the knowledge of the assessor and the current technological milieu. As such, this chapter makes an argument which remains open to revision. Any social reconfiguration of AI will have to go beyond the assessment attempted here and search out such novelties, technological or other, as might be relevant.

Conclusion: Counter-AI

Discussing the democratization of AI, Kevin Scott, CTO at Microsoft, makes the following comparison with the industrial revolution:

> the people who benefited from [steam powered] technology were folks who had the capital to … build factories and businesses around the machines and people who had expertise to design, build and operate them. But eventually … the technology democratized. You don't get any sort of advantage now, as a capital owner, because you can build an engine. And what we … need to do … is dramatically contract that period of time where AI is so hard to do that only a handful of people can do it. (Agarwal 2019)

Scott's sense of democratisation here hinges on the mere generalisation of a technology. The notion seems to be that because, over time, knowledge of how to build steam engines diffused through the population, this technology became democratised – any 'capital owner' can build or go out and buy a

steam engine. But this formulation seems blissfully unaware of the inequalities between capital owners and labour and thus it precisely misunderstands the meaning of democratisation. There are a lot of people in the world without any capital at all. Further, the mere distribution of free AI tools does not ensure democratic control over the centralised means of AI production nor upset the advantage held by the current producers of AI. This chapter has thus explored what it might mean to actually democratise AI, or rather, to socially reconfigure existing AI into an 'AI for everyone'. The central point I have hoped to make is that consideration of the utility of a socially-reconfigured AI should be complemented by consideration of feasibility, which is largely determined by the 'material character of the powers and forces' involved in the technology (Bernes 2018, 336).

Reconfiguring AI entails simultaneous reconfiguration of large chunks of the tech sector, energy infrastructure, advertising industry, data market/ecosystem, and also requires social deliberation over aspects of the material character of AI, such as its apparent need for surveillance. This assessment resonates with that of Huber (2020), who lucidly argues that a reconfiguration of the capitalist food industry is impossible via incremental piecemeal tweaking, but will instead require revolutionising the entire system. Morozov (2019) makes the same case for the 'feedback infrastructure' or the means of producing, harvesting and processing data which are so essential to AI. On the other hand, the data commons movement indicates practical ways in which the data which machine learning systems are built from can be utlised to benefit those not at the helms of big tech capitals. While the data commons movement is occurring within the circuits of capital, it shows how a social reconfiguration of AI might begin. Even if the seizure of AI seems herculean, data commons projects demonstrate a concrete modicum of feasibility.

Finally, if one cannot seize AI today, one can still resist it. It is true, however, that resistance is a wearily overused term for critics of capitalism. What does it mean, in practice, to resist capitalist machine learning? For a final time, I will draw on Bernes (2013), who suggests that we might imagine a 'logistics against logistics, a counter-logistics which employs the conceptual and technical equipment of the industry in order to identify and exploit bottlenecks … This counter-logistics might be a proletarian art of war to match capital's own *ars belli*'. While this chapter cannot adequately explore the idea, it can suggest that there could be a proletarian counter-AI built around the axis of data on which machine learning, and the capital it increasingly powers, runs.

Early forms of this might take the form of rendering data unavailable or unusable to capital. Users might engage in 'data strikes' by deleting or otherwise denying access to their data (Vincent, Hecht and Sen 2019) and they might distort or 'poison' their data by introducing inaccurate or harmful patterns into it (Vincent et al. 2021). But what about the data infrastructure more broadly? It should be possible to determine key bottlenecks in the valorisation processes

of capitalist AI, at which democratic control might be one day exercised, but for now might provide at least an opening for proto-democratic intervention. However, since secrecy is a prime virtue of AI capital, it can be difficult to obtain information on its data-intensive processes. One might thus look into the technical literature on AI for concerns which might be exploited by those resisting capitalist AI, such as 'adversarial attacks' which exploit the pattern recognition properties of machine learning to render model output inaccurate (Samangouei, Kabkab and Chellappa 2018). However, technical problems need to be considered in relation to how they are implicated within valorisation processes. Thus, a fruitful direction for research is business-oriented literature on AI adoption and production. This is generated by the producers of AI commodities – Microsoft's online AI Business School and Google's array of free AI education courses are two examples – but also by a wide variety of industry promoters, from consulting firms, see Accenture's (n.d.) guide to 'AI for Business Transformation' and management-oriented books like *The Executive Guide to Artificial Intelligence* (Burgess 2018). Such sources can reveal what AI capitals are worried about and indicate potential bottlenecks amenable to outsider intervention. Once identified, bottlenecks can be analysed and the social relations which support valorisation via AI therein might be replaced with alternative social relations not amenable to the data-hungry valorisation of AI capital. Finding bottlenecks returns us to the question of visibility, without which strategy cannot be formulated. I hope this chapter will contribute to an incremental increase in visibility and, perhaps, a half step towards a strategy.

References

Accenture. n.d. AI for Business Transformation. https://www.accenture.com/us-en/services/applied-intelligence/business-transformation

Agarwal, D. 2019. Chat with Kevin Scott EVP and CTO of Microsoft. 15 July. *Venture Beat*. Last accessed 5 May 2020: https://www.youtube.com/watch?v=npXxiVaY9p0

Alharby, M. and Van Moorsel, A. 2017. Blockchain-Based Smart Contracts: A Systematic Mapping Study. *arXiv preprint arXiv:1710.06372*.

Alpaydin, E. 2014. *Introduction to Machine Learning*. Cambridge, MA: MIT Press.

Amadeo, R. 2018. Google's Iron Grip on Android: Controlling Open Source by Any Means Necessary. 21 August. *Ars Technica*. Last accessed 5 May 2020: https://arstechnica.com/gadgets/2018/07/googles-iron-grip-on-android-controlling-open-source-by-any-means-necessary

Amodei, D. and Hernandez, D. 2018. AI and Compute. 16 May. *OpenAI Blog*. Last accessed 5 May 2020: https://openai.com/blog/ai-and-compute

Ananthraj, V. 2019. Madhusudan Shekar on how Amazon is democratizing AI. 31 July. *TechCircle*. Last accessed 5 May 2020: https://www.techcircle.in/2019/07/31/madhusudan-shekar-on-how-amazon-is-democratizing-ai

Andrejevic, M. 2020. *Automated Media*. New York: Routledge.

Athanasiou, T. 1985. Artificial Intelligence: Cleverly Disguised Politics. In: T. Solomonides and L. Levidow (Eds.), *Compulsive Technology: Computers as Culture*, pp. 13–35. London: Free Association Books.

Bailey, B. 2019. Power is Limiting Machine Learning Deployments. 25 July. *Semiconductor Engineering*. Last accessed 15 May 2020: https://semiengineering.com/power-limitations-of-machine-learning

Bass, T. and Old, R. 2020. Common Knowledge: Citizen-led Data Governance for Better Cities. Decode Project. https://decodeproject.eu/publications/common-knowledge-citizen-led-data-governance-better-cities

Bastani, A. 2019. *Fully Automated Luxury Communism*. New York: Verso.

Bernes, J. 2013. Logistics, Counter Logistics and the Communist Prospect. *Endnotes* 3. Last accessed 5 May 2020: https://endnotes.org.uk/issues/3/en/jasper-bernes-logistics-counterlogistics-and-the-communist-prospect

Bernes, J. 2018. The Belly of the Revolution: Agriculture, Energy and the Future of Communism. In: B.R. Bellamy and J. Diamanti (Eds.), *Materialism and the Critique of Energy*, pp. 331–375. Chicago: MCM.

Braverman, H. 1998. *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. New York: NYU Press.

Brockman, G. 2019. Microsoft Invest in and Partners with OpenAI to Support us Building Beneficial AI. 22 July. *OpenAI Blog*. Last accessed 5 May 2020: https://openai.com/blog/microsoft

Broussard, M. 2018. Artificial Unintelligence: How Computers Misunderstand the World. Cambridge, MA: MIT Press.

Brynjolfsson, E. and McAfee, A. 2017. The Business of Artificial Intelligence. July. *Harvard Business Review*. Last accessed 10 May 2020: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Burgess, A. 2018. *The Executive Guide to Artificial Intelligence: How to Identify and Implement Applications for AI in your Organization*. London: Palgrave Macmillan.

Castelvecchi, D. 2016. Can we Open the Black Box of AI? 5 October. *Nature News & Comment*. Last accessed 15 May 2020: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731

Cockshott, P. 1988. Application of Artificial Intelligence Techniques to Economic Planning. University of Strathclyde. Last accessed 5 May 2020: http://www.dcs.gla.ac.uk/~wpc/reports/plan_with_AIT.pdf

Cockshott, P. 2017. Big Data and Supercomputers: Foundations of Cyber Communism. 26–28 September. Presented at The Ninth International Vanguard Scientific Conference on 100 Years of Real Socialism and the Theory of Post-Capitalist Civilization, Hanoi, Vietnam.

Cooley, M. 1981. On the Taylorisation of Intellectual Work. In L. Levidow and R. Young (Eds.), *Science, Technology and the Labour Process Volume 2*. London: CSE Books.

Cook, G. 2017. Clicking Clean: Who is Winning the Race to Build a Green Internet? *Greenpeace*. Last accessed 5 May 2020: https://storage.googleapis.com/planet4-international-stateless/2017/01/35f0ac1a-clickclean2016-hires.pdf

Cooper, J. 1977. The Scientific and Technical Revolution in Soviet Theory. In: F. Fleron (Ed.), *Technology and Communist Culture: The Socio-Cultural Impact of Technology Under Socialism*, pp. 146–179. New York: Praeger.

Crawford, K. and Joler, V. 2018. Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources. *AI Now Institute and Share Lab*. Last accessed 5 May 2020: https://anatomyof.ai

Domingos, P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake our World*. New York: Basic Books.

Dyer-Witheford, N. 2013. Red Plenty Platforms. *Culture Machine* 14, 1–27.

Dyer-Witheford, N., Kjøsen, A.M. and Steinhoff, J. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. London: Pluto.

Edwards, P., Jackson S., Bowker, G. and Knobel, C. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Report of a Workshop on 'History and Theory of Infrastructure: Lessons for New Scientific Cyberninfrastructures'. Last accessed 5 May 2020: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf

Feigenbaum, E., McCorduck, P. and Nii, H. 1989. *The Rise of the Expert Company: How Visionary Companies Are Using Artificial Intelligence to Achieve Higher Productivity and Profits*. New York: Vintage Books.

Fingas, J. 2019. Amazon and Google Ask for Non-Stop Data from Smart Home Devices. 13 February. *Engadget*. Last accessed 15 May 2020: https://www.engadget.com/2019-02-13-amazon-and-google-continuous-smart-home-data.html

Garigliano, R. and Mich, L. 2019. Looking Inside the Black Box: Core Semantics Towards Accountability of Artificial Intelligence. In: M.H. ter Beek, A. Fantechi and L. Semini (Eds.), *From Software Engineering to Formal Methods and Tools, and Back: Essays Dedicated to Sefania Gnesi on the Occasion of Her 65th Birthday*, pp. 250–266. Cham, Switzerland: Springer.

Garvey, C. 2018. A Framework for Evaluating Barriers to the Democratization of Artificial Intelligence. Presented at *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. pp. 8079–8080. Last accessed 5 May 2020: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17320/16477

Gitelman, L. (Ed). 2013. *Raw Data is an Oxymoron*. Cambridge, MA: MIT Press.

Google. 2019. The Size and Quality of a Data Set. *Google Developers Machine Learning Crash Course*. Last accessed 5 May 2020: https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality

Gray, M.L. and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York: Houghton Mifflin Harcourt.

Harvey, D. 2007. *A Brief History of Neoliberalism*. Oxford: Oxford University Press.

Haugeland, J. 1989. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Heaven, Will. 2020. Our Weird Behavior During the Pandemic is Messing with AI Models. 11 May. *MIT Technology Review*. Last accessed 5 May 2020: https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/?truid=8f7239a3e8ff7abf667fea197c1218cd&utm_source=the_download&utm_medium=email&utm_campaign=the_download.unpaid.engagement&utm_content=05-13-2020

Hedges, C. and Sacco, J. 2014. *Days of Destruction, Days of Revolt*. Toronto, Canada: Vintage Canada.

Huber, M. 2020. Socialise the Food System. 19 April. *Tribune*. Last accessed 5 May 2020: https://www.tribunemag.co.uk/2020/04/socialise-the-food-system

ImageNet. 2010. About ImageNet. *ImageNet.org*. Last accessed 5 May 2020: http://image-net.org/about-statssrnicke

Jameson, F. 2009. *Valences of the Dialectic*. London: Verso.

Kisner, J., Wishnow, D. and Ivannikov, T. 2017. IBM: Creating Shareholder Value with AI? Not So Elementary, My Dear Watson. 12 July. Jeffries Franchise Notes. Retrieved from: https://javatar.bluematrix.com/pdf/fO5xcWjc

Kittler, F. 1995. There is No Software. *ctheory*. Last accessed 5 May 2020: http://ctheory.net/ctheory_wp/there-is-no-software

Land, N. 2014. The Teleological Identity of Capitalism and Artificial Intelligence. Remarks to the Participants of the Incredible Machines 2014 Conference. 8 March. Formerly online at this address: http://incrediblemachines.info/nick-land-the-teleological-identity-of-capitalism-and-artificial-intelligence

Land, N. 2017. A Quick and Dirty Introduction to Accelerationism. 25 May. *Jacobite*. Last accessed 5 May 2020: https://jacobitemag.com/2017/05/25/a-quick-and-dirty-introduction-to-accelerationism

Larkin, B. 2013. The Politics and Poetics of Infrastructure. *Annual Review of Anthropology*, 42, 327–343.

Lenin, V.I. 1971 [1918]. The Immediate Tasks of the Soviet Government. In: *Lenin Collected Works Volume 42* (2nd English Edition). Moscow: Progress Publishers.

Leonardi, P.M. and Barley, S.R. 2008. Materiality and Change: Challenges to Building Better Theory About Technology and Organizing. *Information and Organization*, 18(3), 159–176.

Li, F. 2017. ImageNet: Where have we been? Where are we going? ACM Webinar. Last accessed 5 May 2020: https://learning.acm.org/binaries/content/assets/leaning-center/webinar-slides/2017/imagenet_2017_acm_webinar_compressed.pdf

Marx, K. 1990. *Capital Volume One*. London: Penguin Classics.

Marx, K. 1991. *Capital Volume Three*. London: Penguin Classics.

Marx, K. 1993. *Grundrisse: Foundations of the Critique of Political Economy*. London: Penguin Classics.

Marx, K. and Engels, F. 1969. Manifesto of the Communist Party. *Marx/Engels Selected Works Volume 1*. Moscow: Progress Publishers.

Mason, P. 2016. *Postcapitalism: A Guide to Our Future*. New York: Farrar, Straus and Giroux.

Medina, E. 2011. *Cybernetic Revolutionaries: Technology and Politics in Allende's Chile*. Cambridge, MA: MIT Press.

Morozov, Evgeny. 2019. Digital Socialism? The Calculation Debate in the Age of Big Data. *New Left Review* 116. https://newleftreview.org/issues/II116/articles/evgeny-morozov-digital-socialism

Pasquinelli, M. and Joler, V. 2020. The Nooscope Manifested: Artificial Intelligence as Instrument of Knowledge Extractivism. KIM HfG Karlsruhe and Share Lab. 1 May. Retrieved from: http://nooscope.ai

Pearce, F. 2018. Energy Hogs: Can World's Huge Data Centers Be Made More Efficient? 3 April. *Yale Environment 360*. Last accessed 5 May 2020: https://e360.yale.edu/features/energy-hogs-can-huge-data-centers-be-made-more-efficient

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S. and Niebles, J.C. 2019. *The AI Index 2019 Annual Report*. Human-Centered AI Institute. Stanford, CA: Stanford University.

Peters, B. 2016. *How Not to Network a Nation: The Uneasy History of the Soviet Internet*. Cambridge, MA: MIT Press.

Phillips, L. and Rozworski, M. 2019. *The People's Republic of Wal-Mart: How the World's Biggest Corporations Are Laying the Foundation for Socialism*. London: Verso.

Polovets, L. 2015. The Value of Data Part 1: Using Data as Competitive Advantage. 27 February. *Coding VC*. Last accessed 5 May 2020: https://codingvc.com/the-value-of-data-part-1-using-data-as-a-competitive-advantage

Ramtin, R. 1991. *Capitalism and Automation: Revolution in Technology and Capitalist Breakdown*. London: Pluto.

Samangouei, P., Kabkab, M. and Chellappa, R. 2018. Defense-gan: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *arXiv preprint* 1805.06605.

Sastry, G., Clark, J., Brockman, G. and Sutskever, I. 2019. Compute Used in Older Headline Results. *OpenAI Blog*. Last accessed 5 May 2020: https://openai.com/blog/ai-and-compute/#addendum

Schmitz, M. 2017. Why Your Models Need Maintenance. 12 May. *Towards Data Science*. Last accessed 5 May 2020: https://towardsdatascience.com/why-your-models-need-maintenance-faff545b38a2

Smith, T. 2009. The Chapters on Machinery in the 1861–63 Manuscripts. In: R. Bellofiore and R. Fineschi (Eds.), *Re-Reading Marx: New Perspectives After the Critical Edition*, pp. 112–127. Basingstoke: Palgrave Macmillan.

Srnicek, N. 2017. The Challenges of Platform Capitalism: Understanding the Logic of a New Business Model. *Juncture*, 23(4), 254–257.

Srnicek, N. and Williams, A. 2015. *Inventing the Future: Postcapitalism and a World Without Work*. London: Verso.

Stanford, S. 2019. Artificial Intelligence (AI): Salaries Heading Skyward. 20 September. *Towards AI*. Last accessed 5 May 2020: https://medium.com/towards-artificial-intelligence/artificial-intelligence-salaries-heading-skyward-e41b2a7bba7d

Steinhoff, J. 2019. Cognition on Tap. *Digital Culture & Society*, 4(2), 89–104.

Sudmann, A. (Ed.). 2020. *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*. Berlin: Transcript-Verlag.

Sylvester-Bradley, O. 2018. The Making of the Cooperative Cloud. 3 April. *The Open Co-op*. Last accessed 5 May 2020: https://open.coop/2018/04/03/making-of-the-coop-cloud

Vincent, N., Hecht, B. and Sen, S. 2019. 'Data Strikes': Evaluating the Effectiveness of a New Form of Collective Action Against Technology Companies. In: *The World Wide Web Conference*, 1931–1943. New York: ACM.

Vincent, N., Li, H., Tilly, N., Chancellor, S. and Hecht, B. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. *arXiv preprint*. 2012.09995.

von Mises, L. 1935. Calculation in the Socialist Commonwealth. In: F.A. Hayek (Ed.), *Collectivist Economic Planning*. London: Routledge.

Williams, A. and Srnicek, N. 2014. #accelerate: Manifesto for an Accelerationist Politics. In R. Mackay and A. Avanessian (Eds.), *#accelerate: The Accelerationist Reader*. Falmouth: Urbanomic.

Wu, H. 2018. China is Achieving AI Dominance by Relying on Young Blue-Collar Workers. *Motherboard*. 21 December. https://www.vice.com/en_us/article/7xyabb/china-ai-dominance-relies-on-young-data-labelers

Yegulalp, S. 2017. Google's Machine Learning Cloud Pipeline Explained. 19 May. *Infoworld*. Last accessed 5 May 2020: https://www.infoworld

.com/article/3197405/tpus-googles-machine-learning-pipeline-explained.html

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* London: Profile Books.

CHAPTER 9

Creating the Technological Saviour: Discourses on AI in Europe and the Legitimation of Super Capitalism

Benedetta Brevini

Introduction

Dominant narratives in public fora, and increasingly within governments, place great importance on nations achieving leadership in artificial intelligence (AI). What is becoming clear is that world leaders are invested in making AI the business opportunity of the future – and thereby selling it as a virtue and a public good (Economist 2017, 2018; World Economic Forum 2018).

Scholars in political economy of communication have shown how discourses around digital technologies have historically been constructed around modern myths (Mosco 2004) with major references to utopian worlds and possibilities. Myths, conceived as the dominant ideologies of our time (Barthes 1993) become powerful devices that normalise conventional wisdom into 'common sense' (Gramsci 1971), thus making the conception of alternatives virtually impossible. As a result, digital developments and policies are adopted without the benefit of an informed debate (Brevini 2020).

Europe is rarely considered a leader in AI developments, but rather, seems to struggle to find its own voice squeezed between China and the United States. However, 2018 was a crucial year in Europe for the advancement of national and EU strategies on AI. The journey to develop an AI strategy in the EU

How to cite this book chapter:
Brevini, B. 2021. Creating the Technological Saviour: Discourses on AI in Europe and the Legitimation of Super Capitalism. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 145–159. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.i.

started in April 2018, when the European Commission presented the 'Declaration of Cooperation on AI' now signed by all 28 Member States including Norway. Member states pledged to work together towards 'a comprehensive and integrated European approach on AI' (EU Declaration 2018). The declaration was followed by two communications reports by High-Level Expert groups on AI (High-Level Expert Group 2019); and a White Paper was published in February 2020.

This chapter aims to unwrap the recurrent myths employed in discourses on AI in Europe. In doing so, this work aims to embrace a research agenda that integrates political economy (Mosco 2004; Fuchs 2015; McChesney 2013) with cultural analysis, thus considering the idea of myth and mythmaking as an essential dimension of inquiry.

Why Dominant Discourses Are Crucial

That technology discourses have a central role in the legitimation of a specific political economic order has been at the centre of scholarship debates for some time (Mosco 2004; Freedman 2002; Brevini 2020). For example, Fisher (2010) has shown how technology discourse legitimated the 'post-Fordist phase' of capitalism characterised by 'the weakening of labour and the state vis-a-vis capital, the liberalization of markets, the privatization of work, and the flexibilization of employment' (Fisher 2010, 234). After all, technological 'fixes' have historically been crucial to solve potential barriers to capital accumulation. As David Harvey (2005) argues, technology becomes 'a prime mover' of capitalist growth (Harvey 2005). Likewise, several studies in the fields of history of technology, sociology and political economy of communication have shown the ideological functions of technology discourse (Mosco 2004; Barbrook and Cameron 1996; Dean 2002).

In this chapter I am drawing, in particular, on analysis that recognises the crucial role of myths in building discourses. In the *Digital Sublime* (2004) Mosco explains how myths are used to claim how digital technology is capable of triggering an historical break: 'Almost every wave of new technology, including information and communication media, has brought with it declarations of the end … Since these tend to take place with no reference to similar proclamations in the previous wave, one cannot help but conclude that the rhetoric of technology, the technological sublime that David Nye so perceptively identifies, is powerful enough to create a widespread historical amnesia' (Mosco 2004, 117).

There are three crucial ways in which myths are used in the context of legitimising the status quo. Firstly, they are used as a weapon to control political debates. Secondly, they are used to depoliticise discourses that would otherwise show their contested political character. Thirdly, they are a crucial component of hegemonies, thus making it difficult for a counter-hegemonic discourse to arise. Barthes (1993) clearly elaborated on this conjoint relation between myths

and the political construction of reality in contemporary democracies. According to Barthes, myths have the ability to construct 'common sense', thus favouring established relations of power. But it was Gramsci (1971) who provided the crucial link between common sense and discourse. For Gramsci, myths are essentially common sense, defined as 'not something rigid and immobile, but … continually transforming itself, enriching itself with scientific ideas and with philosophical opinions that have entered ordinary life. Common sense creates the folklore of the future, that is a relatively rigid phase of popular knowledge at a given place and time.' (Gramsci 1971, 326). Through this process, the values of powerful elites are naturalised, becoming the default position against which, all things are assessed and compared. Thus, myths, here conceptualised as common sense can influence and shape discourse and policy making in way that, as Wyatt notes: 'sometimes today's imaginary becomes tomorrow's lived reality' (Wyatt 2004, 244). It is through the legitimation of dominant discourses (Brevini and Schlosberg 2016; Foucault 1980, 1981) when discourses become hegemonic (Gramsci 1996; Brevini 2020), that they can direct attention from the public, construct and promote digital developments, communication policy and legitimate modes of governance that would not have been possible without the establishment of such a discourse (Brevini and Schlosberg 2016). Incomplete discourses that become dominant can shape how society embraces technological developments.

Tech-Determinism, Tech-Solutionism and AI

The technological deterministic argument that technology can and will fix capitalism – and its intrinsic power to exacerbate inequalities of economic, racial, gender forms – is far from being a recent elaboration (Gilder 1990; Negroponte 1998). To use the words of Mosco, 'one generation after another has renewed the belief that, whatever was said about earlier technologies, the latest one will fulfil a radical and revolutionary promise' (Mosco 2004, 21; Brevini 2020). Mosco (2004) rightly reminds us of James Carey's (1992) work that discussed how machines have often been framed employing a powerful religious ethos: 'in contemporary popular commentary and even in technical discussions of new communications technology, the historic religious undercurrent has never been eliminated from our thought' (Carey 1992, 18).

As a result, technology becomes the most powerful weapon purporting to lift the global capitalist system out of its recurrent crises; and virtually any social problem can be subject to a technical and technological fix (Kurzweil 1985). Development of digital technology, we are reassured, will empower people out of radical inequalities, while naturalising market-based solutions to every issue of governance. Raymond Williams, one of the most established cultural theorists to come out of Britain, offers a fruitful definition of technological determinism as a 'largely orthodox view of the nature of social change' (Williams 1974, 13).

Furthermore, he explains: 'The basic assumption of technological determinism is that a new technology – a printing press or a communications satellite – "emerges" from technical study and experiment. It then changes the society or sector into which it has "emerged"' (Williams 1985, 129). On the contrary, despite William's belief in the opportunities offered by innovation, he held that 'technology is always in a full sense social', thus its development and usage are always shaped by the social relations of the society in which they are adopted (Williams 1981, 227).

Williams was writing at the time when, by the late 1970s, the so called 'information revolution' was just emerging as the new dogma in government and corporate planning (Dyer-Witheford 1999). But the information revolution myth kept getting stronger throughout the 1970s, 1980s and 1990s, 'more attuned to the climate of Thatcherism and Reaganism' (ibid. 21) than to a Keynesian state's framework. This revolution should not come as a surprise since neoliberalism and the information revolution have been endorsed by corporate and governments elites as the solution to the 'growth' crisis of the 1970s. The neoliberal Clinton administration of the 1990s was an aggressive supporter of the technocratic 'information revolution'. In 1994 its congress passed the National Information Infrastructure Bill which launched the world famous 'information superhighway', championed by Al Gore in numerous speeches around the world. Another crucial futurologist of the time stressed once again the link between technological determinism and neoliberal ideologies. Francis Fukuyama's influential book *The End of History* (1992), proclaimed that the end of the Cold War demonstrated the collapse of any reasonable alternative to neoliberalism. Moreover, in order to reinstate the alliance between neoliberalism and technology, in *The Great Disruption* (2017) Fukuyama argues:

> A society built around information tends to produce more of the two things people value most in a modern democracy – freedom and equality. Freedom of choice has exploded, in everything from cable channels to low-cost shopping outlets to friends met on the Internet. Hierarchies of all sorts, political and corporate, have come under pressure and begun to crumble. (Fukuyama 2017, 4)

In sum, this hegemonic Silicon Valley discourse reaffirms again and again that technological progress not only provides newly enhanced individual freedoms but will lead to radical social change.

As a consequence, what has been dubbed as technological solutionism becomes the only logical consequence of late capitalism (Levina and Hasinoff 2016). The term *tech solutionism* has been popularised by Evgeny Morozov in his 2013 book *To Save Everything, Click Here* as:

> Recasting all complex social situations either as neatly defined problems with definite, computable solutions or as transparent and self-evident

> processes that can be easily optimized – if only the right algorithms are in place! – this quest is likely to have unexpected consequences that could eventually cause more damage than the problems they seek to address. I call the ideology that legitimises and sanctions such aspirations 'solutionism'. (Morozov 2013, 5)

From its beginnings in the 1950s, AI has not been exempted from these claims of offering a 'solution' to the inequalities of capitalism (Brevini 2020; Natale and Ballatore 2020; Elish and Boyd 2018). On the contrary, it has been surrounded by evocative claims about the imminent creation of a machine capable of surpassing the potentials of humankind. AI has often been hailed as the magic tool to rescue the global capitalist system from its dramatic failures (Brevini 2020).

Recent studies on popular and public debates on AI have started to show the extent of the dominance of this tech-deterministic ideology, especially in the US (Mayer-Schönberger and Cukier 2013). For example, Elish and Boyd's research (2018) on AI rhetoric, concluded that 'through the manufacturing of hype and promise, the business community has helped produce a rhetoric around these technologies that extends far past the current methodological capabilities' (Elish and Boyd 2018, 58). In exploring public discourse shaping the popular imagination around possible AI futures, Goode (2018) observes that contemporary discourse is:

> skewed heavily towards specific voices – predominantly male science fiction authors and techno-centric scientists, futurists and entrepreneurs – and the field of AI and robotics is all too easily presented as a kind of sublime spectacle of inevitability (…) that does little to offer lay citizens the sense that they can be actively involved in shaping its future. (Goode 2018, 204)

Furthermore, the latest study on media coverage of AI in the UK conducted by the Reuters Institute (Brennen, Howard and Nielsen 2018) showed that the UK media coverage of AI was overwhelmingly influenced by industry concerns, products and initiatives.

Thus, this chapter aims to contribute to these scholarly debates by investigating hegemonic discourses about AI emerging from the European Union's official strategy on AI. In particular, it will highlight the most crucial myths on which hegemonic discourse is based.

Developing AI in Europe

The journey to develop the AI strategy in the EU started in 2018, when the European Commission presented the 'Declaration of Cooperation on AI' now signed by all 28 Member States, including Norway. In the Declaration,

Member States agree to a continuous dialogue to work together towards 'a comprehensive and integrated European approach on AI and, where needed, review and modernise national policies to ensure that the opportunities arising from AI are seized and the emerging challenges addressed' (EU Declaration 2018: 4).

The AI strategy is developed within the context and legislative packages of the Digital Single Market Strategy developed by the EC that include the European Data Economy initiatives, the General Data Protection Directive and, crucially, the European Cloud Initiative. The latter aims to 'make it easier for researchers, businesses and public services to fully exploit the benefits of Big Data by making it possible to move, share and re-use data seamlessly across global markets and borders, and among institutions and research disciplines' (European Cloud initiative 2019).

On 7 December 2018 the European Commission published a coordinated action plan on the development of AI in the EU (European Commission 2018a, 2018b). It pledged to increase its annual investments in AI by 70% under the research and innovation programme Horizon, in order to reach EUR 1.5 billion for the period 2018–2020. In its Communication (European Commission 2018a) the European Commission (EC) reaffirms the belief that 'AI will help us to solve some of the world's biggest challenges', from treating chronic diseases and reducing fatality rates in traffic accidents to fighting climate change and anticipating cybersecurity threats (European Commission 2018a, 2). Therefore, the EC put forward a European approach to artificial intelligence based on three pillars:

- connect and strengthen AI research centres across Europe;
- support the development of an 'AI-on-demand platform' that will provide access to relevant AI resources in the EU for all users;
- support the development of AI applications in key sectors (European Commission 2018b, 1).

In order to support the development of the AI strategies summarised here, the EC established two advisory entities: The High-Level Expert Group on AI (HLEG); and the European AI Alliance. The High-Level Expert Group on AI is charged with developing proposals for the overall EU's AI strategy, policy and priorities. It comprises 23 members from industry, 19 from academia and 10 from civil society; and it is further divided into two working groups: one on ethics; and one on investment and policy. The second advisory entity, the European AI Alliance, is a multi-stakeholder online platform. On the platform, EU members can contribute to ongoing discussions on Al, feeding into the European Commission's policy-making processes. The European AI Alliance is conceived as a tool open to all members of society. Currently, it is composed of members from civil society, trade unions, companies, not-for-profit institutions and consumer organisations.

In the first year after its creation in June 2018, the HLEG released two major policy documents forming the basis of the latest White Paper on AI, adopted in 2020. The first document, Ethics Guidelines on artificial intelligence, put forward the concept of 'Trustworthy AI' and the key requirements that AI systems should meet in order to be trustworthy (High-Level Expert Group 2019a). The second document, Policy and Investment Recommendations (High-Level Expert Group 2019b), developed recommendations for AI towards sustainability, growth and competitiveness and inclusion. On 19 February 2020, the European Commission published a White Paper on artificial intelligence (European Commission 2020) aiming to foster a European ecosystem of excellence and trust in AI and a report on the safety and liability aspects of AI. The White Paper provides a simple, all-encompassing definition of artificial intelligence 'AI is a collection of technologies that combine data, algorithms and computing power. Advances in computing and the increasing availability of data are therefore key drivers of the current upsurge of AI' (European Commission 2020, 2). The White Paper is clear on twofold goals: on the one hand it aims to support the AI uptake and on the other it aims to address the risks linked to particular uses of it. These overall aims will be achieved through coordinated measures that will streamline research, foster collaboration between member states and increase investment into AI development and deployment; and through a policy toolkit for a future EU regulatory framework that would determine the types of legal requirements that would apply to relevant actors, with a particular focus on high-risk applications (European Commission 2020a). Although the White Paper does not set out a concrete framework for new AI legislation, it does set out the Commission's key priorities.

Three Myths in Discourses on AI in Europe

Having outlined the current European Framework developed in the series of communications, High-Level Groups reports and lastly, the White Paper on AI, this section uncovers the recurrent myths employed in official EU plans to develop artificial intelligence. As discussed in the previous section, these myths become crucial components of AI discourse, justifying policy-making within the European Union. Furthermore, as I will argue, these myths construct a discourse that has the ultimate end of reinforcing the current neoliberal ideology of the current stage of capitalism.

Myth #1: Artificial Intelligence as a Solution for Humanity and Capitalism's Biggest Challenges

In its communications of 25 April 2018 and 7 December 2018, the European Commission set out its vision for AI, which supports 'ethical, secure and cutting-edge AI made in Europe' (European Commission 2018a).

The vision could not highlight in a more striking way how AI becomes the solution for humanity's biggest challenge. The following two paragraphs taken from the two official communications (European Commission 2018a) could not be clearer:

> AI is helping us to solve some of the world's biggest challenges: from treating chronic diseases or reducing fatality rates in traffic accidents to fighting climate change or anticipating cybersecurity threats. (ibid. 2)

In more evocative terms, the myth of the revolutionary character of AI is reinforced by a comparison with the 'steam' and electricity 'revolution'.

> Like the steam engine or electricity in the past, AI is transforming our world, our society and our industry. Growth in computing power, availability of data and progress in algorithms have turned AI into one of the most strategic technologies of the 21st century. (ibid. 2)

The High-Level Expert Group on Artificial Intelligence (AI HLEG) goes into even greater detail about the capabilities of AI to make humanity 'flourish', thus solving all problems of society.

> We believe that AI has the potential to significantly transform society. AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes, and supporting how we monitor progress against sustainability and social cohesion indicators. (High-Level Expert Group 2019a, 4)

It's impossible not to see in this mythical discourse the same rhetoric of technocrats of the 1990s (Gilder 2000; Fukuyama 1992; Shirky 2008) that argued how the new communicative opportunities provided by the internet would enhance a new era for democracy (Gilder 2000; Negroponte 1998), the end of history (Fukuyama 1992) and the beginning of a new era of freedom. The same ideological discourse is replicated in current techno-enthusiast claims about the cloud (Nye 1994) more recently debunked by Mosco in his book, *To the Cloud: Big Data in a Turbulent World* (Mosco 2014).

In pure enlightenment fashion, this absolute faith in technology, embraced and supported by cybertarians' Silicon Valley circles (Dyer-Witheford 1999; Brevini 2020) turns into a powerful apology for the status quo and the current structure of capitalism, without any real space for critique.

Myth #2: Creating Urgency and 'Preparing' Society – AI as Ineluctable

The second of the most compelling myths emerging from my analysis of EU strategies on AI is the myth of AI's perceived *ineluctability*, built through a constant emphasis on its urgency. Consider for example this quote, from the European Commission Communication of April 2018:

> The stakes could not be higher. The way we approach AI will define the world we live in. Amid fierce global competition, a solid European framework is needed. (European Commission 2018a, 2)

Moreover, the White Paper – that is the latest policy document adopted by the EC to establish its framework (European Commission 2020) – stresses again the urgency for every sector of Public services to employ AI as soon as possible.

> It is essential that public administrations, hospitals, utility and transport services, financial supervisors, and other areas of public interest rapidly begin to deploy products and services that rely on AI in their activities. (European Commission 2020, 8)

Overall, discourse stressing the need to hurry up on investments – such as 'Europe is behind in private investments on AI' (European Commission 2018a, 5), or 'the European industry cannot miss the train' (European Commission 2018a, 5) – are reiterated throughout the documents developing EU strategy on AI. So fast paced is the race to adopt AI that the opposite would be inconceivable:

> Without such efforts, the EU risks losing out on the opportunities offered by AI, facing a brain-drain and being a consumer of solutions developed elsewhere. (European Commission 2018a, 6)

The myth of *AI ineluctability* is further enhanced by repetition of sentences reaffirming the role of the EU as enabler of AI, with an almost teleological duty to 'better prepare our society for AI' (European Commission 2018b, 5) as if its divine advent on earth was inevitable.

This should remind us of the dawn of AI developments in the 1950s (Roszak 1986), when popular accounts proclaimed the imminent development of intelligent machines capable of outsmarting the human mind amid promises to fundamentally change everything. However, as Goode (2018) recalls, in the last decade we have seen a clear increase of predictions that the arrival of superintelligence is imminent, thus the urgency (Goode 2018) this calls for in producing EU level strategies. Claims like 'The singularity is near', by Ray Kurzweil, futurist and Director of Engineering at Google are indicative of the current 'anxiety surrounding the speed with which the technology appears to be developing, something that some robotics companies are keen to play up' (Goode 2018, 199).

Unveiling this myth of the ineluctability of AI and its urgency, it is impossible not to recall Williams' analysis in *Towards 2000* where he stated that 'The sense of some new technology as inevitable or unstoppable is a product of the overt and covert marketing of the relevant interests' (Williams 1985, 133). In reality technological development is *not* predetermined, and alternative paths to a market-led development that reinforces the current neoliberal status quo are always a possibility (Brevini 2020).

Myth #3: AI Surpassing Human Intelligence

Like every institution that developed a strategy for AI, the EC also had to start by defining AI. The Communication of the Commission clarifies that:

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis, software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications). (European Commission 2018a, 2)

Moreover, what emerges from the EU documents is the underlying assumption that artificial intelligence will outperform human capabilities. In several documents, the EC explains that AI has the capacity to transform 'our world', our 'society', our 'work' (The Communication, European Commission 2018b, 1), thus implying that its abilities will exceed human cognitive functions. Take for example this statement:

> AI needs vast amounts of data to be developed. Machine learning, a type of AI, works by identifying patterns in available data and then applying the knowledge to new data. The larger a data set, the better AI can learn and discover even subtle relations in the data. Once trained, algorithms can correctly classify objects that they have never seen, in more and more cases with accuracies that exceed those of humans. (European Commission 2018b, 6)

In the 1980s Roszak had already implemented the term 'technological idolatry' that propagates a deference to computers 'which human beings have never assumed with respect to any other technology of the past' (Roszak 1986, 45). This clearly reveals how the construction of AI as machines that can outperform human labour helps legitimise current capitalistic structures that are indeed capable of generating the technocratic imperative that see the

subordination of human labour to computers. Of course, as I have discussed above, neoliberalism has long established a privileged relationship with technology as the 'prime mover' of capitalist growth (Harvey 2005).

Conclusion

The AI myths discussed in these pages are very powerful tools for the construction of a discourse that make us perceive AI as the solution to the major problems in our society, including the inequalities brought about by capitalism and other major crises such as climate change and global health emergencies. Through these myths, AI then becomes the technological saviour, whose advent is ineluctable. As such, when the artificial machine arrives – in this future/present which is always inevitably imminent – it will manifest as a superior intelligence to solve the problems that capital economies have themselves created. Eventually, AI will outsmart humans to mend that damage and ameliorate further risks that capitalism inevitably occasions.

The recurrent myths that are omnipresent in the European Framework for AI have two major consequences. Firstly, they structure a hegemonic discourse that makes it impossible to think of alternative paths, framing resistance as futile because technological development is predestined. Accordingly, they legitimise a neoliberal ideology that pushes consumerism and productivity above all values and strips technology from the social relations that are at the basis of technology development (Williams 1985; Brevini 2020). Secondly, they redirect public discourse, by obfuscating and inhibiting a serious debate on the structural foundations of AI, its progressively concentrated ownership and the materiality of its infrastructures. Taken together, these myths of AI, construct a type of discourse that frames the problem of AI in a way that excludes any emphasis on crucial questions of ownership, control and the public interest. It also diverts attention from known problems of inequality, discrimination and bias of data analysed algorithmically that lies at the heart AI systems (Brevini and Pasquale 2020). When these crucial questions are asked, they are only addressed through the 'AI ethical' framework that has little to say about the structural inequalities on which AI is built (Wagner 2018).

This optimism for AI possibilities and achievements so popular in Europe and in the West, is obviously fuelled by extremely effective lobbying efforts by the most powerful technology giants that are already dominating the market and debate. From Alphabet to Amazon, to Microsoft, IBM and Intel, we have evidence that the giants of Silicon Valley are investing billions both on AI developments and on setting the terms of public debates on AI and determining policy outcomes (Benkler 2019). Thus, a central concern of this chapter is the migration of strategic decisions and choices on the direction of AI development from government to corporate board rooms: the privatisation of public policy.

Major lobby groups go in to bat for their vested interests in the policy arena, armed with funded academic research on the benefits of AI and efficiency. For example, a report published in 2019 by the *New Statesman* revealed that in five years Google has spent millions of pounds funding research at British universities including the Oxford Internet Institute (Williams 2019), while DeepMind, Alphabet's own AI company, has specifically supported studies on the ethics of AI and automated decision-making. Correspondingly, Facebook donated US $7.5m to the Technical University of Munich, to fund new AI ethics research centres. Another troubling case is the US-based National Science Foundation program for research into 'Fairness in Artificial Intelligence', co-funded by Amazon (Benkler 2019). As scholar Yochai Benkler explained, the digital giant has 'the technical, the contractual, technical and organizational means to promote the projects that suit its goals' (ibid. 2019). Hence, 'Industry has mobilized to shape the science, morality and laws of Artificial Intelligence' (ibid. 2019).

Moreover, this portrayal of AI as the magic, divine hand that will rescue society also obfuscates the materiality of the infrastructures that are central to the environmental question that has been so consistently and artfully ignored (Brevini 2020). AI relies on technology, machines and infrastructures that deplete scarce resources in their production, consumption and disposal, thus increasing amounts of energy in their use, and exacerbating problems of waste and pollution. AI generates an array of environmental problems, most notably energy consumption and emissions, material toxicity and electronic waste (Brevini and Murdock 2017). Yet these myths help build a discourse that is skewed heavily towards specific voices – predominantly corporate and neoliberal – that build a so-called common sense that is too pervasive to challenge. AI brings us to a present/future in which alternative paths to current capitalism are unthinkable. And so, we surrender to our inevitable destiny of a new world order of wellbeing brought by AI, shaping that future for the benefit of the most powerful who built its technology and framed its hegemonic discourse.

References

Barbrook, R. and Cameron, A. (1996). The Californian Ideology, *Science as Culture* 26: 44–72.

Barthes, R. 1993. *A Barthes Reader.* New York: Random House.

Benkler, Y. 2019. Don't Let Industry Write the Rules for AI. *Nature*, 569(7755), 161.

Brennen, J. S., Howard, P. N. and Nielsen, R. K. 2018. An Industry-Led Debate: How UK Media Cover Artificial Intelligence. Last accessed 1 May 2020, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-12/Brennen_UK_Media_Coverage_of_AI_FINAL.pdf

Brevini, B. 2020. Black Boxes, Not Green: Mythologizing Artificial Intelligence and Omitting the Environment. *Big Data and Society*, 7(2).

Brevini, B. and Murdock G. 2017. *Carbon Capitalism and Communication: Confronting Climate Crisis*. London: Palgrave Macmillan.

Brevini, B. and Pasquale, F. 2020. Revisiting the Black Box Society by Rethinking the Political Economy of Big Data. *Big Data and Society*, 7(2).

Brevini, B. and Schlosberg, J. 2016. Between Philosophy and Action: The Story of the Media Reform Coalition. In Freedman, D. et al. (Eds.) *Strategies for Media Reform*, pp. 123–137. New York: Fordham University Press.

Dean, J. (2002). *Publicity's Secret: How Technoculture Capitalizes on Democracy*. Ithaca, NY: Cornell University Press.

Dyer-Witheford, N. 1999. *Cyber-Marx. Cycles and Circuits of Struggle in High-Technology Capitalism*. Chicago, IL: University of Illinois Press.

Carey, J. 1992. *Communication as Culture: Essays on Media and Society.* New York: Routledge.

Economist, The. 2017. China May Match or Beat America in AI. *The Economist*. Last accessed, 20 February 2020: https://www.economist.com/news/business/21725018-its-deep-pool-data-may-let-it-lead-artificial-intelligence-china-may-match-or-beat-america?zid=291&ah=906e69ad01d2ee51960100b7fa502595

Economist, The. 2018. In the Struggle for AI China will Prevail. *The Economist*. Last accessed 20 February 2020, https://www.economist.com/books-and-arts/2018/09/27/in-the-struggle-for-ai-supremacy-china-will-prevail

Elish, M. C. and Boyd, D. 2018. Situating Methods in the Magic of Big Data and AI. *Communication Monographs*, 85(1), 57–80.

European Cloud Initiative. 2019. *European Commission*. Last accessed 1 March 2020: https://ec.europa.eu/digital-single-market/en/european-cloud-initiative

European Commission. 2018a. Final Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. *Artificial Intelligence for Europe*. Last accessed 20 February 2020: https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF

European Commission. 2018b. Final Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Last accessed 20 February 2020: https://ec.europa.eu/knowledge4policy/publication/coordinated-plan-artificial-intelligence-com2018-795-final_en

European Commission. 2020. On Artificial Intelligence – A European Approach to Excellence and Trust. Last accessed 20 February 2020: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission. 2020a. A European Approach to Artificial Intelligence. Last accessed 1 May 2020, https://ec.europa.eu/digital-single-market/en/artificial-intelligence

EU Declaration. 2018. *Declaration of Cooperation on AI*. Last accessed 1 March 2020, https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence

Fisher, E. 2010. Contemporary Technology Discourse and the Legitimation of Capitalism. *European Journal of Social Theory*, 13(2), 229-252. DOI: https://doi.org/10.1177/1368431010362289

Foucault, M. 1980. *Power/knowledge: Selected Interviews and Other Writings, 1972–1977*. Brighton: Harvester Press.

Foucault, M. 1981. The Order of Discourse. Inaugural Lecture at the Collège de France, 2 December 1970. In: R. Young (Ed.), *Untying the Text, A Post-Structuralist Reader*. Boston: Routledge and Kegan Paul.

Freedman, D., 2002. A Technological Idiot? Raymond Williams and Communications *Technology. Information, Communication & Society*, 5(3), 425–442.

Fuchs, C. 2015. *Culture and Economy in the Age of Social Media*. New York, London: Routledge.

Fukuyama, F. 1992. *The End of History and the Last Man*. New York: Simon and Schuster.

Fukuyama, F. 2017. *The Great Disruption*. London: Profile Books.

Gilder, G. 1990. A Technology of Liberation. In: R. Kurzweil (Ed.), *The Age of Intelligent Machines*. Cambridge, MA: MIT Press.

Gilder, G. 2000. *Telecosm: How Infinite Bandwidth Will Revolutionize our World*. New York: Simon and Schuster.

Goode, L. 2018. Life, But Not as We Know It: AI and the Popular Imagination. *Culture Unbound: Journal of Current Cultural Research*, 10(2), 185–207.

Gramsci, A, 1971. Gramsci: *Selections from the Prison Notebooks of Antonio Gramsci*. London: Lawrence and Wishart.

Gramsci, A. 1996. Quaderni Dal Carcere. In: A. Rosa (Ed.), *Letteratura Italiana Einaudi, Le Opere*. Torino: Einaudi.

Harvey, D. (2005). *A Brief History of Neoliberalism*. Oxford: Oxford University Press.

High-Level Expert Group (HLEG). 2019a. Ethics Guidelines for Trustworthy AI, 8 April 2019. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

High-Level Expert Group (HLEG). 2019b. Policy and Investment Recommendations for Trustworthy AI. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Kurzweil, J. 1985. Artificial Intelligence: An Ideology for the Information Society. *Studies in Communication and Information Technology*, Working Paper #1. Kingston: Queen's University.

Levina, M. and Hasinoff, A. 2016. The Silicon Valley Ethos: Tech Industry Products, Discourses, and Practices. *Television & New Media*, 18(6), 489–495.

Mayer-Schönberger, V. and Cukier, K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Eamon Dolan/ Houghton Mifflin Harcourt.

McChesney, R. W. 2013. *Digital Disconnect: How Capitalism is Turning the Internet Against Democracy*. New York: The New Press.

Morozov, E. 2013. *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems That Don't Exist*. New York: Allen Lane.

Mosco, V. 2004. *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge, MA: MIT Press.

Mosco, V. 2014. *To the Cloud: Big Data in a Turbulent World*. Boulder, CO: Paradigm.

Mosco, V. 2017. The Next Internet. In B. Brevini and G. Murdock (Eds.), *Carbon Capitalism and Communication: Confronting Climate Crisis*, pp. 95–107. London: Palgrave Macmillan.

Murdock, G. and Brevini, B. 2019. Communications and the Capitalocene: Disputed Ecologies, Contested Economies, Competing Futures. *The Political Economy of Communication*, 7(1), 51–82.

Natale, S. and Ballatore, A. 2020. Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence. *Convergence*, 26(1), 3 -18.

Negroponte, N. 1998. Beyond Digital. *Wired*, 6(12), 288.

Nye, D. E. 1994. *American Technological Sublime*. Cambridge, MA: MIT Press.

Roszak, T. 1986. *The Cult of Information*. New York: Pantheon.

Shirky, C. 2008. *Here Comes Everybody: The Power of Organizing Without Organizations*. New York: Penguin Press.

Wagner, B. 2018. Ethics as an Escape from Regulation: From Ethics-washing to Ethics-shopping? In: M. Hildebrandt (Ed.), *Being Profiling. Cogitas Ergo Sum*, pp. 1–7. Amsterdam: Amsterdam University Press.

Williams, O. 2019. How Big Tech Funds the Debate on AI ethics. *New Statesman*. Retrieved from: https://www.newstatesman.com/science-tech/technology/2019/06/how-big-tech-funds-debate-ai-ethics

Williams, R. 1974. *Television: Technology and Cultural Form*. London: Fontana.

Williams, R. 1981. Communication Technologies and Social Institutions. In R. Williams (Ed.), *Contact: Human Communication and its History*. London: Thames & Hudson.

Williams, R. 1985. *Towards 2000*. Harmondsworth: Penguin.

World Economic Forum. 2018. Harnessing Artificial Intelligence for the Earth World Economic Forum System Initiative on Shaping the Future of Environment and Natural Resource Security in Partnership with PwC and the Stanford Woods Institute for the Environment. Last accessed 20 January 2020: https://www.weforum.org/projects/fourth-industrial-revolution-and-environment-the-stanford-dialogues

Wyatt, S. 2004. Danger! Metaphors at Work in Economics, Geophysiology, and the Internet. *Science, Technology & Human Values*, 29(2), 242–261.

CHAPTER 10

AI Bugs and Failures: How and Why to Render AI-Algorithms More Human?

Alkim Almila Akdag Salah

Introduction

When we look at the history of computer art, we see many instances where the artworks were created rather by 'accident' than by carefully worked out processes and deliberately written codes. For instance, one of the first computer artists, Michael Noll, acknowledged that the idea to experiment with computers to achieve artistic patterns came to him after a programming error, which resulted in an interesting graphical output (Noll 1994). The most eye-catching example of this sort comes from an entry for the exhibition *Cybernetic Serendipity*: 'A bug in the programme [sic] effectively randomized the text given to it … but we are not sure as we failed to make the programme do it again. At any rate, this "poem" is all computer's own work.' (McKinnon Wood and Masterman 1968, 54). During those early years of computer art, there was an apparent tendency to 'anthropomorphise' computers, and the split second of humanity that a bug bestows on the computer was maybe the best chance to reach that aim. A computer graphics competition held in 1968 by California Computer Products is cited to publish a statement 'that they were convinced that computer art would be accepted as a recognized art form ... because it gives a humanizing aura to machinery' (Reichardt 1971, 74).

How to cite this book chapter:
Akdag Salah, A. A. 2021. AI Bugs and Failures: How and Why to Render AI-Algorithms More Human? In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives*. Pp. 161–179. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.j.

Today, thanks to the rise of big data, computing power and mathematical advancements, and the introduction of convolutional neural networks (CNNs), we live with intelligent algorithms (i.e. weak AI), in many aspects of life.[1] For example, the effects of these algorithms in digital visual production covers recommendation systems, automatic image editing, analysing and even creating new images, but these are not recognised as 'intelligent' systems (Manovich 2020). What fascinates the human mind are still the observances of failures. A prominent example is the images and videos created with the Deep Dream algorithm, which was originally devised to unearth what lies in the hidden layers of CNNs to capture the workings and failures of the system (Simonyan and Zisserman 2014). These images are hailed by some as artworks on their own (Miller 2019).

Autonomous AI systems such as self-driving cars, or autonomous lethal weapons are expected to work in a framework called 'explainable AI', under meaningful human control, and preferably in a fail-proof way (Santoni de Sio and Van den Hoven 2018). Here, I will discuss case studies where the opposite framework will prove more beneficial, i.e., in certain contexts, such as cultural and artistic production or social robotics, AI systems might be considered more humanlike if they deliberately take on human traits: to be able to err, to bluff, to joke, to hesitate, to be whimsical, unreliable, unpredictable and above all to be creative. In order to uncover why we need 'humanlike' traits – especially bugs and failures – I will also visit the representations of the intelligent machines in the imagination of popular culture, and discuss the deeply ingrained fear of the machine as the 'other'.

The aim of the chapter is twofold: first, by reiterating the history of computer art and comparing it to how artistic production in/with AI is used and interpreted today, I pinpoint how the discourse of artistic (computational) production has changed. Second, by visiting classical definitions of AI and juxtaposing them with the public expectations and fears, I will uncover how the myths about AI are assessed when it is tasked to take on not only human jobs, but human traits as well.

In this chapter, I will build our framework around the famous discussions of the Turing test, the Chinese Room and what it means to have a computational system for creativity and arts. I will then look at the history of computer art by assessing the early artworks, exhibitions as well as magazines devoted to the genre. Especially the latter gives us insight into what the experts' expectations of computers were. I will furthermore delve into the history of sci-fi and build bridges between these early artworks and sci-fi novels and movies of the time to understand the reaction of the public to the idea of intelligent/sensuous (i.e. human like) machines. Moving to today, I will visit two artists working within the framework of AI and Big Data, who proposed two extreme approaches to this framework: Refik Anadol, who enlists Big Data and AI in a black-box fashion for generating big displays of contemporary aesthetics, and Bager Akbay,

who reveals the working of AI by generating instances of occurrences between the audience and the AI.

A Useful Framework for AI or the Ghost in the Machine?

Some sixty years ago, a part of the computer science community embarked upon an ambitious research program named 'artificial intelligence'. Summarily, the task at hand was to write an intelligent computer program; one that could simulate human thinking, and while at it, why not, properly think and even be conscious, just like a human. They had just realised that computers were able to handle arithmetical and logical operations much better than an ordinary human being, and a whole wide world of opportunities opened up before them. But after some initial effort, the researchers saw two things: the aim of implementing intelligence was an ill-posed problem, because there was no satisfactory definition of intelligence. The concept of intelligence, just like many defining characteristic of human beings, is normative and vague. The second realisation would come a little later, as it required more failures: the internal dynamics of many human endeavours were unknown, and misjudged. People thought that understanding and speaking was easy, whereas playing chess was difficult. Thus, chess was seen as a benchmark of intelligence. Years later, when a computer program, Deep Blue (Hsu 2002), was able to beat the world champion of chess, cognitive science had already contributed much to our understanding; nobody claimed that the computer was intelligent, let alone conscious.

The first and foremost effort to frame intelligence came from one of the greatest minds of the era. Alan Turing, in his classical essay 'Computing Machinery and Intelligence' proposed an imitation game, where a computer chats with a person and passes the game if it can keep up the appearance (of being human) for about five minutes (Turing 1950). Turing's expectations of the future of computers was quite close to the mark:

> The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. (Turing 1950, 442)

In 1980, John Searle wrote the most controversial critique of artificial intelligence. His famous Chinese room argument is as follows: suppose there is a room with two small slots, and a person within it. One of the slots is used to pass this person little pieces of paper with Chinese characters on it. There is also a rulebook in the room that tells the person inside what kind of symbols

he should use in order to respond to the incoming 'squiggles'. The response 'squoggles' are put through the second slot. Although the symbol exchange can be seen as a perfectly normal conversation for a person who understands Chinese, the person in the room does not know Chinese. Having consciousness and simulating one are different for Searle, and no machine will ever think, it can just simulate, or act as if it is thinking (Searle 1980, 355).

The responses to Searle are numerous (Harnad 1989), and the debate continued for about ten years. What interests me here is not the debate itself, but rather its emphasis on intentionality. For philosophers of language, what is meant by intentionality is largely an issue of how symbols can have meaning. Searle argues that a computer simulation, no matter how good it is, will never project more than its programmer's intentions, i.e. a computer program can never have intentions (and mental states), because it is written according to some syntactical rules, and it lacks the connections to semantic access. If such written programs can fool us into believing that they are intelligent beings, this does not prove that these programs are operating on the semantic level; it only shows that we are deceived by the programmer's ghost, which acts like a remote-control system through the program.

At the end, Searle's approach comes down to a simple point: an inorganic entity cannot develop intentional states and cannot become conscious. Douglas Hofstadter asserts that Searle's argument comes from a dualist point of view, which is denied fiercely by Searle in his reply to Hofstadter. I think that there is some truth in Hofstadter's claim; after all, a search for a human soul in a digital computer, even if it is run under the name as 'intentionality/consciousness' is suggestive enough for a belief in body/mind distinction. Furthermore, Daniel Dennett (1982) points out that the furious defenders of the Chinese Room argument are known dualists, and the main critics of Searle's argument are de-emphasising the importance of consciousness (even) in humans.

There is the unmistakable Cartesian ego that Dennett and many others see in Searle's argument, the ghost that Gilbert Ryle wanted to abolish, the implicit belief of the superiority of the human that is the hallmark of the modern era. In short, Turing tries to move the definition of intelligent machines outside the realm of human traits (we do not need to measure how intelligent a machine is, we only measure how well it fits within everyday relations with a human). On the other hand, Searle tries to kill the concept of intelligent machines by comparing not the 'behaviour' of these machines to human behaviour, but their 'nature' to that of human nature.

Today, AI is more and more associated with words that are reserved for humans: autonomy, learning and interpretation. For example, Haenlein and Kaplan (2019) state that AI is commonly defined as 'a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation'. Rahwan et al. (2019) called for a new science on machine behaviour as a field that 'is concerned with the scientific study of intelligent machines (i.e. virtual or embodied

AI agents), not as engineering artefacts, but as a class of actors with particular behavioural patterns and ecology'. A computer program that can learn goes beyond what it is programmed to do. The developments in the field (and the transformation of the definitions of AI) notwithstanding, the position of Turing, as well as that of Searle, are not totally overturned yet. Searle's vague definition of Strong AI (to create intentionality artificially) has led to the term 'weak AI', and both terms are still in use today (Morgan 2018). Weak AI delineates an artificial agent that succeeds to reach a goal in a given environment by computing a function from its sensory inputs to its actions. From chatbots we encounter on websites to a washing machine that can use sensors to calculate the washing load and adjust the water usage accordingly, all 'smart' applications and tools fall under weak AI. Strong AI, or artificial general intelligence (AGI), on the other hand, points out to a more human-like intelligence that is flexible enough to learn abstractions of any novel domain relatively quickly, and perform with increasing accuracy on this domain.[2]

Here I would like to draw attention to a recent discussion of strong versus weak computational creativity using AI techniques to visualise 'tracing' line drawings (Al-Rifaie and Bishop 2015). This work invites its readers to a Gedankenexperiment similar to the Chinese Room, where the program generates drawings as outcomes, and for the audience outside the Creativity Room, these drawings are genuine artistic productions. The authors' arguments are similar to those of Searle, and they conclude that it is not possible to create a creative machine in the general sense. However, they transfer the idea of strong/weak AI to machine creativity: 'An analogy could be drawn to computational creativity, extending the notion of weak AI to 'weak computational creativity', which does not go beyond exploring the simulation of human creativity; emphasising that genuine autonomy and genuine understanding are not the main issues in conceptualising weak computationally creative systems. Conversely in 'strong computational creativity', the expectation is that the machine should be autonomous, creative, have 'genuine understanding' and other cognitive states' (ibid.). This differentiation, as we will see, is helpful in assessing both computer artworks and AI-arts, as well as the intentions and goals of the programmer or the artist.

A Useful History: A Look at the First Computer Artworks

For many, the first computer art contest held by the journal *Computers & People* in 1963 marks the beginning of computer art. During the year 1965, similar exhibitions were hosted on both sides of the Atlantic. In February, the Stuttgart gallery of Wendelin Niedlich greeted the art world with the exhibition Generative Computergrafik. On display were the first computer artworks by Georg Nees. Following this exhibition, Georg Nees took part in another exhibition, Computergrafik, on 5 November, this time with his co-scientist Frieder Nake.

Their works were on display for the entire month of November. Meanwhile, Michael A. Noll and Bela Julesz from Bell Laboratories were asked to show their works on the other side of the ocean, and the first computer art exhibition of America came to life in Howard Wise's New York gallery between 6–24 April. This exhibition had a simple and descriptive name, just like its forerunner; it was called Computer Generated Pictures.[3]

The driving force behind the German wing of computer art was Max Bense, who was an influential aesthetician and a well-known figure among art circles. He was greatly interested in cybernetics and information theory, and his attempts to make use of these concepts in art theory gave the impetus and the right climate for scientists like Nees and Nake to publish their works as art. Bense's objective was to use computer generated pictures for finding 'quantitative measures of the aesthetics of objects' (Candy and Edmonds 2002, 8). Soon after the first exhibitions, Nees became a student of Bense, and wrote his dissertation entitled 'Generative Computer Graphics'. The thesis most notably included several highly principled computer programs to generate graphical output, based on Bense's ideas as outlined in his book *Aesthetica* (Nees 1969; Bense 1965).

Computer art relied on two mechanisms in the beginning: the artist's intentions, and the use of computer, respectively. During the early years of the movement, computers were found only in research centres and could not be operated by a single user, thus naturally implying coaction between the artist and the scientist. Hence, computer art, by definition, has a hybrid character, combining elements of art and technology, oscillating between working in the constraints of scientific agenda and creating art products. The investment needed to digest/understand latest scientific research and apply these ideas, methods and tools into arts asks for a new type of artist/scientist (well-versed in both discourses), or teamwork consisting of scientists and artists who have developed a hybrid language to overcome the problem of illiteracy that arises due to the lack of a common terminology. The early computer artworks were produced by scientists who were both interested in pushing the abilities of the computers, as well as searching for quantifiable means of aesthetic production and experience.

Although Georg Nees actively sought procedural descriptions for basic principles of aesthetics, most of the first artworks were created rather by 'accident' than by carefully thought processes and deliberately written codes. For instance, Frieder Nake's first art objects were the results of a test run for a newly designed drafting machine. In a similar vein, Michael Noll acknowledges that the idea to experiment with computers to achieve artistic patterns came to him after a programming error, which resulted in an interesting graphical output: 'Lines went every which way all over his plots. We joked about the abstract computer art that he had inadvertently generated' (Noll 1994, 39).

A majority of these first artworks of the movement came either from research laboratories or universities, and their accidental nature is highly relevant in

understanding the genesis of the movement. Noise, errors (or bugs as they are called in computer engineering jargon) and accidents were an unavoidable part of these experiments and Noll was by no means the only one getting excited about tracing the outcomes of such bugs to generate artworks. An 'erring' computer – if it provides an appealing output – looks like acting on free will, running against the orders by providing such an unanticipated result. The computer of course does not have a free will, and it was usually not unpredictable at all. It did not go beyond the programmed code, but it could get input from the world, and use its uncertainty to drive its own unpredictability. However, the unexpected result was mostly the outcome of an error in the code. But when the expectations of the users are not answered due to such a bug, and especially when the users are novices (like most of the computer artists were), and therefore have no clue of what the error in the code might be, the impact of the bug is unavoidable: the computer becomes human in the eyes of the novice. The perception of computers in popular culture, i.e. in the eye of the novice and the layman, is best read in the science fiction novels produced during the initial years of computer science. To see the expectations of the experts (in arts, as well as in sciences) a summary of publications on intelligent and spiritual machines suffices.

Intelligent or Spiritual Machines: Which One is More to be Feared?

Leonardo, a journal devoted to the intersection of art and sciences, published various papers on the issue of intelligent machines and their relation to art. Especially the early papers show a high degree of expectancy and a belief in the unlimited potential of computers. Michael Thompson declares in 1974 that 'in order to be of value to artists, computers must be perceptive and knowledgeable in visual matters. Being "perceptive" means that they should be programmed to deal with phenomena that artists perceive and find interesting. Being "knowledgeable" means here that the computer can use information stored in it to take appropriate courses of action' (Thompson 1974). In 1977, Michael Apter raised the stakes higher by conceptualising a computer that develops an aesthetic taste (Apter 1977). Even during the so-called AI winter, i.e. when the expectations of AI research were not met and the funding for AI plummeted, Marthur Elton claimed that we can build creative machines that will allow us to understand 'ourselves and machines' better (Elton 1995, 207). Robert Mueller elaborated that once realised, such devices will 'mark the death of the personal human imagination'. He nonetheless concluded that creative machines could pave the way to new art venues (Mueller 1990). The most preposterous claim put forward around that time came from McLaughlin's (1984) prediction that in 100 years intelligent machines will dominate the earth.

There was an opposite view against this belief in computers' abilities to develop intelligence and creative abilities in the (near) future. A great many saw computers only as 'symbol processors', i.e. machines that are a little better than calculators. Both sides had little understanding of what computers really were, and what could be expected from them. Harold Cohen, who is famously known for his artificial painter program AARON,[4] observes this diversification in the audience of his exhibitions: 'The public seemed to be divided by pretty evenly between un-sceptical believers and unbelieving sceptics. The believers were happy to believe that computers could do anything and consequently accepted the idea, with neither difficulty nor understanding that mine was doing art. The sceptics thought computers were just complicated adding machines and consequently, experienced insurmountable difficulty and equally little understanding, in believing that mine was doing what I said it was doing' (Cohen 2002, 97).

The public opinion oscillated between these two ends; both of which were equally dangerous since both were open to wild speculations or predictions about the future of computers and their role in the society. The science fiction novels of those times are full of telling examples about this oscillation. In many science fiction novels, the computer is depicted either as a giant machine controlling the human society, in a sense replacing the government (this is the exaggerated version of the belief that computers were symbol processors) or as a substitute for a human, where the computer or the robot takes on specific roles like the teacher, police, surgeon, adviser, etc. (following on the belief that computers could do anything). In both instances, the computer is portrayed as superior to humans, and the only way to make humans triumph over computers is to overemphasise certain human traits.

For example, in 'The God Machine' (Caiden 1989), the supercomputer collapses because it cannot bluff as the human opposing it; in both 'Variable Man' (Dick 1957) and 'Fool's Mate' (Sheckley 2009) the computer is defeated because it cannot predict human actions; or in 'The Moon is a Harsh Mistress' (Heinlein 1966) the super machine cannot understand why a joke is funny. The SF literature is full of these examples, but the one example that is most relevant to the topic of the present work is the one where the computer (or the robot) develops beyond being a calculating machine and gains a very peculiar human ability: creativity.

One of Asimov's best stories, 'The Bicentennial Man' is based on this idea (Asimov 2000). The hero of the story is one of the earlier robots crafted for general usage and sold into a most wealthy household. When it develops the ability to make art-pieces, the producing company acknowledges this as a defect, and offers to replace the robot. Its owner decides to keep this peculiar robot, and gives it the privilege to earn money through selling its works. As the story unfolds, the robot becomes more and more human and demands to have more rights; first it fights for its freedom, then it asks to be called a human. However, the price for humanity is very high. It is not enough to be creative, to

have the wish and need for freedom, or the longing for humanity; it is not even enough to look and act like a human. The price for humanity is the humble attitude of giving up all the superior abilities; and in this case, the most superior (and dangerous) faculty of the robot is its immortality. Thus, in order to become human, it has to accept death.

With every attempt to move computers into the territory of human intelligence, the definition of intelligence or the understanding of human abilities changes. In the literature of AI history, this dilemma is called as the 'AI Effect' (Haenlein and Kaplan 2019). However, since the source of the problem is rather in the disinclination of humans to accept the capabilities of computers in taking on faculties that are attributable to humans only, even the Turing test which was devised to avoid this problem, cannot offer a tangible solution. At the bottom of this disinclination lies the narration of humans as superior beings in the universe. This belief, which has religious roots, shapes the world view of its adherents in such a way that there is no place for computers beating down humans in logical operations, let alone in more delicate traits like writing poetry, or making art. We should not forget that the definition and understanding of human intelligence has been shaped by AI research as well, and there are ample examples in science fiction to accommodate the discourse around posthumanism (Hayles 2008) and transhumanism movements.

A Provocative Experiment

The fear of computers, the fear of intelligence in an 'other' that is capable of thinking and creating, played a role in forming a certain reluctance to associate any kind of art with computers in the public mind. Obviously, there were other problems, most importantly the fact that a normative definition of art involves the intentions of the artist. Computer art as a genre followed this normative definition, and put emphasis on the intentions of the artist/scientists. Therefore, intentionality, or the lack thereof, was a much more relevant issue.

During 1960s, the art world was discussing the relevance of chance occurrences in creating art. The idea of randomness as opposed to intentions has surfaced now and again throughout many 20th century art movements and made quite an impact. Surrealists tried to let their subconsciousness take over by giving up their self-control over their minds. A similar approach – albeit with different reasons and results – was followed by Dada artists during the 1920s. During the late 1960s randomness and chance were important factors in the artworks of major figures like John Cage. As the computers entered the stage, they offered an easier way to explore these chance occurrences. In an article published in 1968 about art and technology, Douglas Davies particularly emphasises the effect of chance and its role in the history of aesthetics, as well as its immediate relation with technology and control (Davies 1968).

Random occurrences, or the ability to create randomness in an artwork was seen as one of the advantages offered by the computer. This is one of the most debated topics of early computer art. Reichard, as many others, refutes the idea of putting computer-generated randomness on a par with the chance occurrences sought by action painters like Alan Davie. As an example, Reichardt refers to one of Alan Davie's paintings where the words cat and mouse are added to the painting, because a cat entered the room and walked over the painting while Davie was working on the painting by pouring paint onto the canvas. Reichardt is of the idea that such an occurrence cannot be duplicated or mimicked by computers (Reichardt 1971). However, according to Max Bense's theory of Generative Aesthetics, 'randomness involved in computer graphics replaces that aspect in art which is described as intuitive' through computer procedures. 'Thus the randomizing procedures in computer technology are analogous to an artist's intuition' (Davies 1968, 8).

When Georg Nees displayed his randomly distributed geometric shapes in an art context, the reaction of the art community in Stuttgart was quite fierce: 'Some of them (the artists) became nervous, hostile, furious. Some left. If the pictures were done by use of a computer, how could they possibly be art? The idea was ridiculous! Where was the inspiration, the intuition, the creative act? What the heck could be the message of these pictures? They were nothing but black straight lines on white paper, combined into simple geometric shapes. Variations, combinatorics, randomness … but even randomness, the artists learned, was not really random but only calculated pseudo-randomness, the type of randomness possible on a digital computer. A fake, from start to end, christened as art!' (Nake cited in Candy and Edmonds 2002, 6). Many artists were not ready to accept a randomness created by computers for real. Actually, the fact that these works were showcased as art was not as puzzling and disconcerting as the realisation that the same works could really be passed as made by human hands.

In his book *AI Aesthetics* (2020), Manovich compares early computer art with AI-arts, and makes an interesting observation. The early computer artworks are abstract in nature, not related to human affairs except the concern on aesthetics, whereas today, we see more and more works that mimic many layers of arts. Manovich furthermore proposes three ways to define AI art. The first proposition comes from designing a 'Turing AI arts' test, the other two definitions are asked to not only mimic existing art in a convincing way, but to go beyond the cultural production of today, and generate truly innovative products. The definitions here differ only how they achieve these innovations. But if we return to a Turing AI-test that follows the conditions 'if art historians mistake objects a computer creates after training for the original artifacts from some period, and if these objects are not simply slightly modified copies of existing artifacts, such computer passed "Turing AI arts" test … In this definition, art created by an AI is something that professionals recognize as valid historical art or contemporary art.'

Actually, history of computer art already witnessed such a Turing AI arts test. The artworks generated by the computer back then were not 'sufficiently' different than original artworks, but considering the time frame when the experiments were run, this might be excused. Michael Noll wrote a program to simulate Mondrian's 'Composition with Lines', which is a black–white composition. The end result 'Computer Composition with Lines' was quite similar to the original, and Noll used it in an experiment performed with 100 subjects. In the experiment, the subjects were shown Mondrian's and Noll's compositions, and subsequently were asked to about the authenticity of the pictures shown; the subjects had to identify Mondrian's picture, and they were asked to give explanations on why they chose one picture over the other. The third page was called the 'preference' questionnaire, asked the subjects which picture they liked more, encouraging them to give specific reasons for the preference. Noll published the results as an article in *The Psychological Record* in 1966, and this article is reprinted in various edited collections since then (Noll 1966).

In his paper, Noll carefully noted his methods for designing the experiment to explicate how a control group was formed to see whether his subjects were prejudiced against computer art. To prevent any such prejudice, the control group was first asked to choose a painting, and only then expected to identify the Mondrian painting. The statistical analysis showed that the order of questions did not have any significant bearing on the preference of the subjects. The results were quite 'thought-provoking' as Noll noted in his paper. Of the subjects, 59% preferred the computer-generated image. Moreover, only 28% of the subjects were able to identify Mondrian's painting correctly, and most of them had a 'technical' background. Noll's explanation for the higher correct identification rate by subjects with technical backgrounds was that they were familiar with computer programming and had an advantage at guessing which picture was generated by a computer. On the other hand, Noll was also convinced that Mondrian's painting was carefully planned and conducted according to an algorithm, which he himself was unable to discover. In comparison with this calculated painting, Noll's computer design struck the eye as being more 'random'. Consequently, a higher percentage of the subjects preferred the computer-generated image. Noll concluded that randomness creates a feeling of creativity, and especially for the non-technical subjects that was equivalent to an indication that a human crafted the painting.

Noll's explanation to the rather astonishing preference for the computer-generated image was that all the subjects were familiar with computer technology and were using it in their everyday lives.[5] The subjects were recruited from his own work environment, and thus represent a biased sample. It was quite unnatural in 1965 to have so many subjects familiar with computers, as the majority of the population had not even encountered a computer in their lives. According to Noll, the subjects did not have any prejudice against using computers for creating art as a result of this familiarity. He further commented that the results may have been quite different if the subjects had been coming

from an artistic background, anticipating a negative reaction against computer usage in arts. Although later he also did an experiment on subjects with artistic backgrounds, the experimental setup was quite different, and does not lead to a direct comparison (Noll 1972).

Noll's experiment is particularly relevant, because it demonstrated that a computer-generated picture could be confused with or even preferred over a human-generated artwork. Through the experiment, he challenges the possibility of reading the artistic intentions behind an artwork (and questions one of the basic assumptions of art history as a discipline): 'The experiment compared the results of an intellectual, non-emotional endeavour involving a computer with the pattern produced by a painter whose work has been characterised as expressing the emotions and mysticism of its author. The results of this experiment would seem to raise some doubts about the importance of the artist's milieu and emotional behaviour in communicating through the art object' (Noll 1966, 10). If the artwork cannot mirror its creators' intentions, thoughts and ideas, can we still claim that the artwork reflects its era, more than any everyday object?

Where Are We Now: Computer Art, Aesthetics and AI Art

Within computer vision and multimedia retrieval, computer-based analysis of artworks has received increasing attention in the last two decades (Spratt and Elgammal 2014). The research focused on creating automatic programs that, given an artwork, can identify the artist, the style or the production date, as well as search for stylistically similar artworks in a collection (Stork 2009). While some of this research followed reductionist perspectives and was heavily criticized for losing sight of critical content, the fact remains that computer vision provided art historians with tools that can be used in locating visual materials with certain aspects successfully. For instance, Crowley and Zisserman's retrieval system allowed one to search for simple concepts (e.g. 'train', 'dog', 'flower', 'bridge') in painting databases, without requiring annotations for these concepts. It works by collecting keyword-indexed images from the internet and learning from them the appearance of the concept on the fly (2014). It became possible to retrieve and visualise paintings of a particular period that show a certain visual quality, or contain a certain object or feature. With the introduction of style transfer algorithms (Gatys, Ecker and Bethge 2016; Sanakoyeu et al. 2018), one more step was taken: the content of a picture could be separated from the style of the painting.

All these steps paved the way to AI algorithms contributing more and more to todays' aesthetic and artistic production and appreciation. We use various applications for getting recommendations to the artworks we like, for automatically 'beautifying' photographs we take, or for assessing aesthetically pleasing

photographs with an explanation of the reason behind the assessment, or for designing our PowerPoint slides and even for automatic creation of short videos from our photographs or videos. The way from the early computer artworks to today was a long and winding one. We have seen that within computer art, aesthetics was an important research venue. The combinatorial possibilities offered by the computer lead the artists to create variations of simple geometric patterns, and many possible combinations of a single composition, from which the most aesthetically pleasing ones could be selected. With this approach, philosophers like Max Bense (Bense 1965) and Abraham Moles (Moles 1966) pioneered the search for mathematical rules governing aesthetics, and their theories were influential. Today, the aesthetical assessment of images are done thanks to the help of image archives that are used in the supervised training of machine learning algorithms; in its essence this means that the work of the programmer has changed tremendously. Instead of 'programming/coding' rules about aesthetics, the current algorithms are programmed to discover statistical patterns in huge image datasets, where the algorithms 'learn' by comparing images to each other and guessing the correct answer to a question on aesthetics (e.g. which image is more liked by people) or content (e.g. if the image has a bird in it) by minimising an error function. The programmer does not supervise the learning progress, instead, she provides the algorithm with information about the image data sets. The success of these statistical algorithms is simple to assess, they are all widely used.

If we compare distinct artworks from the earlier era of computer art to AI-art of today we might capture the transformation more clearly. Harold Cohen is one of the most widely recognised electronic artists, and he let AARON evolve through more than 25 years (from 1973 to early 2000s) to its present state of maturity. In his words, AARON was originally 'a program designed to investigate the cognitive principles underlying visual representation' (Cohen 1988, 846). In 25 years of its artificial life, AARON 'learned' to draw, like a child's first scribbles slowly transforming into a modernist painter's stylistic abstractions. The processes developed by Cohen for AARON to create its paintings can be inspected to discover patterns and clues about 'creativity', but not everyone who watches AARON paint will find sufficient evidence to call it 'creative', nor did Cohen ever claim that AARON is creative. There have been debates about the definition of creativity, and whether it is possible to concede that an artificial intelligence (AI) program can be creative like a painter, or not. After all, if there are rules or a procedural description for the artistic activity, then there is no reason why a computer program cannot be written to produce art. An important issue here is that many humans contribute to the production of a final artwork, and the AI algorithm is not an encapsulated unit, yet the language used in their description (e.g. thought vectors, consciousness priors, attention) anthropomorphises the algorithms and creates a conceptual problem (Epstein et al. 2020).

Cohen's AARON is an early example of AI-based artwork creation, I will visit two more recent examples, representing two ends of a spectrum (low-budget to high-budget). Bager Akbay's recent AI-artworks of *Deniz Yilmaz, The Robot Poet*,[6] took a different approach to the problem at hand, and bypassed the definitions of creativity, as well as questions on whether autonomous computational creativity is possible or not. What Akbay proposed is to generate a 'learning' poet, just like AARON the painter. However, unlike AARON, the underlying program of Deniz Yilmaz is based on the processing of a big data set of published poems in a literary magazine, *Posta Gazetesi*. Akbay expected the outcomes of this algorithm to be similar in nature to the dataset, generating 'average' poems. Similarly, while crafting Deniz Yilmaz's identity (which has its own Facebook page, and now in search for its citizenship), he used the photographs that were published along the poems in Posta Gazetesi, and generated an average photograph for Deniz. Today, Deniz Yilmaz has multiple exhibitions (just like AARON) and a book publication. Like AARON, for which Cohen had designed various printing and painting devices, Akbay designed a handwriting style and ways of 'writing' poems for Deniz Yilmaz. But the similarity between AARON and Deniz Yilmaz ends here. Whereas Cohen's ambition was to find ways to write a code that learns aesthetic principles, and a way to develop itself, Akbay's focus was exploring robot rights and leading conceptual discussions around the entity of Deniz Yilmaz, asking: 'Can a robot poet be considered autonomous, and if so, what are the mechanisms to enable this transition?' To that effect, Akbay wrote another algorithm, a manager for Deniz Yilmaz's artistic endeavours, which invited various people to assemble a board of directors to manage Deniz Yilmaz's dealings within the art world. Akbay's ambition is for Deniz to have a life of its own, where its earnings will be transferred to a bank account bearing its own name. He refers to Deniz Yilmaz as a failed experiment, as his name as the creator of the robot poet still is on the foreground.

Refik Anadol, who uses big data, as well as big displays to showcase AI-artworks, is a well-known name for his (and his team's) unique approach to different data sources and the way he transforms these into new imaginations. Anadol explains that he views 'machine intelligence not only as a new medium, but as a collaborator, allowing us to re-examine not merely our external realities, but rather an alternative process to which we attribute artistic consciousness' (Anadol 2019). For example, for *Latent History*,[7] a recent work on the history of Stockholm, he used archival data consisting of the city's photographs from the last 150 years combined with current photographs. He maintains that the classical approaches to displaying such a plethora of data fails short, whereas machine learning generates 'a time and space exploration into Stockholm's past and ultimately present … on a multidimensional scale' (ibid.). These kinds of explorations make the audience enter a new type of reality along with a new type of aesthetics. Still, with every new artwork, even though Anadol claims to use machine learning as a collaborator, the artist's decisions on what type of

data to use for which purposes, and how, renders the results as artworks, not the other way around.

Whereas Cohen's, Anadol's and Akbay's artworks carry the stigmata of their creators, Deep Dream animations generated by style transfer technology have a different position. First of all, like the first computer artists, the creators of Deep Dream were scientists, and initially, they were not after designing AI algorithms to create artworks. As a similar story unfolds, they found the results of their algorithms were beyond their expectations, and worthy of investigation: 'In the summer of 2015, we also began to see some surprising experiments hinting at the creative and artistic possibilities latent in these models' (Agüera y Arcas 2017). As the resulting 'artworks' (see Mordvintsev, Olah and Tyka 2020 for examples) are quite different then what is ever hailed as art, the audience was both fascinated, and sceptical. From computer art to AI-arts we are still within the realm of weak AI, i.e., AI algorithms are used as tools to create artworks, and we still see the oscillation between 'un-sceptical believers and unbelieving sceptics', and the impact of the Deep Dream comes from the unexpected results it generated, which bestows an autonomous position to the algorithm.

By Way of Concluding

Science fiction takes the idea of computers with human capabilities to its extremes. However, Jameson points out that science fiction genre is akin to utopias, which actually never attempt 'to represent or imagine a real future but rather to denounce our inability to conceive one, the poverty of our imaginations, the structural impossibility of our being able to generate a concrete vision of a reality that is radically different from our current society' (Jameson 1982). Lacking the means to open new ways to future realities, science fiction rather takes on the role to unveil 'a particular historical present' (Thacker 2001, 156).

When we look at contemporary science fiction productions, especially the revisits to old TV series or films such as 'Battlestar Galactica' or 'Westworld', we see that the fear of the strong AI that looks and acts just like humans – but stronger and smarter in nature – remains unchanged. When it comes to intelligent machines, the particular historical presents do not change, even though the technology has developed considerably in the past decades, and we live in a world where weak AI has permeated into the daily life. As a marketing policy, tools and algorithms we use are not necessarily tagged as AI, and this might have helped their dissemination without any resistance by the public (Tascarella 2020).

For the interests of what I have presented in this chapter, i.e. the assessment of artworks generated with/by computers and/or computational creativity, we see the same trend: the audiences, as well as the creators still prefer to be amazed by the unexpected results of human–machine collaborations (like in

Deep Dream), and are not so much interested in bestowing an autonomy and creative status on the programs themselves (like in the case of Deniz Yilmaz), but they are much more open to new imaginations that are created with AI (like in the case of Latent History). Throughout the chapter, I have referred to computer art and AI-arts as two separate art movements. However, when we follow the broad definition of computer art, i.e. artworks generated by artists/scientists with the aim of challenging the boundaries of arts as well as sciences, we see that AI-arts still fall under the umbrella of the latter. Of course, a lot has changed since the first computer artworks, in intention, goals and challenges. More importantly, today, AI-arts offer a platform where computers could become more than tools, and collaborators, and maybe in the future, sole artists. As Mark Weiser (1995) in his now famous *Scientific American* article noted, 'the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it', and the weak AI creativity already disappeared. The question remains when the strong computational creativity will achieve the same status, and how it will be hailed when it does.

Notes

[1] In 2003 only, Menzies (2003) listed topics where AI-research was successfully implemented when he described the road ahead. He emphasised the ubiquity of the tools AI researchers could use and combine. Today, the tools mentioned in that list are already operational in daily life. These tools are considered successful as weak AI examples: they operate fairly well on the tasks they are designed to accomplish, but they do lack a general intelligence. They cannot do anything but what they are designed to do.

[2] For a historical overview of AI, please see Russell and Norvig (2002).

[3] For a detailed history of computer art see Franke (1971) and Noll (1994). This section summarises research on technoscience art (Akdag Salah 2008).

[4] To see various artworks generated by AARON, please visit: http://www.aaronshome.com/aaron/index.html.

[5] To see the Mondrian and Noll's artworks used in the experiment, please visit: http://dada.compart-bremen.de/item/artwork/5.

[6] To see the list of exhibitions of the robot poet Deniz Yilmaz, please visit: https://www.poetryinternational.org/pi/poet/29478/Deniz-Yilmaz/en/tile. You can access Yilmaz's published book from here as well: https://drive.google.com/drive/folders/0B6I_wTbmgoBMeUp2OExhQVVUWEU. The poems are generated in Turkish.

[7] To see a sample of Latent History, please visit: https://www.fotografiska.com/sto/en/news/refik-anadol-latent-history

References

Agüera y Arcas, B. 2017. Art in the Age of Machine Intelligence. *Arts* 6(4), 18. Multidisciplinary Digital Publishing Institute.

Akdag Salah, A.A. 2008. *Discontents of Computer Art: A Discourse Analysis on the Intersection of Arts, Sciences and Technology*. PhD Diss., University of California, Los Angeles.

Al-Rifaie, M.M. and Bishop, M. 2015. Weak and Strong Computational Creativity. In: Besold, T. R., Schorlemmer, M. and Smaill, A. (eds.) *Computational Creativity Research: Towards Creative Machines,* pp. 37–49. Paris: Atlantis Press.

Anadol, R. 2019. Latent History. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1138–1138.

Apter, M. J. 1977. Can Computers Be Programmed to Appreciate Art? *Leonardo* 10, 17–21.

Asimov, I. 2000. *The Bicentennial Man*. New edition. London: Gollancz.

Bense, M. 1965. *Aesthetica: Einführung in die Neue Aesthetik*. Baden-Baden: Agis-Verlag.

Berkeley, E. C. 1974. Can Tigger think? Can Peder think? *Computers & People* September, 8.

Caiden, M. 1989. *The God Machine*. New York: Baen Books.

Candy, L. and E. Edmonds. 2002. Introduction. In L. Candy and E. Edmonds (Eds.), *Explorations in Art and Technology*, pp. 5–21. New York: Springer.

Cohen, H. 1988. How to Draw Three People in a Botanical Garden? *Proceedings of the Seventh National Conference on Artificial Intelligence*, 846–855.

Cohen, H. 2002. A Million Millennium Medici's. In: L. Candy and E. Edmonds (Eds.), *Explorations in Art and Technology*, pp. 91–105. New York: Springer.

Crowley, E. J. and Zisserman, A. 2014. In Search of Art. In *European Conference on Computer Vision*, pp. 54–70. Cham: Springer.

Davies, D. M. 1968. Art and Technology – Toward Play. *Art Journal* 56, 46–48.

Dennett, D. C. and Searle, J. 1982. The Myth of the Computer: An Exchange. *NY Review Books* 29(11), 56.

Dick, P. 1957. *The Variable Men*. New York: Ace.

Elton, M. 1995. Artificial Creativity: Enculturing Computers. *Leonardo* 28(3), 207–213.

Epstein, Z., Levine, S. Rand, D. G. and Rahwan, I. 2020. Who Gets Credit for AI-Generated Art? *iScience* 23(9), 101515.

Franke, H. W. 1971. Computers and Visual Art. *Leonardo* 4(4), 331–338.

Gatys, L. A., Ecker, A. S. and Bethge, M. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Haenlein, M. and Kaplan, A. 2019. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review* 61(4), 5–14.

Harnad, S. 1989. Minds, Machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence* 1(1), 5–25.
Hayles, N. K. 2008. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, IL: University of Chicago Press.
Heinlein, R. 1966. *The Moon was a Harsh Mistress*. New York: G. P. Putnam's Sons.
Hsu, F.H. 2002. *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton, NJ: Princeton University Press.
Jameson, F. 1982. Progress Versus Utopia; or, Can We Imagine the Future? *Science Fiction Studies*, 9(2)147–158.
Manovich, L. 2020. *AI Aesthetics*. Moscow: Strelka Press.
McKinnon Wood, R. and Masterman, M. 1968. Computer Poetry From CLRU. In Reichardt, J. (Ed.), *Cybernetic Serendipity: The Computer and the Arts*, pp. 55–56. London: Studio International.
McLaughlin, W. I. 1984. Human Evolution in the Age of the Intelligent Machine. *Leonardo*, 17(4), 277–287.
Menzies, T. 2003. 21st-Century AI: Proud, Not Smug. *IEEE Intelligent Systems* 18(3), 18–24.
Miller, A. I. 2019. *The Artist in the Machine: The World of AI-Powered Creativity*. Cambridge, MA: MIT Press.
Moles, A. A. 1966. *Information Theory and Esthetic Perception*. Urbana: IL: University of Illinois Press.
Mordvintsev, A., Olah, C. and Tyka, M. 2020. Inceptionism: Going Deeper into Neural Networks. Google Research Blog. 17 June. Last accessed on 20 June 2020, https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Morgan, J. 2018. Yesterday's Tomorrow Today: Turing, Searle and the Contested Significance of Artificial Intelligence. In: I. Al-Amoudi and J. Morgan (Eds.), *Realist Responses to Post-Human Society: Ex Machina*. Abingdon: Routledge.
Mueller, R. E. 1990. The Leonardo Paradox: Imagining the Ultimately Creative Computer. *Leonardo*, 23(4),427–430.
Nees, G. 1969. *Generative Computergraphik*. Berlin: Siemens.
Noll, M. A. 1966. Human or Machine: A Subjective Comparison of Piet Mondrian's 'Composition With Lines' (1917) and a Computer-Generated Picture. *The Psychological Record* 16(1), 1–10.
Noll, M. A. 1972. The Effects of Artistic Training on Aesthetic Preferences for Pseudo-Random Computer-Generated Patterns. *The Psychological Record* 22(4), 449–462.
Noll, M. A. 1994. The Beginnings of Computer Art in the United States: A Memoir. *Leonardo*, 39–44.
Rahwan, I. et al. 2019. Machine Behaviour. *Nature* 568(7753), 477–486. DOI: https://doi.org/10.1038/s41586-019-1138-y
Reichardt, J. 1971. *The Computer in Art*. London: Studio Vista.

Russell, S. and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice Hall.

Sanakoyeu, A, Kotovenko, D., Lang, S. and Ommer, B. 2018. A Style-Aware Content Loss for Real-Time HD Style Transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 698–714.

Santoni de Sio, F. and Van den Hoven, J. 2018. Meaningful Human Control Over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, 5, 15.

Searle, J. R. 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences* 3(3), 417–457.

Sheckley, R. 2009. *Fool's Mate*. Kindle Edition.

Simonyan, K. and Zisserman, A. 2014. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556.

Spratt, E. L. and Elgammal, A. 2014. Computational Beauty: Aesthetic Judgment at the Intersection of Art and Science. In European Conference on Computer Vision, 35–53. Cham: Springer.

Stork, D. G. 2009. Computer Vision and Computer Graphics Analysis of Paintings and Drawings: An Introduction to the Literature. In: *International Conference on Computer Analysis of Images and Patterns*, pp. 9–24. Berlin, Heidelberg: Springer.

Tascarella, P. 2020. Robotics Firms Find Fundraising Struggle, with Venture Capital Shy. *Pittsburgh Business Times*. Last accessed 20 June 2020: https://www.bizjournals.com/pittsburgh/stories/2006/08/14/focus3.html

Thacker, E. 2001. The Science Fiction of Technoscience: The Politics of Simulation and a Challenge for New Media Art. *Leonardo* 34(2), 155–158.

Thompson, M. 1974. Intelligent Computers and Visual Artists. *Leonardo* 7(3), 227–234.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236), 433–460.

Weiser, M. 1995. The Computer for the 21st Century. *Scientific American* 272(3), 78–89.

PART 3

AI Power and Inequalities

CHAPTER 11

Primed Prediction: A Critical Examination of the Consequences of Exclusion of the Ontological Now in AI Protocol

Carrie O'Connell and Chad Van de Wiele

The dominance of the machine presupposes a society in the last stages of increasing entropy, where probability is negligible and where the statistical differences among individuals are nil. Fortunately we have not yet reached such a state.

– Norbert Wiener (1989, 181)

Introduction

Norbert Wiener (1989) concludes his seminal work, *The Human Use of Human Beings: Cybernetics and Society*, with a warning. The thermodynamic universe, as he envisioned it, was evolving towards an entropic fate, as natural systems do. As entropy and progress are at odds, and ever the champion of purposive progress, Wiener applies the Darwinian principle of natural selection as a guide for a progressive *cybernetic* future. Wiener's concept of *negentropy*, or the mitigation of such natural entropic determination (Faucher 2013), is premised on the optimism that tailored feedback within cybernetic systems could teach

How to cite this book chapter:
O'Connell, C., and Van de Wiele, C. 2021. Primed Prediction: A Critical Examination of the Consequences of Exclusion of the Ontological Now in AI Protocol. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 183–201. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.k.

machines to redirect course towards more organised and error-reduced (rather than error-free) outcomes. The point of such tailoring – in the sense that it serves as a blueprint for algorithmic prediction – is to model possibilities of human behaviour relating to the socio-cultural. However, at what point does the simulation of human behaviour become just a more consumable way of saying, 'shaping behaviour through technology'?

The primary purpose of this chapter is to explore the shortcomings of modern-day applications of Wiener's cybernetic prediction – the theoretical foundation of artificial intelligence (AI) – particularly in terms of capture technologies that remain ubiquitous as a method of data collection for feeding such systems. We argue that such data are not impartial or necessarily explanatory, but rather evidence of third-order simulacra, *simulation*, as conceptualised by Jean Baudrillard (1994). We examine what cybernetic prediction, as outlined by Wiener, excludes; namely, an attendance to the complex ontological now, which Baudrillard warned against in his analysis of the order of simulacra – particularly the role technological innovations play in untethering reality from the material plane, leading to a crisis of simulation of experience. Secondly, we explore the potential psychosocial consequences associated with machine learning systems predicated on a cybernetic theorem that foundationally relies on human repetition – specifically, that reliance upon such repetition leads to the very entropy that Wiener warned against. As Mumford (1972) notes in his essay, *Technics and the Nature of Man*, human nature may be subsumed, 'if not suppressed' (77), by the technological organization of intelligence into technological systems. From this perspective, any machine learning system rooted in Wiener's view of cybernetic feedback loops risks creating outcomes through a process of subjective priming, more so than predicting it.

The Genesis of Cybernetics

Emerging mid-century, and inspired – in part – by the technological advancement of both machinery and intelligence-gathering systems that emerged during WWII, Wiener's theory of cybernetics focuses on the diffusion of communication in terms of control imposed by constraints and allowances afforded by the networks through which messages spread. Inspired by the 17th-century philosopher and progenitor of *modal metaphysics*, Gottfried Leibniz – in part because of Leibniz's explication of language as a computational system, and in part due to Leibniz's fascination with the potential of automata – Wiener envisioned a system of feedback in which man and machine are indistinguishable when considering message input and output. Like living organisms which have 'a tendency to follow the patterns of their ancestors' (Wiener 1989, 27), cybernetic systems, too, in their ability to be shaped by external stimuli, can leverage feedback as a 'method of controlling a system by reinserting into it the results of its past performance' (61). The past, in other words, can inform and correct future outcomes.

To Wiener, the goal of understanding communication feedback as a computational system wasn't simply to reflect upon the human condition, but leverage that reflection as a tool of prediction for future events. The 'divine intermediary' in Wiener's calculation wasn't a Leibnizian pre-established harmony ordained by God or natural law – these were the prescriptions that lead to the entropy he warned against. Instead, the intermediary would take human form, ordained by a prescription of diverse input from not just the scientist, but also the 'philosopher and anthropologist'. From a 21st-century perspective, with decades of applied cybernetic prediction as evidence, it is necessary to wonder, however, if the very entropy Wiener warned against has come to fruition via the exclusion of input variety by those who design, operationalise, and ultimately capitalise on predictive technologies which surveil, capture, and predict human behaviour and events.

One focus of contemporary concern regarding the application of cybernetics is that, as the theoretical foundation for AI, its principles are often applied beyond the 'negligibly small' domain of truly closed systems. As Faucher (2013) argues, 'The utility of cybernetics is confined to very local and specific contexts, and in a universe of increasing complexity, cybernetics will not necessarily save us' (206). Yet, today, cybernetic principles undergird algorithms designed to predict everything from global economies to recidivism in the arena of criminal justice. The question, however, is whether such cybernetic-based systems objectively reflect potential probability in an effort to prune towards progress, or 'play an active role in steering the likelihood of an event' (Faucher 2013, 211), thereby priming behaviour, both machine and human, towards future outcomes.

The Algorithm: Third-Order Simulacra

Wiener analogizes machine learning to the neurological process of receiving input, stimulating synaptic flares, recording memory (or, *taping*), and ultimately evolving future responses to stimuli. To explain this *taping* mechanism 'which determines the sequence of operations to be performed', (65) he refers to the recreation of this physiological function in digital form as the 'mechanical simulacra of the brain', (65). This function of the human brain provides an apt blueprint for Wiener's vision of machine learning on two fronts. On the one hand, the analogy provides an elegant heuristic for understanding the learning process in easily accessible terms. On the other, it reminds us that the neurological process of recording memory is hidden from plain sight – shrouded by a vessel of skin and bones, nerves and blood flow. In mechanical terms, this shrouding, or 'black-boxing', is done via bits and code. Cautious of the all-or-nothing binary that might be gleaned from this analogy for learning, Wiener recognised that we must treat the human subject as a cultural creation, not just an agent of neurotransmission that records memory, or data, to be analysed. However, as machine learning has advanced, the genesis of such cultural

creation is called into question. As Martin Hand (2014) notes, 'Algorithmically produced data now accesses us, intervening and mediating nearly all aspects of everyday life whether we know it (like it) or not' (8). Thus, a new social ontology has emerged, consistent with Baudrillard's definition of third-order simulacra: We exist in a 'dataverse' in which the world is literally made of data – so too, our cultural knowledge.

As Baudrillard (1994) describes, there are three orders of simulacra – which can be historically paralleled against the epochal transitions from the pre-Industrial, to Industrial, to Digital eras, and manifold scientific advances which anchored each. The first order comprises those simulacra which are naturalist, counterfeit images of reality that still 'aim for the restitution or the ideal institution of nature made in God's image' (Baudrillard 1994, 121). In the pre-Industrial Age, the nature of being, as inspired by God, defined natural reality and universal truth. The second order of simulacra are materialised as products, made possible by the advanced machinations of the Industrial Age, which generated the expansion of globalisation. Scientific invention, as materialised by the machine, prominently figured as a technical form of magic in the scientific imagination during the Industrial era. The imbued power of God which had defined ontological reality in the pre-Industrial mind was now replaced with the power of science fiction premised on a future made possible by the Promethean power of industrial technologies. The third order, and arguably the most confounding, are the simulacra of *simulation* – that which is 'founded on information, the model, the cybernetic game' (121), and whose aim is total operational control.

Fundamentally, Baudrillard's explication of the order of simulacra is a quest for the provenance of ontological truth. Due to the emergence of the technological 'other' in the form of *simulation*, we are on the precipice of a cultural hiatus, distortion, or rift of ontology. Today, 'truth' has been subsumed into a self-referential system of binary code by those who seek to operationalise, predict, and ultimately control human behaviour. Such cybernetic 'truth' is not inspired by nature (vis-à-vis 'God'), or the Modern principles of human imagination that provoked scientific inquiry, but feedback loops that selectively include and exclude data input for reasons obscured or 'black-boxed' from the end-user. The power of God that defined ontological reality in the first order of simulacra, as well as the power of scientific imagination that defined ontology in the second, has now been firmly replaced by a new mode of instantiation – the *algorithm*.

An investigation of Baudrillard's concept of *simulation* to explore the power imbalance created by modern technology is not without precedent. In her book, *Paper Knowledge: Towards a Media History of Documents*, Lisa Gitelman (2014) examines the troubled ethos behind digital simulation – the site of the disappearance of meaning and tangible representation. Similarly, Castillo and Egginton (2017) argue that, in the digital era, what is 'real' and what is a constructed 'copy' has become increasingly difficult for the human user to

distinguish due to the black-boxed, bits-based nature of production in a cybernetic world. Similarly, in his analysis of the legacy of the Automaton Turk on current perceptions of AI, Ashford (2017) notes that machines are capable of 'projecting illusions that can undermine our very ontologies' (139), and suggests that computational technology might soon eclipse human agency in shaping history. Uricchio (2017) echoes this concern in his analysis that what defines subject (human) and object (technological artefact) has been confounded by modern-day algorithmic intermediaries that are capable of self-learning. In other words, in the 21st century, as machine learning evolves, authorship of – not just output, but the system itself – has been taken from the hands of humans who have become passive contributors of data. Soon, algorithms will know so much about our behaviour that such agency will no longer be foundational to the cybernetic relationship between a technical system and human interlocutor.

In many respects, Wiener envisioned this algorithmic future. Fascinated by the idea that black boxes, or those cybernetic units 'designed to perform a function before one knew how it functioned' (Galison 1994, 246), Wiener – in the philosophical vein of Descartes – thought it possible to create hardware that replicated the function of the human brain. As Jeffrey Sconce points out in *The Technical Delusion*, Wiener himself envisioned a 'brain-in-a-jar' form of cybernetics: 'Theoretically, if we could build a machine whose mechanical structures duplicated human physiology, then we could have a machine whose intellectual capacities would duplicate those of human beings' (Wiener as cited in Sconce 2019, 234). Yet, as Galison (1994) notes, critics of Wiener's black box project saw the potential for 'the elimination of inner states of human intention, desire, pleasure, and pain in favour of purely observable manifestations' (252). At the heart of cybernetic prediction is the belief that to understand human beings, it is first essential to understand how patterns of information are created, stored, retrieved, and organised (Hayles 2008). However, such cybernetic prediction is a narrow, self-referential system focused on the past and future in which information input plays a privileged role in hiding 'the real behind a veil of digital representations designed to take command of life itself' (Faucher 2013, 211). And, as Hand explains, 'The dataverse promises a new descriptive-predictive analytics of pattern and correlation, prioritized over meaning and causation' (2014, 10). That is, rather than producing meaning, algorithms – black boxes that house and take as input information that becomes simulatory – merely produce more information.

Critically, this process and the technical systems that facilitate it closely align with what Philip Agre (1994) describes as *capture*. According to Agre, capture serves as both a linguistic metaphor (opposite the visual metaphors of surveillance, as articulated by Orwell and Foucault) and material process of tracking used to characterise the institutional, technical logic whereby human activities are captured and represented, or tracked, within sociotechnical systems. Capture technologies, Agre explains, comprise five interlocking processes through

which sociotechnical systems represent, constitute, direct and/or transform human activity through its purported 'discovery'; these processes include: (1) analysis, (2) articulation, (3) imposition, (4) instrumentation, and (5) elaboration. First, the activity in question is analysed and ontologically rendered into basic, programmatic terms (objects, relations, variables, etc.) for the subsequent articulation of grammars of action, which delineate 'the ways in which those units can be strung together to form actual sensible stretches of activity' (Agre 1994, 746). Next, these grammars are socially and/or technically imposed upon those engaged in that activity (i.e., made legible by the capture system) and recorded via some means of instrumentation. Lastly, captured records of that activity may be elaborated upon (audited, modelled, merged, stored) for optimisation. Capture, as Agre clarifies, may thus be deployed for either the archiving of data as input and/or the abstraction of 'semantic notions or distinctions, without reference to the actual taking in of data' (744), as with AI-based systems. Thus, as Chun (2016) explains, 'An AI program has successfully "captured" a behaviour when it can mimic an action ... without having to sample the actual movement' (59–60).

As Malik (2010) argues, however, 'control in the cybernetic sense does not mean absolute control of every detail. It is more like steering, directing and guiding' (33). To aid in this guidance requires a broad brush applied to cull information into categories. Take, for instance, AI-based risk assessments – built upon the fallible premises of cybernetic prediction – that are accurate only insofar as they produce risk as *simulation via capture* (i.e., of past behaviour) by categorising individual risk in terms of broad sociological data. Cathy O'Neil (2016) describes various public and private domains within which predictive models obscure – and ultimately magnify – human bias, such as the use of recidivism software for criminal sentencing just mentioned. As O'Neil argues, 'sentencing models that profile a person by his or her circumstances', including socioeconomic status and familial/social ties, 'help to create the environment that justifies their assumptions' (O'Neil 2016, 29). Accordingly, the risk of recidivism is primed using narrow parameters that often exacerbate racial and class-based disparities. In a recent interview, Wendy Hui Kyong Chun similarly discusses the proclivity for credit monitoring systems to reify the purported 'risks' they aim to detect and avoid (i.e., [in]ability for repayment; Chun and Cotte 2020). Based on various factors (beyond the borrower's credit/financial history, such as educational attainment and social network ties, etc.) risk assessment models designed to predict creditworthiness are, in effect, programming the very conditions they claim to eschew – an outcome of benevolent surveillance described elsewhere by Marion Fourcade and Kieran Healy (2007; 2017). What these cases demonstrate is the relationship between *capture* and *risk*, whereby risk as simulation becomes embedded within technical systems of capture intended to predict and mitigate future risk.

The Self-Referential Learning Machine

From earlier approaches to AI (i.e., 'expert systems' built upon 'if-then' rules with limited scalability), presently dominant approaches rely upon unsupervised machine/deep learning, leveraging information theory and connectionism for scalable prediction and decision-making (for a comprehensive discussion of AI paradigms and their evolution, see Russell and Norvig 2016). Among the myriad public and private domains wherein these AI-based systems *prime* social outcomes, perhaps the most consequential and ethically questionable is the criminal-legal system. In the U.S., algorithmic decision-making programs, predictive policing applications, and targeted/anticipatory surveillance technologies have become standard fare. Wiener recognised the potential for human actors – governments, militaries, and other cultural hegemons – to leverage the power of the *learning machine* against its citizenry, and cautioned as much. To mitigate such domination – both of the machine and the human actors who seek to leverage its power, Wiener (1989) heeds that 'we must know as scientists what man's nature is and what his built-in purposes are, even when we must wield this knowledge as soldiers and as statesmen; and we must know why we wish to control him' (182). It is not just the scientist, he notes, that should be responsible for our new technological future, but also the anthropologist and philosopher, if we are to prevent such an entropic reality.

Complicating the relationship between information input and predictive outcomes is the problem of data categorization that is foundational to capture technologies. For example, as applied to risk assessments for criminal offenders, a qualitative understanding of the perpetrator, as well as those individually particular antecedents which may have factored into the commission of a particular crime, are secondary (if considered at all) to the broad categories within which a perpetrator may fall. Data such as age, race, and socioeconomic status are far more valuable to the cybernetic game because they may be reduced to easily quantifiable statistics. The propensity for AI-based, cybernetic systems to *prime* (i.e., 'prune') human behaviour has been explored by several scholars, albeit in different ways: From reproducing essentialist social categories and magnifying their attendant (institutional, economic, etc.) disparities, to transposing notions of risk and the institutional handlings thereof. In *Coming to Terms with Chance*, for instance, Oscar Gandy Jr. (2009) describes cross-sector technologies of 'rational discrimination' that 'facilitate the identification, classification and comparative assessment of analytically generated groups in terms of their expected value or risk' (55). Such techniques, leveraging actuarial risk models and statistical evidence for purposes of prediction, serve to emphasise and reify race as an essential category (via proxy measures; see also Harcourt, 2015).

Cybernetics, at its core, is the acute science of subjective choice reduction as a means of avoiding entropy, which makes such categorization attractive.

As Faucher argues, 'Cybernetics does not drive toward the ultimate truth or solution, but is geared toward narrowing the field of approximations for better technical results by minimizing on entropy' (2013, 206). Yet, as modern applications of algorithmic and AI-based risk assessment systems illustrate, the push towards determining a predicted 'truth' or 'solution' has achieved the opposite, partly due to the reliance upon categories of data – rather than a variety – as the heuristic which guides machine learning. Wiener (1989) illustrates the value of variety of external input in digital systems, warning that closed systems run the risk of homogeneity, thereby increasing entropy, or a devaluation of output. To illustrate this point, and simultaneously argue that systems will only be as good as their human creators make them, Wiener envisions a digital remaking of Maelzel's chess-playing Automaton Turk as an example of where the future of machine learning may lead, if variety in external output is considered:

> A chess-playing machine which learns might show a great range of performance, dependent on the quality of the players against whom it had been pitted. The best way to make a master machine would probably be to pit it against a wide variety of good chess players. (177)

His reference to the Automaton Turk is quite apt, as it is seen both in its day and in hindsight as an iconic example of technological deception at the hands of a skilled human operator, able to fool the audience based on both sleight of hand theatrics, as well as a keen insight into predictable human behaviour.

To Wiener (1989), exposing a novel computerised version of the 'Turk' to a variety of chess master challengers offers hope that the system can learn from mistakes, recalling past defeats in an effort to not repeat them. This exposure to variety, thus, unburdens the chess-playing automaton – once the controlled object of a single human operator – from its storied narrative of being nothing more than an inauthentic representation of communicative exchange between subject (human audience) and machine object. The machine may escape an entropic fate by gaining new information via the continued interaction with a variety of experts. Yet, from a 21st-century perspective, Wiener's optimism falls short two-fold: (1) machine learning is capable of self-propagation (Uricchio 2017), reducing the role of human input to that of passive data source, rather than active participant in the creation of knowledge, and (2) the basis of machine learning as Wiener envisioned it – that of cybernetic feedback loops informed by past action to predict future outcomes – allows for applied interpretations that dismiss present context (Halpern 2014). Additionally, the produced information output itself – conforming to a grammar of action imposed to maintain 'compliance between system records and ongoing events' (Agre 1994, 748) – is reified as truth, rather than simply more information. It is this reification, evidenced in the practice of risk assessment technologies, that steers the use of these technologies away from the aim of cybernetic negentropy and

towards what Wiener cautioned against: homogeneity within the closed system that will ultimately undermine it.

Capturing Behaviour, Programming Risk

> *The computers won, but not because we were able to build abstract models and complex situations of human reasoning. They bypassed the problem of the agent's inner life altogether. The new machines do not need to be able to think; they just need to be able to learn.*
>
> – Fourcade and Healy (2017, 24)

In order for machines to learn, they must be able to correct prior errors. In order to correct such errors, those missteps must be recorded as feedback in order to inform the feed-forward. The philosophical underpinning as to why and how such errors can be recorded stems from Wiener's assessment of biological memory as the by-product of synaptic flares that imprint on the human mind due to the physiological gravity of experience. In other words, memories stick – and may even aid in shaping how we approach future events – when they are derived from heightened sensory experience. As an analogy: I may not recall what I ate for breakfast on an otherwise insignificant and random day a decade ago, but I can tell you precisely the colour of the bike and the sensation of pain that I experienced when first riding and crashing a bicycle. It is not the narrative of the event that imprints the memory, but the connection of that event to a physiological sensation experienced emotionally or tactically. It is this reflection upon past experience that paves the way for understanding the mind monad as a system of learning, which Wiener believes could be replicated in machine form.

Like Leibniz, Weiner qualifies the relationship between the mind and experience (past and present) as a communicative process, though goes a step further to suggest that 'the organism is not like the clockwork monad of Leibnitz with its pre-established harmony with the universe, but actually seeks a new equilibrium with the universe and its future contingencies' (Wiener 1989, 48). Simply put, like the pruning of Darwinian natural selection, the potential for robust cybernetic systems to weed-out frailties in the organism prepares, or as we argue, *primes* the subject for future environments.

Unfortunately, the data upon which these systems operate are often biased, incomplete or simply unqualified. For example, in the sentencing of convicted criminals, factors beyond the individual's crime – such as broader recidivism rates based on socioeconomic and demographic data – are used to predict the likelihood an individual may be a repeat offender, thereby influencing sentencing (Hillman 2019). Accordingly, it is fair to question whether such potential 'predicted' outcomes are primed via the algorithmic encoding of emotional triggers Weiner believed encouraged behavioural repetition.

As Halpern (2014) argues, the basis for Wiener's belief in the possibility of prediction is that humans, under duress, act repetitively. When applied to law enforcement, this logic produces an ostensible feedback loop whereby, for instance, statistical models based on prior (individual) arrest rates – already contaminated by racial/demographic assumptions vis-à-vis crime (e.g., over-policing of Black neighbourhoods; see for example Crawford 2018; Pasquale 2015) – 'generates the data that validate its hypotheses about race without necessarily involving animus based on features unrelated to criminal behaviour' (Gandy 2009, 125). In such an algorithmic scheme – aptly described by Frank Pasquale as a 'reputation system' – based on cybernetic principles of prediction, the individual is reduced to mere data points of *past* behaviour coupled with macro-level sociological data in a decision-making feedback loop bereft of present context. In the language of capture, this produces a grammar of action that reorients and superintends – through imposition and instrumentation – the activities of those within a given socio-technical system (in the case of law enforcement, both officer and suspect); that is, human activity becomes systematised around a standard ontology for 'maintaining the correspondence between the representation and the reality' (Agre 1994, 742).

To elucidate the psycho-social consequences of this process, we consider data policing/management programs and actuarial risk assessment tools for criminal sentencing, as these most readily clarify the notion of risk as simulation; indeed, as Harcourt (2015) explains, 'risk today has collapsed into prior criminal history' (237). Examples of these tools – particularly in the United States – are innumerable and continue to gain traction among state and federal law enforcement agencies. Introduced by the New York Police Department (NYPD) in 1995, CompStat (short for computer and/or comparative statistics) was developed to capture and index, in real-time, crime-related data that law enforcement may use to inform and direct policing efforts (Bureau of Justice Assistance 2013). Similarly, PredPol[1] was developed by the Los Angeles Police Department (LAPD) and researchers at the University of Southern California in 2012 (PredPol n.d.). Unlike CompStat, which initially relied only upon historical (macro-level) crime data to track and prevent crimes, PredPol was designed to anticipate when and where crime *might* occur (PredPol n.d.). Since their inception, the prevalence, sophistication and purported accuracy of these and similar tools has increased: As of 2016, 20 of the 50 largest law enforcement agencies in the US reported using *at least* one form of predictive policing (Jouvenal 2016), demonstrating a broader shift toward 'algorithmic governance' and data policing (Završnik 2019, 2). That is, a reliance upon automated data analysis and prediction (e.g., via AI-based systems) for decision-making by law enforcement and intelligence agencies, which, according to Završnik, is supported by the neoliberal emphasis on objectivity, legitimacy and efficiency (see also Benbouzid 2016; Wang 2018).

Presented as an affordable and reliable solution to limited police resources, predictive policing, by all accounts, appears neutral and accurate. In 2013, following the forced downsizing of the police department in Reading, Pennsylvania, police chief William Heim implemented PredPol in order to streamline law enforcement efforts; one year later, the number of reported burglaries decreased by 23 percent (O'Neil 2016). Despite this and other seemingly positive outcomes, critics have warned of the potential consequences of predictive policing programs; namely, for their reliance upon skewed and self-reinforcing crime-related statistics from the over-policing of communities of colour (Ferguson 2017; Hinton 2016; Jouvenal 2016). That CompStat has been critically associated with the 'broken windows' theory of policing[2] further clarifies the inherent social biases and prejudicial animus – whether implicit or overt – embedded within such tools and their attendant practices (e.g., Eterno and Silverman 2006).

Consider the following scenario: In an effort to stymie crime in a poverty-stricken, urban neighbourhood – itself a historical product of multi-layered and intersecting patterns of social and economic disenfranchisement, usually along racial lines – police engage in round-the-clock patrols of that area. As a result, and by virtue of institutionalised pressure to tangibly reduce crime (Giacalone and Vitale 2017), police stops and summonses become more frequent. Consequently, reported crime rates for that neighbourhood increase, feeding police management databases and prediction tools context-deprived data points, thereby prompting further patrols, arrests and so on, thereby triggering 'cascading disadvantages' (Pasquale 2015). Thus, the 'CompStat mentality' (Giacalone and Vitale 2017) – impelled by blind faith in the capture/analysis of quantitative data, corresponding to the neoliberal underpinnings of 'algorithmic governance' (Završnik 2019) – may be understood as a grammar of action that *primes* law enforcement toward decontextualised metrics of productivity, obscuring the connotations of physical violence within the 'capture' metaphor (Agre 1994). As crime becomes untethered from its social dynamics through this grammar of action, so too does law enforcement become estranged from the communities it claims to serve and protect.

Unlike crime management tools (e.g., CompStat) and predictive policing software (e.g., PredPol), which, at their core, aim to 'prevent' criminal activity by forecasting who is most likely to commit what type of crime, when and where – using *historical* crime data – criminal risk assessment programs assess the likelihood of recidivism (i.e., that a convicted criminal will re-offend). According to Carlson (2017), such tools include 'actuarial instruments, or models that predict risk of recidivism by studying the common traits of paroled inmates responsible for committing multiple crimes' (305). Although many risk assessment programs available today rely upon AI and algorithmic models, predictive assessments of criminal risk have been used in the US since the 1930s[3] and have steadily gained traction among law enforcement since (Harcourt 2007). In fact,

the National Institute of Corrections, a subdivision of the US Justice Department, *encourages* law enforcement agencies to incorporate risk assessments at each stage of the legal process (Angwin et al. 2016). Given the seeming potential to reduce incarceration rates and correctional costs by ranking offenders according to probable threat (Harcourt 2015), risk assessment is among the leading forms of predictive decision-making within the criminal justice system.

Investigations of risk assessment programs and their outcomes, however, have revealed the very inequities and ethical issues detailed earlier (e.g., Harcourt 2007; 2015; O'Neil 2016). As Casacuberta and Guersenzvaig (2018) explain, the utlisation of these algorithms is predicated upon an assumption of fairness and objectivity, though such outcomes are not necessarily guaranteed. Take, for example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool – an extension of the Level of Service Inventory (LSI), the leading risk assessment instrument among law enforcement agencies (Angwin et al. 2016; Northpointe 2015; Wykstra 2018). In 2016, an investigation conducted by *ProPublica* revealed the degree to which implicit racial bias impacted risk assessment scores via COMPAS (Angwin et al. 2016; Wykstra 2018). Using results from over 7,000 arrestees in Broward County, Florida, Angwin and colleagues reached several conclusions: Not only were risk assessment scores unreliable for projecting violent crimes, they were also unevenly distributed between Black and White defendants. As the researchers concluded, COMPAS 'was particularly likely to falsely flag Black defendants as future criminals, wrongly labelling them this way at almost twice the rate as White defendants' (Angwin et al. 2016, para. 15). Perhaps unsurprisingly, Northpointe, the for-profit organisation behind COMPAS, maintains that the program does *not* consider racial categories in calculating risk; however, as Harcourt (2015) argues, other factors included within risk assessment models serve as proxies for race. Specifically, the COMPAS model, as well as other risk models such as LSI and LSI-R (the Level of Service Inventory-Revised), includes educational attainment (both of the individual and their family members), employment status and income, and prior criminal history in determining risk, which distribute unevenly along racial lines and thus reflect – and augment – the pathologising effects common among earlier policing practices (see Hinton 2016). O'Neil (2016) aptly illustrates this scenario:

> A person who scores as 'high risk' is likely to be unemployed and to come from a neighbourhood where many of his friends and family have had run-ins with the law. Thanks in part to the resulting high score on the evaluation, he gets a longer sentence, locking him away for more years in prison where he's surrounded by fellow criminals – which raises the likelihood that he'll return to prison. He is finally released into the same poor neighbourhood, this time with a criminal record, which makes it that much harder to find a job. If he commits another crime,

> the recidivism model can claim another success. But in fact, the model itself contributes to a toxic cycle and helps to maintain it. (27)

As this reabsorption makes clear, risk outputs derived from algorithmic and AI-based programs – designed to analyse, predict, and mitigate or prevent future damages (harms, losses, etc.) – do little beyond programming and, arguably, *ensuring* future risk; particularly when risk scores are introduced during criminal trials. In bypassing crucial facets of the human experience and other exogenous factors (e.g., prison cycling), risk is reduced to a sequence of quantitative variables – a grammar of action imposed upon those whose activities have been captured (Agre 1994) – that, taken together, merely (re)produce the hyperreal, the imaginary, and the immanent. As Baudrillard (1994) suggested in delineating third-order simulacra, 'The models no longer constitute the imaginary in relation to the real, they are themselves an anticipation of the real, and thus leave no room for any sort of fictional anticipation' (122). That is, in attempting to model and predict risk (of recidivism), the gap between real and imagined risk yawns, producing risk as simulation – an imitation of risk simulated and reabsorbed through retrospective cybernetic systems and practices thereof.

Conclusion: Lifting the Cybernetic Veil

Wiener acknowledges the potential for the abuse of cybernetic systems by external forces when he warns that 'the machine's danger to society is not from the machine itself, but what man makes of it' (1989, 182). 'The great weakness of the machine' (1989, 181), he states – the weakness that would prevent the domination of humankind by machines and, subsequently, those human agents who seek to leverage the power of cybernetics for control over populations – is that the machine itself cannot account for the myriad conditions that qualify human existence. Leibniz considered these myriad conditions, as well as future possibilities, to be contingencies that were accounted for, organised and even predicted by a pre-established harmony vis-à-vis God, or divine intermediary. To Leibniz, the power of his new cybernetic 'calculus' of communication was as a tool of ontological reflection; however, this is where Wiener breaks from Leibniz.

In order to understand our current path towards a socio-psychological entropic fate at the hands of cybernetic prediction, it is necessary to reflect upon the gravity of the philosophical detour Wiener (1989) takes from the 'patron saint of cybernetics'. Leibniz, like Descartes before him, made early headway into the question of substance dualism, or the distinction between the *mind* (the thinking substance) and *body* (the extended substance) as separate, though dependent, entities. To Descartes, these created substances are relational, working in perfect union of mind and matter to form the subject. What distinguishes Leibniz's approach to the mind-body problem is his

rejection of the notion of the body as extended *substance*, and therefore subject. In his theory of *monads* – or, simple, unextended substances – Leibniz agrees with Descartes that the mind (or soul) qualifies as substance, or monad. Monads are independent from causal extension; therefore the body does not qualify. Additionally, a monad's properties are naturally continually active, changing, and evolving over time.

On the surface, this argument seems contradictory to his rejection of Cartesian dualism. If both the mind and body evolve, how can they not be both seen as *substance*? The answer lies in Leibniz's definition of the natural world. As de Mendonça (2008) states in her explication of Leibniz's concept of nature, Leibniz distinguishes between material nature, or that 'which is produced in nature according to mechanical principles', and that which is natural to the soul, 'and explained by its own principles – namely, the principle of perfection' (187). The distinction between mind and body, then, is found in the genus of each. The soul, as natural perfection, is created by God. The body, as material form, is merely organised and transformed by the laws of nature. In this sense, the mind and body are not equal, causal entities; rather, the mind, or soul, is the ultimate conductor of the subject.

The implications to current applications of cybernetics, in general, and AI, specifically, are thus called into question. To see the body as extended substance, as Wiener did, provides the philosophical foundation upon which one can justify the mechanical object as a replication of the human brain. The cybernetic brain-computer model, as Sconce (2019) notes, while perhaps deeply flawed in analogy and application, is something 'we all believe' (235) to be self-evident. Echoing Hayles (2008), Sconce (2019) continues: 'Underlying the cybernetic dream of uploading consciousness is a magical positivism born of a panicked materialism, a belief that any and all questions can be resolved through the accumulation of sufficient data' (235). Yet, as the aforementioned examples of cybernetic risk assessment illustrate, this accumulation of data is often far from sufficient, and more often than not subjectively reductive.

Perhaps it is time to revisit the mind-body problem as it relates to cybernetic principles, and explore the merits of predictive technology from this philosophical foundation. In his new calculus, Leibniz introduced the mind-body problem that 'included the new concept of the differential within the field of significations' (Serfati 2008, 127). To Leibniz, meaning is a complex negotiation between both what is tangibly present, tangibly missing, and the qualitative significance of that difference. External 'substances' therefore, cannot be regarded as true subjects, but rather as modes or states of presentations of an assemblage. By biasing towards the thesis that the mind (or soul) is the single natural source of human substance, and everything else an ever-evolving assemblage of material, perception and transformation, Leibniz paves the way for understanding the pitfalls of cybernetic prediction as it is applied today. Such a critique is echoed in the work of contemporary scholars like Orit Halpern, and is ripe for continued critical examination. Perception, to Leibniz, is a complex calculus

between the representation of the object, the subject perceiving that object, *and* the discursive properties of that interaction. Yet, it is the discursive nature of communication systems that cyberneticians often fail to consider. As Halpern (2014) notes, Wiener understood that not all forms of information (e.g., metaphorical representations, connotative meaning, denotative descriptions, etc.) could be recorded into cybernetic systems, thereby making the foundation of prediction recognised today wholly incomplete.

In our contemporary context, and with an ever-increasing black-boxed world subsuming ontological truth, revisiting the theological investigation of the provenance of 'natural' or universal truth is necessary. As Sconce (2019) keenly observes,

> In this post-human universe of secular data management, the immateriality of information replaces the ontological infinitude of God as the occult field of magical omniscience, promising its acolytes, through the transubstantiating miracle of magical positivism, the possibility of deliverance from the mortal humiliations of material existence. (235)

This is not to say that a purely theological view of truth, or life itself, should be embraced. Rather, the same rigour of inquiry that has defined metaphysical philosophy since Descartes must be applied to contemporary instantiations of often unquestioned truth: The black box, the algorithm, the cybernetic veil of AI. In his essay, 'The Technology of Enchantment and the Enchantment of Technology', Alfred Gell (1992) urges such an approach by examining the *process* of creation. Creators, through imbued skill and cultivated craft, are often revered as gods amongst their human peers. Yet, when the artefact is a technological system, the question emerges: Should we allow the unquestioned sovereignty of those who create the systems that ever-increasingly seek to orchestrate and prime our daily outcomes simply because those creators possess a skill we do not? What Gell advocates, perhaps unintentionally, is something that many – from philosophers of technology to everyday consumers – grapple with today: The godlike status those who create are granted, often passively, by those who rely upon the skilled to navigate an increasingly technologically dependent society.

During an era wherein human and technological systems have become evermore intertwined – often to the point of obscurity – a critical understanding of this godlike and unquestioned role humans play in developing technological systems is increasingly necessary; particularly as these systems have become the hidden blueprint of our sociological condition. As Wiener states:

> Those who would organize us according to permanent individual functions and permanent individual restrictions condemn the human race to move at much less than half-steam. They throw away nearly all our human possibilities and by limiting the modes in which we may adapt

> ourselves to future contingencies, they reduce our chances for a reasonably long existence on this earth. (52)

As is the case with software like PredPol, using broad categories of social data to predict individual behaviour not only misapplies cybernetic principles of learning vis-à-vis feedback beyond its narrowly defined parameters, but risks limiting human possibility as Wiener warned. Instead of fetishising algorithmic futures, researchers should continue the endeavour of philosophical questioning of algorithmic contingencies and the point of creation, as well as practical inquiry into the genesis of the data, how that is accrued, and implications of relying upon categories to 'predict' individual action.

We must actively and critically embrace that humans, not sublime or other godlike manifestations, are the creators of artefacts that mitigate our ontologies. The implications of this acknowledgment are philosophically far-reaching, upending a culturally-entrenched power dynamic between creator of technology and unquestioning consumer that persists even today – an era saturated with information, simulation and, ultimately, primed prediction.

Notes

[1] Short for 'predictive policing', for which the program has its own definition: 'The practice of identifying the times and locations where specific crimes are most likely to occur, then patrolling those areas to prevent those crimes from occurring' (PredPol n.d.).

[2] Essentially, the 'broken windows' theory of policing argues that if minor offenses or criminal acts are left unattended, thus indicating a lack of regard, more serious criminal activity and 'urban decay' will follow; see Kelling and Wilson 1982.

[3] As Harcourt (2007) notes, the first risk assessment instrument was introduced in Illinois in the 1930s.

References

Agre, P. E. 1994. Surveillance and Capture: Two Models of Privacy. *The Information Society*, 10(2), 101–127.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. 2016. Machine Bias: Risk Assessments in Criminal Sentencing. *ProPublica*, 23 May. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ashford, D. 2017. The Mechanical Turk: Enduring Misapprehensions Concerning Artificial Intelligence. *The Cambridge Quarterly*, 46(2), 119–139. DOI: https://doi.org/10.1093/camqtly/bfx005

Baudrillard, J. 1994. *Simulacra and Simulation*. Ann Arbor, MI: University of Michigan Press.

Baudrillard, J. 1995. The Virtual Illusion: Or the Automatic Writing of the World. *Theory, Culture & Society*, 12(4), 97–107.

Benbouzid, B. 2016. Who Benefits from the Crime? *Books & Ideas*, 31 October. Retrieved from: https://booksandideas.net/Who-Benefits-from-the-Crime.html

Buchli, V. 2010. Presencing the Im-Material. In: M. Bille, F. Hastrup and T. Soerensen (Eds.), *An Anthropology of Absence*, pp. 185–203. New York: Springer.

Bureau of Justice Assistance. 2013. *Compstat: Its Origins, Evolution, and Future in Law Enforcement Agencies*. Retrieved from: https://www.bja.gov/publications/perf-compstat.pdf

Carlson, A. M. 2017. The Need for Transparency in the Age of Predictive Sentencing Algorithms. *Iowa Law Review*, 103, 303–329.

Casacuberta, D. and Guersenzvaig, A. 2018. Using Dreyfus' Legacy to Understand Justice in Algorithmic-Based Processes. *AI & Society*, 1435–5655, 1–7. DOI: https://doi.org/10.1007/s00146-018-0803-2

Castillo, D. R. and Egginton, W. 2017. *Medialogies: Reading Reality in the Age of Inflationary Media*. New York: Bloomsbury Academic.

Chun, W. H. K. 2016. *Updating to Remain the Same: Habitual New Media*. Cambridge, MA: MIT Press.

Chun, W. H. K. and Cotte, J. 2020. Reimagining Networks: An Interview with Wendy Hui Kyong Chun. *The New Inquiry*, 12 May. Retrieved from: https://thenewinquiry.com/reimagining-networks

Crawford, C. E. 2018. Minorities, Space, and Policing. In: C. E. Crawford (Ed.), *Spatial Policing: The Influence of Time, Space, and Geography on Law Enforcement Practices* (2nd ed.), pp. 73–92. Durham, NC: Carolina Academic Press.

de Mendonça, M. 2008. Leibniz's Conception of Natural Explanation. In: M. Dascal (Ed.), *Leibniz: What Kind of Rationalist? Logic, Epistemology, and the Unity of Science*, pp. 183–197. New York: Springer.

Eterno, J. A. and Silverman, E. B. 2006. The New York City Police Department's Compstat: Dream or Nightmare? *International Journal of Police Science and Management*, 8(3), 218–231.

Faucher, K. X. 2013. *Metastasis and Metastability: A Deleuzian Approach to Information*. Rotterdam, NL: SensePublishers.

Ferguson, A. G. 2017. The Police are Using Computer Algorithms to Tell if You're a Threat. *Time*, 3 October. Retrieved from: http://time.com/4966125/police-departments-algorithms-chicago

Fourcade, M. and Healy, K. 2007. Moral Views of Market Society. *Annual Review of Sociology*, 33, 285–311. DOI: https://doi.org/10.1146/annurev.soc.33.040406.131642

Fourcade, M. and Healy, K. 2017. Seeing like a Market. *Socio-Economic Review*, 15(1), 9–29. DOI: https://doi.org/10.1093/ser/mww033

Galison, P. 1994. The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision. *Critical Inquiry*, 21(1), 228–266.

Gandy, Jr., O. H. 2009. *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*. Farnham: Ashgate Publishing.

Gell, A. 1992. The Technology of Enchantment and the Enchantment of Technology. In: J. Coote and A. Shelton (Eds.), *Anthropology, Art, and Aesthetics*, pp. 40–63. Oxford: Oxford University Press.

Giacalone, J. L. and Vitale, A. S. 2017. When Policing Stats do More Harm than Good. *USA Today*, 9 February. Retrieved from: https://www.usatoday.com/story/opinion/policing/spotlight/2017/02/09/compstat-computer-police-policing-the-usa-community/97568874

Gitelman, L. 2014. *Paper Knowledge: Towards a Media History of Documents*. Durham, NC: Duke University Press.

Halpern, O. 2014. *Beautiful Data: A History of Vision and Reason Since 1945*. Durham, NC: Duke University Press.

Hand, M. 2014. From Cyberspace to the Dataverse: Trajectories in Digital Social Research. In: M. Hand and S. Hillyard (Eds.), *Big Data? Qualitative Approaches to Digital Research*, pp. 1–27. Bingley: Emerald Group Publishing Limited.

Harcourt, B. E. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago, IL: University of Chicago Press.

Harcourt, B. E. 2015. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter*, 27(4), 237–243. DOI: https://doi.org/10.1525/FSR.2015.27.4.237

Hayles, N. K. 2008. *How We Become Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago, IL: University of Chicago Press.

Hillman, N. L. 2019. The Use of Artificial Intelligence in Gauging the Risk of Recidivism. *The Judges' Journal*, 58(1), 36–39.

Hinton, E. 2016. *From the War on Poverty to the War on Crime: The Making of Mass Incarceration in America*. Cambridge, MA: Harvard University Press.

Jouvenal, J. 2016. Police are Using Software to Predict Crime. Is it a 'Holy Grail' or Biased Against Minorities? *The Washington Post*, 17 November. Retrieved from: https://www.washingtonpost.com/local/public-safety/police-are-using-software-topredict-crime-is-it-a-holy-grail-or-biased-against-minorities/2016/11/17/525a6649-0472-440a-aae1-b283aa8e5de8_story.html?utm_term=.ed1fc4745464

Kelling, G. L. and Wilson, J. Q. 1982. Broken Windows: The Police and Neighborhood Safety. *The Atlantic*, 1 March. Retrieved from: https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465

Malik, C. 2010. *Ahead of Change: How Crowd Psychology and Cybernetics Transform the Way We Govern*. Frankfurt, DE: Campus Verlag.

Manovich, L. 2017. Cultural Analytics, Social Computing and Digital humanities. In M. T. Schäfer and K. van Es (Eds.), *The Datafied Society: Studying*

Culture through Data, pp. 55–68. Amsterdam, NL: Amsterdam University Press.

Mumford, L. 1972. Technics and the Nature of Man. In: C. Mitcham and R. Mackey (Eds.), *Philosophy and Technology*, pp. 77–85. New York: The Free Press.

Northpointe. 2015. *Practitioner's Guide to COMPAS Core*. Retrieved from: http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

PredPol. n.d. Retrieved from: https://www.predpol.com/about

Russell, S. and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach* (3rd ed.), London: Pearson.

Sconce, J. 2019. *The Technical Delusion: Electronics, Power, Insanity*. Durham, NC: Duke University Press.

Serfati, M. 2008. Symbolic Inventiveness and 'Irrationalist' Practices in Leibniz's Mathematics. In: M. Dascal (Ed.), *Leibniz: What Kind of Rationalist?*, pp. 125–139. New York: Springer.

Uricchio, W. 2017. Data, Culture, and the Ambivalence of Algorithms. In: M. T. Schäfer and K. van Es (Eds.), *The Datafied Society: Studying Culture through Data*, pp. 125–137. Amsterdam, NL: Amsterdam University Press.

Wang, J. 2018. *Carceral Capitalism*. Cambridge, MA: MIT Press.

Wiener, N. 1989. *The Human Use of Human Beings: Cybernetics and Society*. London, UK: Free Association Books.

Wykstra, S. 2018. Philosopher's Corner: What is 'Fair'? Algorithms in Criminal Justice. *Issues in Science and Technology*, 24(3), 21–23. Retrieved from: https://issues.org/perspective-philosophers-corner-what-is-fair-algorithms-in-criminal-justice/

Završnik, A. 2019. Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings. *European Journal of Criminology*, 1–20. DOI: https://doi.org/10.1177/1477370819876762

CHAPTER 12

Algorithmic Logic in Digital Capitalism

Jernej A. Prodnik

Introduction

In recent years research in social sciences and related academic fields has attributed increased importance to algorithms and their impact on social relations and our everyday lives. While algorithms are nothing particularly new and can be closely related to computing or even mathematics as such, debates have slowly but surely moved beyond the narrow confines of the so-called hard sciences. They are now taking centre stage when authors analyse topics such as political communication, electoral campaigning and mass micro-targeting of potential voters (Moore 2018; Vaidhyanathan 2018), automated trading in stock markets and various other types of financial transactions (Pasquale 2015, Ch. 4; MacKenzie 2017; 2018), or the impact of technological innovations on journalism (Diakopoulos 2019). Their influence is emphasised in healthcare, loan approvals, transportation, traffic-control, city urbanization, education, employment, policing, security and even military conflicts (Fisher 2020; Bridle 2018; Moore 2018; Munn 2018; Mosco 2014). Critical analysis has demonstrated their impact in constructing 'digital poorhouses', since they have become prominent in state administration and eligibility systems for poverty management (Eubanks 2017). It is also impossible to ignore them when considering technologies forming the Internet of Things and cloud computing (Bunz 2014; Mosco 2014), search engines, digital social networking platforms and various recommendation systems, or ranking, reputation and personalisation

How to cite this book chapter:
Prodnik, J. A. 2021. Algorithmic Logic in Digital Capitalism. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 203–222. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.l. License: CC-BY-NC-ND 4.0

tools aimed at tracking and controlling of behavioural patterns (Mager 2012; Gillespie 2014; Prodnik 2014; Kitchin 2017; Srnicek 2017; Fuchs 2019).

This is to name only some of the most prominent issues that recent research has focused on, with many more aspects of our lives affected on a daily basis (Willson 2016). There seems to be little doubt algorithms now play one of the central roles in almost all spheres of society, from politics and economy to culture and interpersonal relationships, subsequently raising various types of ethical issues (Mittelstadt et al. 2016; Coeckelbergh 2020).

In digital environments algorithms overlap and mutually influence each other, forming what can be considered layered algorithmic systems or ensembles of algorithms (cf. Kitchin 2017, 18–21). In this chapter I will not explain individual algorithms in an abstract manner, but rather focus on the key characteristics and social consequences of such ensembles of algorithms in their current hegemonic social form (for practical reasons I will simply refer to them as algorithms). This will hopefully shed some light on the reasons for their increasing social influence.

All technologies are inevitably embedded in – and influenced by – the social context in which they are developed, so my analysis will consider ensembles of algorithms as part of the competitive and inherently unstable capitalist society (Streeck 2012), or to put it more narrowly, as part of digital capitalism (Fuchs and Mosco 2015; Fuchs 2019). My contribution therefore aims to provide some answers on how algorithms work in digital capitalism, what are the key reasons for this and what is their impact for society at large. Focusing on digital capitalism assumes a theoretical framework of the political economy of communication, which points at the power asymmetries in society in an overarching manner, while taking on board the fact there is nothing 'natural' in these characteristics of algorithms. It also helps to move the analysis beyond abstract notions that have a limited explanatory value in specific historical contexts.

Understanding Ensembles of Algorithms in Capitalism

In contrast to many other topics there is a large degree of overlapping in how authors define algorithms. Bunz (2014, 7), for example, notes that an algorithm is 'a set of rules to be followed by calculations'. This definition does not differ significantly from either Bucher's (2017, 31), in which an algorithm 'is just another term for those carefully planned instructions that follow a sequential order', or Kitchin's (2017, 14), for whom algorithms are 'sets of defined steps structured to process instructions/data to produce an output'. In this sense all computer software and digital technologies are fundamentally composed of algorithms (ibid.). Even though they were put forward by social scientists, such definitions are quite abstract and cannot explain by themselves why the social impact of algorithms has been so significant in recent years, especially since there is no inherent technical necessity for their increased omnipresence.

As already noted, my aim is not to look for universal characteristics of algorithms – even if that were possible or made sense in social sciences – but to understand them as part of the existing historical epoch, where they are bundled together in vast and overlapping digital ensembles, predominantly under the control of powerful capitalist corporations. Not to interpret them as technical or mathematical constructs, but through their social causes, purposes and consequences when implemented and executed (cf. Mittelstadt et al. 2016, 2–3). This choice comes close to popular definitions of algorithms and has obvious downsides. It leaves much room for ambiguity and either risks making the scope of analysis too expansive, or puts too much focus solely on what could be called 'mega-algorithms', while ignoring the more basic ones. It is exactly these algorithms, however, that are most influential and consequential. As such, they must be subject to thorough scrutiny.

Algorithms as Narrow Artificial Intelligence

Before continuing I must note that for the purposes of this chapter I consider algorithms as part of a narrow form of artificial intelligence. They have limited autonomy beyond the tasks which they were made for. While the so-called Artificial General Intelligence has the capacity to behave intelligently in a wide variety of contexts and use knowledge in novel situations, emulating intelligence of human beings, it remains in the realm of speculation (Boden, 2016; Dyer-Witheford et al. 2019, Ch. 1; Mitchell 2019, Ch. 3; Coeckelbergh 2020). What is sometimes called narrow AI, however, is already widely present and exists in our everyday lives. It can be connected to algorithmic processes that normally address narrow tasks, which means that their application cannot be generalised to other domains of functioning. State of the art AI still lacks real understanding and thus flexibility to operate outside the frontiers of their own design (ibid.).

Because algorithms 'don't know what they don't know' human beings have an advantage especially in complex communication, expert thinking, and creative tasks (Diakopoulos 2019, 29–30, 122; cf. Bunz 2014, 17; Mitchell 2019). It is also very challenging for computers to perform non-routine tasks, as human beings have large reservoirs of tacit and contextual knowledge, which they are not even aware of (so-called Polanyi's paradox). The situation is similar with our most basic and unconscious sensorimotor abilities, including walking, manipulating objects or understanding complex language, which may be very simple tasks for human beings but are amongst the biggest challenges for engineers (Moravec's paradox). These issues are currently generating considerable engineering bottlenecks (see Frey 2019, 233–236).

The currently dominant paradigm in AI is machine learning, for example, via artificial neural networks which try to mimic human brains. Instead of being built top-down as a set of logical rules for handling data, machine learning

systems use an inductive approach for finding patterns, which are often based on statistical calculations and probability. A statistical pattern-recognition approach presupposes pattern extraction from data, with these systems creating their own models of inference. Developed solutions are therefore based on the data itself and on what these algorithms have previously learned (see Boden 2016; Mitchell 2019; Dyer-Witheford et al. 2019, 8–15; Bridle 2018, Ch. 6).

The fact that machines now continuously learn on data means that actors and institutions, that have access to quantitatively more and/or qualitatively better information, are in an advantageous position. They can improve the quality, effectiveness and capacities of their algorithms. This is an important point I will return to when describing the characteristics of algorithms in digital capitalism. Nonetheless, as a narrow form of AI these systems can currently generalise only on data they were trained for, and therefore merely simulate real intelligence.

Embedding Algorithms in Capitalism

To say that algorithms have to be considered as part of capitalist society may seem fairly inconsequential, as I noted at the start of this chapter. But this is a system with certain tendencies and basic characteristics that influence all phenomena operating within it. Even though these tendencies can be countered or partially neutralised in many ways, most obviously through politically enacted regulation, they are the result of existing and dynamic social structures. They do not pre-determine the outcomes, but they do set the framework and delimit the level of possibilities within that system (cf. Collier 1994). In other words, capitalism has a specific logic in how it operates, and the impact of that logic can be identified and analysed in various phenomena that work within this system.

A concise definition of capitalist society is provided by Streeck (2012), who argues that this 'is a society that has instituted its economy in a capitalist manner, in that it has coupled its material provision to the private accumulation of capital, measured in units of money, through free contractual exchange'. Similar to social scientists in the 19th century, he emphasises that there cannot be any strict empirical separation between society and economy because of their interrelatedness. Furthermore, economic relations are constantly attempting to consume non-economic relations through commodification, since this is a system that needs to expand constantly, paradoxically staying stable only when being in movement (cf. Prodnik 2016). Competitiveness, permanent revolutionising – presupposing continuous change, innovations, instability and uncertainty – and expansion of capital are therefore part and parcel of this system and influence all social relations (Streeck 2012, 5–9).

In a critical and holistic approach of the political economy of communication, it would thus not only be disadvantageous, but quite impossible to completely

dissemble algorithms from the wider capitalist context. Major ensembles of algorithms today are developed and owned by some of the biggest corporations in the world (Mosco 2014). Alphabet (Google), Facebook, Microsoft, Apple or Amazon might be seen only as tech companies, but they are expanding into and influencing numerous other branches of the economy and consequently our lives. Many other companies that primarily function as digital platforms, such as Uber or Airbnb, have brought similar economic disruption not only to the most prominent geographic locations, but also to peripheral ones (cf. Srnicek 2017). Even the automotive company Tesla views itself first and foremost as an innovative tech company, conspicuously basing its headquarters in Silicon Valley.

It is not only that algorithms play one of the most important roles in all of the cases mentioned above, they also clearly demonstrate it has become impossible to speak of 'digital-only' projects, that would somehow be separated from the non-digital world. In digital capitalism many formerly clear borders and demarcation lines have converged or completely collapsed as commodification seeps into every part of our lives, social practices and relations (Prodnik 2016).

Even though major digital corporations are in many ways breaking new ground, they are not entirely dissimilar from corporations of the old. They are in perpetual quest of either short-term or long-term profits and new areas where they could expand to, while constantly struggling to innovate and increase their market share. These very basic pursuits largely delimit the manner in which they design algorithms and why they are developed in the first place (cf. Mager 2012; Gillespie 2014, 176–177). Bilić (2018) points out that this is one of the central reasons why algorithms cannot be seen simply as technical artefacts. In the case of Alphabet, for example, algorithms are 'also business strategies for market control and dominance' (ibid. 71). This should be taken aboard before pondering further about the characteristics of algorithms, as it is relevant throughout the chapter.

Characteristics of Algorithms and their Structural Reasons

A study of the literature on algorithms cited in the previous sections makes it possible to define four basic characteristics of algorithms in digital capitalism: (1) opacity and obfuscation, (2) datafication, (3) automation, and (4) instrumental rationalisation. There are both structural reasons for these characteristics as well as wider consequences they could have on social relations and social totality. While these characteristics can be analytically separated, they are thoroughly interconnected in practice and frequently reinforce each other. The key point of emphasis mentioned earlier is that these characteristics should not be seen as universally inherent to algorithms, since they are to a considerable degree a product of the existing social order – digital capitalism. To put it differently, in a different political-economic context, there could be other

structural reasons at play, thus leading to changes in these basic characteristics or at least in their prominence.

Opacity and Obfuscation

The first fundamental characteristic of algorithms is opacity and obfuscation, which is mainly an outcome of their secrecy and restrictiveness, but also of technological complexity and multiplicity. In essence, how algorithms actually operate is to a large degree incomprehensible and difficult to understand, often even for experts. While we have basic ideas about the major algorithms, discerning details of how exactly they work, what data they are collecting, how it is used, why certain results finally appear, or who has access to them, is much more difficult or even impossible. Pasquale (2015), for example, notes that algorithms are secretive and restrictive black boxes. This seems like an apt metaphor, since it denotes both a recording device and 'a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other' (ibid. 3). Even though algorithms have wide-ranging consequences for the shape and direction of our societies, this means they are 'opaque and inaccessible to outside critique; and their parameters, intent and assumptions indiscernible' (Willson 2016, 4).

There are three major structural reasons for this characteristic. Firstly, they are privately owned and subject to various types of intellectual property rights (copyright, patents, trademarks, etc.), that generally presume secrecy (ibid.). As emphasised by Pasquale (2015, 61), 'the huge companies resist meaningful disclosure, and hide important decisions behind technology, and boilerplate contracts' (cf. Kitchin 2017, 20). While most companies that own algorithms have obvious commercial reasons to keep them opaque – intellectual work, as part of them, can serve as an important market advantage when competing with other companies (cf. Bilić 2018) – making them completely transparent could also lead to security breaches and attempts of manipulation. An obvious example is the gaming of search engines by rogue websites.

Even if there was full transparency on how algorithms work, most internet users would have serious problems if they tried to meaningfully comprehend them (Obar 2020; Willson 2016, 10). A digital divide can therefore be seen as the second structural reason for opacity, one that can be connected to the lack of expert digital literacy and programming knowledge of lay users. Power asymmetries and social inequalities are leading to exclusion in the digital as well as non-digital spheres as most people have vast difficulties with much more basic online understanding than complexities of algorithmic procedures. A poll conducted by the Pew Research Center (2018), for example, revealed that the majority of Facebook users had almost no knowledge of how their news feeds work. One can therefore only imagine how far from being able to understand the complexities of algorithms most internet users are.

Comprehending how algorithms operate, however, is not difficult only for the average user of the internet but even for experts. Several factors contribute to this, including the fact that they are 'always somewhat uncertain, provisional and messy fragile accomplishments', often worked on by large teams of programmers that constantly change and have a highly specialised division of labour between them, making an overview of the whole programming process difficult (Kitchin 2017, 18, 21; Bridle 2018, 40). In an analysis of financial algorithms, Pasquale (2015, 123; cf. 32), for instance, noticed how sometimes 'black boxes are so effective they even 'fool' their creators'. This is the third structural reason for opacity, which can be closely connected to the fact we are speaking about large ensembles of layered algorithms that are interconnected, mutually influencing each other and constantly expanding, multiplying and changing (cf. Willson 2016). Furthermore, results of sub-symbolic AI systems, for example deep learning neural networks that are increasingly used in machine learning, are very difficult to unpack, because they do not use symbols and a logic that is understandable to human beings (Mitchell 2019).

Datafication

Most algorithms make little sense or cannot even operate without the data which they process. Algorithmic decisions are made on this basis, meaning that the effectiveness of algorithms is 'strongly related to the data sets they compute' (Bunz 2014, 7). This is why, as I noted earlier, data is an increasingly important commodity in digital capitalism.

There are, again, several reasons for this characteristic, the most obvious being that decisions made by algorithms are based on computational calculations that can usually be made only via quantifiable information. This inherent dependency on data has as a consequence a clear tendency towards the datafication of various practices and relations. Or to put it differently, the transformation of social reality and the world into structured data schemes that generally exclude nuance and wider context (Diakopoulos 2019, 117). Datafication should not be seen as something static, but as a continuous process; it is an important characteristic of major algorithms because they require constant flows of data (i.e. Big Data) to perform their key functions – a tendency so prominent because of the increased computing power and the near total ubiquity of digital networks and their tracking capacities (Prodnik 2014).

It could certainly be argued that dependency on data holds true for algorithms as such, but it seems clear they truly gained relevance only with the availability of large troves of information that enable complex inferences, correlations and predictions drawn on a large scale (and beyond the scope and capacities of human beings). This also leads us to the second structural reason for datafication, one that is closely related to the first one: there is a constant need for enhanced capabilities and effective algorithms in a competitive

environment, at least if they are to produce improved results and operate better. Frey (2019, 304) notes that 'data can justly be regarded as the new oil'. Even though this illustration is overused and can at best be understood as a somewhat faulty metaphor, it is true that 'as big data gets bigger, algorithms get better' (ibid.). With the development of machine learning, exposing algorithms to more examples leads to improvements in how they perform tasks (Mitchell 2019, Ch. 6).

Datafication has become so pronounced only with the development of digital capitalism, where the constant need for more and more data has become both a self-perpetuating cycle and one of the central factors in the production process. If we borrowed a phrase from the Marxist conceptual apparatus, we could say that the sum total of the forces of production in a specific historical context had to be developed to a certain level to make this a real possibility. Viewing this as a universal characteristic of algorithms would therefore be difficult, since they do not need – as a necessity – vast quantities of data to perform the most basic functions. It is only when they become crucial in the production process that this becomes the case. Institutions and actors employing algorithms typically do so because they want to predict large scale trends, patterns and risks or try to exert control, again pushing towards datafication to come closer to this objective. This is closely related to the properties of digital capitalism and can be seen as the third structural reason for datafication (cf. Mosco 2014; Prodnik 2014).

Automation

Datafication is directly associated with the automation of processes, functions and decision-making. Automation is what makes algorithms so appealing in the first place, but it also 'means that information included in the database must be rendered into data, formalized, so that algorithms can act on it automatically' (Gillespie 2014, 170). Automation and datafication are therefore mutually intertwined. An attempt of automating decisions structurally pushes towards more datafication, with access to (more) data often enabling more intensive and extensive automation.

Automation is a very discernible characteristic that makes algorithms into an interesting option for various actors and institutions. It enables them to 'make high-quality decisions, and to do so very quickly and at scale' (Diakopoulos 2019, 19). This can lead both to a qualitative jump in acceleration of functions or procedures and their considerable scalability. Again, it may seem perfectly reasonable to claim this characteristic is universally inherent to algorithms and has little to do with digital capitalism; however there are three interrelated, but analytically distinguishable structural factors that counter this seemingly commonsensical notion.

First, algorithms and their capacities can bring increased competitive advantages to the companies employing them. Fractions of seconds, incomprehensible

to humans, can bring literally millions in financial trading or vast reductions in labour costs (Pasquale 2015, Ch. 4; MacKenzie 2017; Bridle 2018, 106–109).

Second, existing resources can be put to a much better use and can help move decision-making process beyond the limitations that are inherent to human beings. Formerly laborious operations are simplified and made easy, often literally a click away. Today, it seems almost incomprehensible to imagine a manual harvesting of data, the indexing of the internet or non-automated searching, which would be performed solely by human beings and did not happen instantly.

Third, and related to the previous reasons, increased overall efficiency is another enticing prospect for firms when applying algorithms. Attempts of automation have of course been a constituent part of the industrial capitalist society and there is a constant tendency on the part of capital to replace workers with machines and reduce labour costs (Dyer-Witheford et al. 2019, Ch. 1; cf. Marx 1867/1990). Algorithms thus present merely another step in that direction, but quite possibly a qualitatively new one, since human beings increasingly have difficulties competing in cost, efficiency and speed with automated systems, leaving the door open for the automation of entire labour processes.

Instrumental Rationalisation

Development and application of technologies is highly dependent on the wider power relations, values and ideologies in society. While this is no place to go into the theoretical nitty gritty of it, most critical approaches today acknowledge this fact. The political economy of communication, for example, emphasised the historical interrelation between the US military and industry in the development of ICTs, and how these technologies were remodelled to fit capitalist social relations (Prodnik 2014; Dyer-Witheford et al. 2019, 3; Fuchs 2019). Even if we would disregard the long history of ICTs – which, after all, have to be seen as constitutive for the development of AI – development of algorithms is usually just a means to reach a very narrowly defined end. In other words, it is highly rationalised and instrumentalised. As an example, we can take the digital social media that critical authors view first and foremost as attention machines, aimed at catching and producing consumers. But as noted by Vaidhyanathan (2018, 87), how they work goes beyond distraction and exhaustion. It also dehumanises users, since 'it treats us each as means to a sale rather than as ends in ourselves'.

Algorithms cannot be seen merely as technical artefacts, because this would fail to explain their social role and influence – something I underscored earlier in the chapter. As stressed by Bilić (2018, 60), they must be seen as expressions of a specific technological rationality predominant in capitalism. They are embedded within it 'as a mode of production, a specific form of capitalism–algorithmic capitalism' (ibid.). Other kinds of technological rationalisations are

always possible, but in capitalism imperatives of this system are predominantly imposed on technologies. Examining how search engines are constructed, Mager (2012) for instance noticed that boundaries emerging from capitalist social relations were woven into the practicalities and the operation of algorithms behind them. This produced specific biases and altered the whole digital ecosystem, producing what she called algorithmic ideology.

The capitalist logic can therefore be seen as the main structural reason for instrumental rationalisation, being one of the fundamental characteristics of algorithms (cf. Fuchs 2019, 59). This describes not only the central reason behind this characteristic, but in many ways also key reasons for opacity, datafication and automation. For Fuchs (2009, 8), instrumental reason is 'oriented on utility, profitableness, and productivity', with its objectives reduced to cost-benefit calculations. At least to a degree this is present in all characteristics delineated above, and all are therefore contributing to the intensification of instrumental rationalisation.

The Algorithmic Logic and its Social Consequences

It is possible to identify a range of conceivable consequences resulting from the four characteristics of algorithms. A schematic overview of structural reasons and their social consequences is provided in Table 12.1. The list of consequences is far from exhaustive and their relation to the characteristics may not be as direct as presented. It should, however, capture at least the essential features of what can be called algorithmic logic in digital capitalism.

What I am describing here are tendencies that are real in the abstract, but can in practice be counteracted in various ways, hence forming potential countertendencies that would limit their actual social impact. Social struggles and protests could for instance force governments into political measures that would lead to a shortening of the workday, which in turn could ease the pressure on unemployment; regulation could curb mass surveillance and data harvesting; court decisions could limit the dominant market position of certain corporations and its platforms or put a stop to facial recognition and so on. Various countermeasures are to be expected, but they should not lead us to believe these tendencies were not 'real' or present in the first place (cf. Collier 1994).

Incomprehensibility, Lack of Accountability and Preserving the Status Quo

An important consequence of opacity and obfuscation of algorithms is their incomprehensibility for both lay users and often also for experts. In essence, these are secretive artefacts in more than one meaning of the word, since their complexity is an important and non-intentional contributing factor to

Table 12.1: Algorithmic logic in digital capitalism.

Structural reasons		Basic characteristics	Social consequences	
Intellectual Property Rights		Opacity and obfuscation	Incomprehensibility and secrecy	Reproducing status quo; reinforcing power asymmetries and inequalities
Digital divide (expert illiteracy)			Lack of oversight and control	
Multiplication, layering, constant change			No democratic accountability and legitimacy	
Decisions made via mathematics and rules		Datafication	Intensified quantification and concentrated ownership of data	
Improving capabilities and effectiveness			Mass and ubiquitous surveillance as a norm; privacy breaches	
Predicting trends and exerting control			Constructed and biased processes appearing as objective and neutral	Naturalisation
Increasing competitive advantages		Automation (of processes, functions and decisions)		
Increased efficiency of labour, processes and decisions			Further push in social acceleration	
Better use of existing resources, moving beyond limitations of human beings			Changes in the (re)production of space	
			Wide-ranging effects on employment and labour relations	
Profit seeking; competitiveness; capitalist colonisation of technologies	↓ Opacity	Instrumental rationalisation	Social atomization; commodification; control/domination; reification; alienation	
	↓ Datafication			
	↓ Automation			

their secrecy (Coeckelbergh 2020, Ch. 8). It is often the case that we do not understand how algorithms function and interrelate, what exactly their operation encompasses, what impact they have on our lives and under what conditions this happens. This is why algorithms can lead to results and consequences that might not be intended in the first place and sometimes cannot even be adequately explained.

There have been numerous cases of encoded biases in algorithms such as racist profiling or sexism (Bridle 2018, 142), which were a consequence of comparable biases historically existing in society. The poet Joy Buolamwini, for example, criticized them in a project *AI, Ain't I A Woman* (www.notflawless.ai), which focused on grave failures of facial recognition when it came to black women. A myriad of such incidents demonstrates both that algorithms are far from neutral artefacts, a point I return to later, but also that even their designers in many cases have difficulties understanding why certain results materialised in the first place. In one of the more famous instances, Grindr was linked as a related application to an app which was aimed at finding sex offenders, revolting the LGBT community. What is telling is that this and many other similar examples usually surprised designers of algorithms themselves. Increasingly sophisticated, extensive and complex algorithmic processes mean that 'unintended and unanticipated consequences are an obvious, and will be an increasingly common, outcome' (Willson 2016, 8).

According to Pasquale (2015, 14) strategies of secrecy and obfuscation in algorithms are aimed at the consolidation of power and wealth. This cannot be seen as surprising, since applying intellectual property rights can bring the owners competitive advantages. Many authors have advocated for more transparency as a solution to the problem of algorithms being black boxed, which is a worthy cause. But making them transparent does not in itself bring any meaningful understanding of how they function (Willson 2016; Coeckelbergh 2020, Ch. 8; Obar 2020). Since they are layered and complex systems, these properties represent difficulties even for experts, not to mention activist groups or regulators that would have the capacity to curtail them. Neither does the transparency of algorithms touch on an even graver problem – the commodification and privatisation of data.

Social scientists have started warning about the dangers of algorithmic procedures for democracy, especially when it comes to the influence of the biggest digital social networking sites (Moore 2018; Vaidhyanathan 2018). This happened because nobody beyond their owners has a real oversight over how these algorithms are used, even though they have vast influence over the political process. This lack of accountability can be seen as a fundamental problem, because legitimation is at the core of all publicly relevant decisions in democratic societies (cf. Coeckelbergh 2020, Ch. 10). Pasquale (2015, 16) goes even as far as to claim that 'transactions that are too complex to explain to outsiders may well be too complex to be allowed to exist'. In his opinion the information imbalances have gone too far, particularly since corporations that own algorithms have become the new sense-makers of our world. The Big Data they

collect brings big dangers, with even the smallest of oversights potentially creating life-changing reclassifications in algorithmic decision-making processes (for examples see Eubanks 2017; Coeckelbergh 2020).

What seems apparent, therefore, is that datafication in many ways helps to reproduce or even reinforce the status quo, and with it the existing power asymmetries and social inequalities.

Mass Ubiquitous Surveillance in a World of Privatised Data

It goes almost without saying that a logical consequence of the ever-present datafication is mass and ubiquitous surveillance, with severe breaches of privacy as the final outcome. In the last two decades digital surveillance via various ICTs has practically become a norm, which led to a formation of a whole new research subfield with Surveillance studies. In 2013 this became an even more vigorously debated topic after the Snowden revelations. There is no need to repeat the main arguments of these debates, beyond the fact that digital surveillance opens the door for new ways of sorting, classifying, profiling, segregating and thus also discriminating people, which again reinforces existing inequalities and brings about new social disadvantages (see Prodnik 2014; Mosco 2014; Fuchs 2019).

It is essential to underscore that data is not simply one of the resources in what Srnicek (2017) calls platform capitalism or what Fuchs (2019) defines as Big Data capitalism. It has become *the* resource for major companies, especially in the case of machine learning (Coeckelbergh 2020). This is why datafication – and correspondingly Big Data and mass surveillance – is not simply an optional thing. If you block surveillance the effectiveness of algorithms plummets and many of the existing business models start to collapse. Surveillance and privacy breaches are therefore a necessary part of the algorithmic logic in digital capitalism. They are not a bug but a constituent feature that powers its development.

A continuous push for datafication also brings about a highly unequal concentration of the ownership of the data, which is syphoned off using digital surveillance (cf. Mosco 2014). These information inequalities are even more intensive than in the past, when Perelman (2002, 5) pointed out that 'intellectual property rights have contributed to one of the most massive redistributions of wealth that has ever occurred'. He based this assessment on the fact they were owned almost exclusively by the rich and the powerful. Processes occurring with algorithmic datafication, however, are accentuating and intensifying this problem even further.

Neutrality of Algorithms and their Naturalisation

Various studies have attested to the fact that algorithms are far from neutral technical artefacts (Willson 2016, 9–10). This is both because human biases are present in their development and because they are created with certain

purposes in mind, for example 'to create value and capital; to nudge behaviour and structure preferences in a certain way' (Kitchin 2017, 18). Who creates algorithms and with what underlying aims is far from irrelevant. Facebook's algorithms, for instance, highly value content that arouses strong emotional reactions (Vaidhyanathan 2018), which was not a neutral engineering decision of its creators. While this may make Facebook into a powerful tool for motivation – but especially for grabbing users' attention – it also means it 'is a useless tool for deliberation' (ibid. 132, 144). It mainly sparks shallow declarations and potentially destabilises democratic procedures.

As noted by Diakopoulos (2019, 18) 'the judgments that algorithms make are often baked in via explicit rules, definitions, or procedures that designers and coders articulate when creating the algorithms'. They are of course neither neutral nor objective, but what is true is that 'they will apply whatever value-laden rules they encode consistently' (ibid.). This contributes to the illusion of their neutrality, even though it merely moves discrimination, prejudices, stigmatization and disadvantages upstream (Pasquale 2015, 35).

How Google sorts its search results or how Facebook organises its news feed may seem self-evident and almost natural for their users, a normal order of how things stand, even though it was based on very real human decisions of how these platforms present and sort content. Many of our activities and practices of course become naturalised when they become part of our everyday routines and we accept them without necessarily questioning the power relations constitutive for them (Willson 2016, 2). It would indeed be impossible to live our lives if we always scrutinised every step we took, even the most mundane ones. However, this is not the only reason for naturalisation of algorithms; both datafication and automation are contributing to the fact that algorithmic decisions appear neutral. They are based on objective calculative procedures, which indeed have no intrinsic biases in themselves. This 'mathematical, computational and rational design', which is necessary for algorithms and is acquired through datafication, creates 'an aura of universality of reason, an aura of calculable, efficient and truthful solutions to given problems' (Bilić 2018, 59). Since these decisions are simultaneously also automated, they obtain what could be called epistemic purity, and with it a halo of authority (Diakopoulos 2019, 118). This can be related to a phenomenon called automation bias, in which automated procedures are perceived as more trustworthy than nonautomated ones or even our own experiences (Bridle 2018, 40). This is particularly true in case of ambiguous situations, since 'automated information is clear and direct, and confounds the grey areas that muddle cognition' (ibid.).

Temporal and Spatial Changes

Automation will also produce noticeable changes in temporal compression and the way space is (re)produced. When processes, decisions and functions

increasingly become automated, they also get accelerated. Especially in the case of intangibles, the level of acceleration facilitated by algorithms cannot be measured only quantitatively. The change is primarily qualitative in nature, because it leads beyond limitations inherent to humans. The most obvious example is High-Frequency Algorithmic Trading in the financial markets, which is highly unstable and has been largely automated, with human traders becoming more or less obsolete. Decisions are now made in microseconds, leading to 'one of the most dramatic increases in speed in recent times', going 'beyond those perceptible by human beings' (MacKenzie 2017, 55; cf. Pasquale 2015, 128–132; Wajcman 2015, 17–21).

Nevertheless, acceleration in trading cannot be explained solely with technological advances in algorithms. It was a result of carefully planned decisions at the time these algorithms were designed, with speed purposively at the core of how they function (see MacKenzie 2017). It would therefore be both theoretically and empirically wrong to make a direct causal connection between acceleration and changes in technologies, as if the latter were constructed in a social vacuum. As emphasised by Wajcman (2015, 3), 'temporal demands […] are built into our devices by all-too-human schemes and desires'.

In Rosa's (2013) general theory of modernity, social acceleration is a constitutive and unavoidable part of modern societies, but technological acceleration is only one of the three dimensions in what he calls the acceleration-cycle. The other two are acceleration of social change and acceleration of the pace of life. Technological acceleration is indeed based on technological innovations like algorithms, with competition providing incentives for their development and adoption (what Rosa calls the external economic motor). However, in isolation, technological acceleration could not by itself lead to social acceleration. In most cases new technologies enable us to save time and should therefore – if anything – contribute to a general deceleration. It is only in relation to the other two dimensions and the fact we live in a competitive (capitalist) society that technological breakthroughs in fact lead to social acceleration (ibid.).

In a similar manner, algorithms may actually slow down the way certain sectors function. MacKenzie (2017, 57–58), for instance, discovered that work in the trading sector has slowed down considerably. It became much less hectic, but this was down to the fact that the work itself changed completely. It was not performed by human traders anymore, but by programmers that developed algorithms. Even with such contradictory examples, the general effect of the adoption of algorithms will almost surely be further social acceleration, in line with other similar technological advances.

Algorithms are also changing public and private spaces, and how we perceive and interact with them (Mittelstadt et al. 2016, 1). Algorithms are at the core of smart cities, they are creating new knowledge about space, they are (re)directing traffic, procuring navigation and rewriting how we understand certain geographical locations (Fisher 2020). Alexa's algorithms are, for example, reshaping how we live in our private homes, while Airbnb is fundamentally transforming

how people see their dwellings, simultaneously changing city geographies (Munn 2018). In essence, algorithms are already remodelling time and space configurations.

(Un)employment and Automation

Several studies are warning that the current pace of automation could have a serious impact on future unemployment and global labour markets. It is expected that a combination of algorithms, robotics and computers will increasingly make human labour redundant, even without development of Artificial General Intelligence (Coeckelbergh 2020, 136–144). There are many technical problems connected to automation, but they are slowly being overcome with machine learning and by making simple tasks even simpler. This solution was already used in factory automation during the industrial revolution, when previously unstructured tasks were subdivided and simplified. Whereas there is certainly a lot of unwarranted hype connected to algorithms and AI, a long history of technological innovations, identified already by Marx (1867/1990, 562–563), attests to capital's constant tendency to make labour superfluous through automation. As noted by Dyer-Witheford et al. (2019, 4) the 'dismissal of automation as a "charade" is deeply ahistorical'. In the past, 'capital has made people and indeed entire populations disposable'.

A research paper by Frey and Osborne, published in 2013, for example, tried to estimate the probability of computerisation for 702 detailed occupations in which 97 per cent of the American workforce was employed at the time (Frey 2019, 319). They estimated that nearly half of all employments were at risk, with low-income jobs that required lower education to perform hit the hardest (ibid. 319–321). Frey (ibid. 322) analysed other studies and they concurred it was especially unskilled jobs that were most exposed to the risk of automation. A policy brief by OECD (2018) forecasted less drastic impact of automation, with 14 per cent of the jobs in OECD countries highly automatable and 32 per cent facing substantial change in how they are done. But their analysis also warns that the tasks AI cannot do are rapidly shrinking, with some jobs becoming entirely redundant (ibid.).

It is unlikely all occupational areas will go through such a radical transformation in the mid-term as jobs in financial trading (MacKenzie 2017), but it seems that only a few will remain unaffected (Frey 2019, Pt. 5). While estimates regarding the proportion of occupations under direct threat remain speculative and vary because of differences in methodologies, it is highly doubtful they will all be offset by completely new occupations. Collins (2013) is amongst the authors that are convinced capitalist societies are facing the end of the middle-class work as we knew it because of technological displacement. He predicts even starker inequalities. Considering how deeply unequal societies today are,

and how uneven ownership of the algorithmic means of production is, we have every reason to be sceptical that the benefits of these processes will be evenly shared by the majority of the population.

Conclusion: Algorithmic Necessity?

Once it is formed, a system takes on a life of its own.
– Haruki Murakami (1Q84)

In a growing number of social domains decisions are influenced or directly made by algorithms. It remains to be seen how far reaching their influence will be in the long-run, but it seems increasingly likely that different corporate actors and state institutions will either adopt algorithms or use them even more widely than they currently do. This tendency can be called *algorithmic necessity*, indicating that it is increasingly inevitable that different institutions will employ algorithms. Their adoption can have significant advantages on the market or can help to 'rationalise' administrative functions, which is always portrayed as a worthy cause in the neoliberal state. Non-adoption can similarly bring disadvantages, as companies that are incapable of innovation fall behind their competitors or simply fail to meet their quarterly goals. When one company uses large quantities of personal data to improve their algorithms in an attempt to gain a competitive edge, others are likely to follow, which forms an almost self-propelling cycle.

What Marx (1867/1990, 433) called 'the coercive laws of competition', this iron cage of capitalist society, will therefore have direct influence on the general expansion of algorithms and how they are developed. Competition between different capitals that are structurally forced to constantly increase their accumulation, for example, pushes them into technological innovation (cf. Streeck 2012, 5). With algorithms, this can lead to increases in productivity (preferably through automation), improvements in efficiency, or speeding up of the circulation of capital. As noted by Wajcman (2015, 17), 'the faster that money can be turned into the production of goods and services, the greater the power of capital to expand or valorize itself. With capitalism, time is literally money, and "when time is money, then faster means better" and speed becomes an unquestioned and unquestionable good'.

The mythological aspects of implementing technological innovations should not be overlooked either, even in the case when they might not be economically rational at all. It is easy to simply dismiss the hype surrounding technological breakthroughs, but in Mosco's (2014, 5) view, such appraisals are mistaken: 'The marketing hype supports myths that are taken seriously as storylines of our time. If successful, they become common sense, the bedrock of seemingly unchallengeable beliefs'. Socially dominant myths acquire their own power and tend to become self-fulfilling prophecies.

In digital capitalism the implementation of algorithms follows the logic of instrumental rationalisation that produces 'irrational results' and 'impoverishes human experience' (Bilić 2018, 59–60). Authors of the Frankfurt School closely related instrumentalisation to the development of capitalism and the predominance of economic rationality in this system. They warned that intensification of these processes will lead to further social atomization, reification, domination and alienation. These are some of the most fundamental consequences of algorithms as artefacts of digital capitalism.

These critical observations should not be taken as some Luddite rejection of technological progress, where the only path is either acceptance of algorithms or their complete rejection. Instead, there is no doubt that algorithms of a different sort can serve democratic means, reduce human toil, reduce inequalities and help to bring about overall improvements in the quality of our lives. But this presupposes their fundamental reimagining in how they are made and for what purposes, together with political struggles that take into account the fact they can – and should – be changed if this is to happen. And this cannot be done without a change in who has control and ownership over these systems. In other words, this presupposes social relations that go beyond those imposed by digital capitalism.

Funding

This research was supported by the 'New Modes and Global Patterns of Online News (Re)production' project (N5-0086) funded by the Slovenian Research Agency (ARRS).

References

Bilić, P. 2018. The Production of Algorithms and the Cultural Politics of Web Search. In: P. Bilić, J. Primorac and B. Valtýsson (Eds.), *Technologies of Labour and the Politics of Contradiction*, pp. 57–76. Basingstoke, New York: Palgrave Macmillan.

Boden, M. A. 2016. *AI: Its Nature and Future*. Oxford: Oxford University Press.

Bridle, J. 2018. *New Dark Age: Technology and the End of the Future*. London: Verso.

Bucher, T. 2017. The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms, *Information, Communication & Society*, 20:1, 30-44, DOI: 10.1080/1369118X.2016.1154086.

Coeckelbergh, M. 2020. *AI Ethics*. Cambridge, MA: The MIT Press.

Collier, A. 1994. *Critical Realism*. London, New York: Verso.

Collins, R. 2013. The End of Middle-Class Work: No More Escapes. In: I. Wallerstein, R. Collins, M. Mann, G. Derluguian, C. Caljhoun, *Does Capitalism Have a Future?*, pp. 37–70. Oxford, New York: Oxford University Press.

Diakopoulos, N. 2019. *Automating the News*. Cambridge, MA: Harvard University Press.
Dyer-Witheford, N., Kjøsen, A. M. and Steinhoff, J. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. London: Pluto Press.
Eubanks, V. 2017. *Automating Inequality*. New York: St. Martin's Press.
Fisher, E. 2020. Do Algorithms Have a Right to the City? *Cultural Studies*. DOI: https://doi.org/10.1080/09502386.2020.1755711
Frey, C. B. 2019. *The Technology Trap: Capital, Labor, and Power in the Age of Automation*. Princeton, NJ: Princeton University Press.
Fuchs, C. 2009. A Contribution to Theoretical Foundations of Critical Media and Communication Studies. *Javnost–The Public*, 16(2), 5–24.
Fuchs, C. 2019. Karl Marx in the Age of Big Data Capitalism. In: D. Chandler and C. Fuchs (Eds.), *Digital Objects, Digital Subjects*, pp. 53–71. London: University of Westminster Press.
Fuchs, C. and Mosco, V. (Eds.). 2015. *Marx in the Age of Digital Capitalism*. Leiden: Brill.
Gillespie, T. 2014. The Relevance of Algorithms. In: T. Gillespie, P. J. Boczkowski and K. A. Foot (Eds.), *Media Technologies*, pp. 167–193. Cambridge, MA: The MIT Press.
Kitchin, R. 2017. Thinking Critically About and Researching Algorithms. *Information, Communication & Society*, 20(1), 14–29.
MacKenzie, D. 2017. Capital's Geodesic. In: J. Wajcman and N. Dodd (Eds.), *The Sociology of Speed*, pp. 55–71. Oxford: Oxford University Press.
Mager, A. 2012. Algorithmic Ideology. *Information, Communication & Society*, 15(5), 769–787. DOI: https://doi.org/10.1080/1369118X.2012.676056
Marx, K. 1867/1990. *Capital: A Critique of Political Economy, Volume One*. London: Penguin Books.
Mitchell, M. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. Penguin: London.
Mittelstadt, B. D. et al. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society, 3(2)*. DOI: https://doi.org/10.1177/2053951716679679
Moore, M. 2018. *Democracy Hacked: Political Turmoil and Information Warfare in the Digital Age*. London: Oneworld Publications.
Mosco, V. 2014. *To the Cloud: Big Data in a Turbulent World*. Boulder, CO: Paradigm.
Munn, L. 2018. *Ferocious Logics: Unmaking the Algorithm*. Lüneburg: Merson Press.
Obar, J. A. 2020. Sunlight Alone is Not a Disinfectant. *Big Data & Society,*. 7(1). DOI: https://doi.org/10.1177/2053951720935615
OECD. 2018. Putting Faces to the Jobs at Risk of Automation, March. Retrieved from: www.oecd.org/employment/Automation-policy-brief-pdf, 20 June 2020.
Pasquale, F. 2015. *The Black Box Society*. Cambridge, MA: Harvard University Press.

Perelman, M. 2002. *Steal This Idea*. New York: Palgrave.
Pew Research Center. 2018. Many Facebook Users Don't Understand How the Site's News Feed Works. Retrieved from: www.pewresearch.org/fact-tank/2018/09/05/many-facebook-users-dont-understand-how-the-sites-news-feed-works, 20 June 2020
Prodnik, J. A. 2014. The Brave New Social Media. *Teorija in praksa*, 51(6), 1222–1241.
Prodnik, J. A. 2016. 3C: Commodifying Communication in Capitalism. In: C. Fuchs and V. Mosco (Eds.), *Marx in the Age of Digital Capitalism*, pp. 233–321. Leiden: Brill.
Rosa, H. 2013. *Social Acceleration*. New York: Columbia University Press.
Srnicek, N. 2017. *Platform Capitalism*. Cambridge: Polity.
Streeck, W. 2012. How to Study Contemporary Capitalism? *European Journal of Sociology*, 53(1), 1–12.
Vaidhyanathan, S. 2018. *Antisocial Media*. New York: Oxford University Press.
Wajcman, J. 2015. *Pressed for Time*. Chicago, IL: The University of Chicago Press.
Willson, M. 2016. Algorithms (and the) Everyday. *Information, Communication & Society*, 20(1), 137-150 DOI: https://doi.org/10.1080/1369118X.2016.1200645

CHAPTER 13

'Not Ready for Prime Time': Biometrics and Biopolitics in the (Un)Making of California's Facial Recognition Ban

Asvatha Babu and Saif Shahin

Introduction

On 8 October 2019, Governor Gavin Newsom signed a law forbidding California police departments from using facial recognition (FR) software on body cameras. The decision was welcomed widely, especially by civil society groups that have long called for outlawing 'an invasive and dangerous tracking technology that undermines our most fundamental civil liberties and human rights' (ACLU 2019a). AB-1215, or The Body Camera Accountability Act, came on the heels of bans on government use of FR in five US cities earlier that summer: San Francisco, Berkeley, and Oakland in California, and Cambridge and Somerville in Massachusetts (Cagle 2020).

FR is a form of biometric artificial intelligence that involves 'the automated process of comparing two images of faces to determine whether they represent the same individual' (Garvie, Bedoya and Frankle 2016, 9). Attempts to use computers to identify human faces go back at least half a century (Goldstein, Harmon and Lesk 1971) and FR technology has become commonplace in recent years. We use it every day to unlock our mobile phones or tag friends on social media. Companies employ it to improve user profiles for targeted

How to cite this book chapter:
Babu, A. and Shahin, S. 2021. 'Not Ready for Prime Time': Biometrics and Biopolitics in the (Un)Making of California's Facial Recognition Ban. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 223–245. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.m.

advertising. Law enforcement agencies mostly rely on FR for two purposes: *face verification* to confirm a claimed identity and *face identification* to 'identify an unknown face' (Garvie, Bedoya and Frankle 2016, 10).

Although both are contentious, it is the latter application that has raised the most eyebrows. Back in 2016, the Georgetown Law Center on Privacy & Technology reported that FBI searches for identity using FR were 'more common than federal court-ordered wiretaps' (Garvie, Bedoya and Frankle 2016, 25). The faces of nearly 117 million Americans were already in federal law enforcement databases and every other American adult had had their photos searched in this manner. Law enforcement agencies in the United States have dabbled in FR-based surveillance projects since 2001 (Gates 2011). Now, with advances in technology, police departments in various cities in the US, in conjunction with technology corporations, have initiated FR-based surveillance programs that use existing infrastructure like CCTV cameras such as Detroit's Project Greenlight, under which 'cameras all over the city keep an eye on the populace' (Colaner 2020). Others are experimenting with using FR on body cameras and mobile phones (Naughton 2020). The US Customs and Border Protection agency began deploying FR cameras at airports in 2017 (Oliver 2019) while immigration officers started running FR searches on driving license photos to identify undocumented immigrants since at least 2019 (Chappell 2019).

There is no federal law regulating the collection and use of biometric data in the United States. Illinois, Texas and Washington are the only states so far to pass comprehensive legislation regulating state and private collection of biometric information (Pope 2018). At the same time, facial data has become ubiquitous on the internet with the sharing of photos and videos on social media. An ecosystem of businesses, such as the controversial Clearview AI, have sprung up that take advantage of this ubiquity and regulatory lag to build massive databases and cheap tools for the state to use (Mann and Smith 2017; Naughton 2020; Hill 2020; Kak 2020).

California's AB-1215 law, which came into effect in January 2020, is part of a slew of local, state and federal attempts to check this proliferation. In 2019, San Francisco, home to global tech giants and one of the most technologically advanced cities in the world, became the first US city to ban FR use by law enforcement. It was quickly followed by Somerville, Massachusetts and Oakland, California (Metz 2019). In October 2019, California became the third state to issue a ban, following Oregon and New Hampshire (Thebault 2019). As the namesake of the Californian Ideology (Barbrook and Cameron 1996), an uncritical 'technological solutionism' (Morozov 2012) and belief in 'dotcom' neoliberalism, it is particularly interesting that California has been one of the earliest movers in attempting to regulate this emerging technology.

In the wake of nationwide – indeed global – protests after the killing of 46-year-old George Floyd in May 2020 by the Minneapolis Police, reports of FR use by law enforcement to identify and arrest protestors has further fuelled

demands to remove the technology from the arsenal of law enforcement agencies around the country (Colaner 2020). In June 2020, the Detroit police came under fire for arresting an innocent Black man after FR technology flagged him as a shoplifting suspect. As activists mounted pressure on the city council to reject the proposed extension of the police FR contract, the Detroit Police Chief admitted that the FR system fails to accurately identify faces approximately 95% of the time (Cameron 2020; Colaner 2020; Ferretti 2020). Around the same time, Boston became the biggest city on the US East Coast to ban FR for municipal use. Two US senators have proposed federal legislation calling for 'a full stop to facial recognition use by the government' at all levels nationwide (Ng 2020b). Even companies such as Amazon, IBM and Microsoft announced that they would not be selling the technology to law enforcement (NPR 2020).

And yet, it may still be too early for critics of FR and other forms of algorithmic surveillance to rejoice. Widespread calls to abolish carceral technologies and practices that disproportionately affect Black Americans (Benjamin 2016) have put pressure on private industry, not necessarily to stop building FR for the government but to at least manage the optics of their involvement. Seen in that light, some of these moves appear to be little more than short-lived attempts at corporate image management. Amazon, for instance, made it clear that its moratorium on the sale of its FR tool Rekognition to law enforcement would only be for a year.

Without the visibility provided by these protests and the sustained pressure of activists, it is possible that some of these uses may have gone on unchecked and unexamined. Indeed, many of these companies continue to sell FR technology to governments outside of the United States where there is little pressure from civil rights groups or movements like Black Lives Matter (Barik 2020). Documents released a week after Microsoft's announcement revealed it had previously been trying to sell FR technology to federal agencies without any regard for human rights, contrary to sentiments expressed by the company's president (Ng 2020a).

Even AB-1215 may not be the 'victory' (Guariglia 2019) it seems at first glance. Initially intended to prohibit Californian police's use of FR on body cameras permanently, the bill was gradually diluted and defanged. The version that was passed into law proscribes FR use for only three years. In this chapter, we adopt a *law and society* approach – which views legislation as a social phenomenon (Ewick and Silbey 1998) – to explain how and why this came about. In this approach, researchers study law and *legality* as situated in society and laced through its culture by examining both the 'formal and informal settings where legal activity – in all its guises – may unfold' (Seron and Silbey 2004, 30). Our study traces the trajectory of AB-1215 from its introduction as a 'spot bill' on 21 February 2019 through its signing into law eight months later – both as legislative action within the state assembly and as public deliberation outside the corridors of the capitol. We critically examine legal documents as well as

reports from the civil society and news media as a social discourse with a view to understanding (1) how FR technology was perceived and presented by different stakeholders, specifically in terms of its benefits and harms and (2) what factors and actors contributed to the diminution of the law.

We next discuss a range of socio-political concerns raised by FR and outline a conceptual framework for thinking about FR regulation in the light of these concerns. Then, following a brief discussion of the passage of AB-1215, we turn attention to our empirical analysis of the bill as a social discourse. We conclude with an assessment of how and why the bill was defanged and consider how it can inform future research and resistance against algorithmic governance.

Inaccuracy and Mass Surveillance

The application of FR for law enforcement raises three interrelated concerns. The first is the *inaccuracy* of the technology itself, or the possibility that an individual is not who the FR algorithm identifies them as. That is because 'biometric recognition is an inherently probabilistic endeavour' (Pato and Millett 2010, 9). FR technology doesn't *see* faces – the way humans do – but offers probable *matches* based on geometric representations of facial features. Making computers 'see' – i.e. engineering computer vision – is a difficult task that scientists and engineers have mulled over for decades. Advances in the ability to store and crunch large amounts of data as well as in machine learning algorithms have led to breakthroughs in recent years (Demush 2019). At first, the algorithm is trained to recognise which facial features are more likely than others to indicate similarity by analysing different images of the same person from a large training data set. This is known as machine learning. It then applies this learning to identify images of the same individual in real-life law-enforcement scenarios (Huang et al. 2008). But this machine learning involves 'millions of variables' – such as lighting, picture quality, and subject distance, for example – and is never perfect (Garvie, Bedoya and Frankle 2016). The FR algorithm, therefore, does not produce the 'right' match for an image but a series of more or less likely matches.

In 2018, ACLU – the American Civil Liberties Union – tested the accuracy of Amazon's Rekognition, a top FR platform used by government agencies throughout the United States. The FR tool failed spectacularly: 'the software incorrectly matched 28 members of Congress, identifying them as other people who have been arrested for a crime' (Snow 2018). ACLU repeated the test in August 2019. This time, Rekognition misidentified 26 California legislators as criminals – among them Phil Ting, the author of AB-1215 (Chabria 2019). This lack of accuracy means that when law enforcement uses FR uncritically, many innocent people could end up on their radar, go to jail, be deported or, in countries such as the US, face the death penalty.

The second concern is their propensity for *mass surveillance and violation of privacy*. This happens in two ways. One, law enforcement agencies collect photographs of people to populate their FR databases – and they do so indiscriminately. Referring to the FBI program, the Georgetown Law report noted, 'Never before has federal law enforcement created a biometric database – or network of databases – that is *primarily* made up of law-abiding Americans' (Garvie, Bedoya and Frankle 2016, 20, emphasis added). So far, these photographs have been mined from people's driving licences, ID cards and even social media accounts. As FR is employed via cameras at airports, on police personnel and so on, people will be subjected to facial 'tracking from far away, in secret' and en masse (10). This makes it nearly impossible for people to manage their boundaries, a key practice in maintaining dynamic privacy relationships. This, in turn, makes it difficult to practice critical citizenship and aspire to the ideals of liberal democracy (Cohen 2012).

Two, a significant feature of 'high-dimensional' – or individualised – data collection is that it allows cross referencing of multiple data sets (boyd and Crawford 2012). In other words, information about an individual in one data set can be used to find more about that individual by linking it with other data sets. So, while law enforcement agencies might compile FR databases putatively for identifying individuals in situations where a law is violated, they could easily use the photographs to track various other activities of the individual – including activities that may be perfectly legal but politically undesirable for authorities, such as participating in protests against police violence. As was revealed by NSA whistle-blower Edward Snowden, this information can be easily obtained by the US government through social media and phone records (Greenwald, MacAskill and Poitras 2013). When this issue is coupled with FR's inaccuracy, it means that '[e]ven if you're sitting at home on your couch, there's a chance you could be arrested for protesting' (Shwayder 2020).

Biopolitics of Facial Recognition

A third problem with FR systems is their proclivity to perpetuate and amplify *social discrimination* against marginalised communities. This is only partly a consequence of its technical flaws, which are also not arbitrary: 'if a training [data] set is skewed towards a certain race, the algorithm may be better at identifying members of that group as compared to individuals of other races' (Garvie, Bedoya and Frankle 2016, 9). The Silicon Valley's whiteness and apparent 'colour blindness' means FR training data are overwhelmingly White and therefore several times more likely to misidentify Black people (Buolamwini and Gebru 2018; Simonite 2019) – driving up their already disproportionate rates of incarceration.

But social discrimination is not solely a function of machine learning or technology design. Immigrants and minorities – Black, Latinx and Muslim people

in particular – are already much more likely than White people to be singled out for surveillance via FR at the level of policymaking (Bedoya 2019) and disproportionately represented in law enforcement and intelligence watchlists (Devereaux 2019). Browne (2010) argues that the practice of making a body visible – or 'legible' – has always been an exercise in power, with political and economic ramifications. The branding of enslaved Africans on American plantations, for instance, was a means of 'accounting and of making the already hyper-visible body legible' – not exactly the same, but also not altogether different from the contemporary practice of making 'bodies informationalized by way of biometric surveillance' (139). Biometric technology reimagines the body as flows of data and patterns of communication (van der Ploeg 2002). Objectified and digitized 'individuals are broken down and reinterpreted in terms of the information provided by their body, instead of as agential social beings' (Hood 2020, 158). These data points are logged in a virtual register, making the bodies themselves legible, accountable, and thus controllable (Andrejevic, 2019).

Considered from this perspective, FR is only the latest chapter in a long history of authorities using technology to subjugate, racialise and dehumanise people by acting upon their bodies – the newest arena of 'biopolitics'. A biopolitical view (Foucault 2003) brings to surface the *systemic* nature of social discrimination. In this view, the technological inaccuracy of FR is itself a consequence of institutionalised racism and classism – evident in everything from education to hiring practices to law enforcement – that keeps marginalised bodies outside of Silicon Valley offices and inside of prisons. This view is at once micro and macro: it segues from the datafication of human body to map the geography of social belonging that such data enables. Gandy Jr. (1993) showed how the economic value of a person's data differed depending on the social group they belonged to and how this differentiation reproduced social inequalities. As he noted, surveillance goes beyond social control and into the realm of sorting and differentially targeting people based on their positionality in the socio-economic hierarchy.

Petit (2017) makes a distinction between discrete and systemic 'externalities' that accrue from artificial intelligence systems. Externalities could be either harms or benefits to third parties. *Discrete externalities* are 'personal, random, rare or endurable' (26). They take place at the level of the individual, may affect anyone with an equal chance, are low in frequency and neither ruin nor radically improve the affected individual's life. Examples include a malfunctioning robot mistaking a garden-variety rodent for a parasite and spraying it with pesticide. *Systemic externalities* are 'local, predictable, frequent or unsustainable' (26). In other words, they are foreseeable, take place repeatedly, affect 'a non-trivial segment of the population' and can cause a long-term 'reduction or increase in well-being of the local population class under consideration' (26). For instance: China's reported use of automation and cutting-edge technological tools to surveil and detain its largely Muslim Uighur population (Taddonio 2019).

This distinction helps us think normatively about AI regulation. Petit (2017) recommends that discrete externalities 'should be left to the basic legal infrastructure' (28). That is, emergent problems could be resolved on a case-by-case basis in an *ex-post* manner – or after they occur – by applying specific laws that are already in place. But systemic externalities require *ex-ante* consideration. As they are not only predictable but also significant in the scale of their impact, their repercussions need to be anticipated and lawmakers ought to institute regulations to mitigate the harms they might cause.

Are FR's externalities discrete or systemic? A purely technological view that is restricted to FR's inaccuracy would consider them to be discrete – random, rare and occurring at the level of the individual. But a biopolitical view, as outlined above, enables us to see that FR's externalities are in fact *systemic* in nature – causing frequent and permanent harms to large and identifiable populations. FR therefore requires ex-ante regulation that anticipates these harms and prevents them.

A Brief History of AB-1215

In a blogpost about ACLU's 2018 test on Rekognition which demonstrated its fallibility, Jacob Snow, Technology and Civil Liberties Attorney with the ACLU, wrote: 'These results demonstrate why Congress should join the ACLU in calling for a moratorium on law enforcement use of face surveillance' (Snow 2018). This test from ACLU, combined with mounting academic research on the discrete and systemic harms of FR, has become a cornerstone for calls to ban the use of facial recognition technology by government agencies and law enforcement organisations (Metz 2019). AB-1215, in California, was one such call.

AB-1215 was signed into law in October 2019 and came into effect in January 2020 for a three-year period. It states that: '[a] law enforcement agency or law enforcement officer shall not install, activate, or use any biometric surveillance system in connection with an officer camera or data collected by an officer camera' (Body Camera Accountability Act 2019e, 3).

The spot bill for AB-1215 was introduced in February 2019 by Democratic member of California's State Assembly, Phil Ting. With the support of ACLU of Northern California and several other civil society organisations, Ting introduced the first substantive draft of the bill in the State Assembly on 8 April. This version of the bill was considerably different from the final version that would eventually be passed. For instance, it prohibited the installation, activation or use of biometric surveillance in connection with an officer camera *indefinitely*. It also made the law enforcement agency or official liable for damages up to US$4000 in addition to attorneys' fees (Body Camera Accountability Act 2019a). The bill made its arguments for a ban by drawing attention to the threat to civil liberties and the constitutional right to privacy and anonymity posed by FR, the possible chilling effect on free speech in public spaces,

FR's lack of accuracy in identifying people of colour and women, and the disproportionate impact of this technology on overpoliced communities. The bill also repudiated law enforcement's co-optation of tools meant to ensure their accountability (body cameras) into tools of dragnet surveillance (Body Camera Accountability Act 2019a).

On 23 April, the bill was debated and amended in the Assembly Public Safety Committee with the recorded support of 24 civil society organisations (including the ACLU of Northern California) and opposition from law enforcement groups including the California Police Chiefs Association and California State Sheriffs' Association. It passed through to the next stage with an amendment removing the specific amount that had to be paid in damages. When it passed with a 45–17 vote in the California State Assembly on 25 April, the ban on the use of FR in police body cameras was still indefinite (Body Camera Accountability Act 2019b).

In June, it was taken up in the California State Senate Committee on Public Safety, where it passed 5–2 without amendment. A couple of months later, while the bill was still on the backburner at the Senate, the ACLU of Northern California ran its second test using Rekognition which matched 26 California lawmakers including Phil Ting with criminal mugshots (Gardiner 2019a). Almost two weeks later, on 27 August, the bill was amended to include a sunset clause that repealed the bill on 1 January 2027. The amendments also excluded from the ban 'internal editing procedures for redaction purposes'; and the 'lawful use of mobile fingerprint scanning devices' (Body Camera Accountability Act 2019c, 3–4). In September, at its final reading before passage, the bill was amended to shorten the sunset period from seven to just three years and was set to expire on 1 January 2023 (Body Camera Accountability Act 2019d).

Thus, when the bill was signed into law, it was not so much a *ban* on the use of FR by law enforcement as a relatively short *moratorium* on a very specific use of the technology. While this was still considered a win by civil rights groups around the country, ACLU and Assembly member Ting's original intent to keep FR out of the arsenal of law enforcement in perpetuity was thwarted. How, and why, did this happen? And what can we learn from this qualified success to guide future attempts for regulating FR and other forms of algorithmic governance?

Research Design

To answer the questions above, we used the law and society approach (Seron and Silbey 2004) to track AB-1215 as social discourse, taking place both within and without the corridors of the state assembly, involving lawmakers as well as civil society groups, technology companies, police unions and the mass media. The passage of AB-1215 was a months-long process that involved these multiple stakeholders engaged in shaping the conversation

through reports, blog posts, press releases, media presence and participation in the legislative process. These artefacts offer important insight into the way the social discourse surrounding FR and its regulation evolved. Keeping in mind our objective to track this evolution, we conducted a qualitative content analysis of these artifacts.

We triangulated data from three sources: legislative documents such as bill and floor analyses; communication materials from organisations listed in legislative documents as supporting or opposing the bill; and local news coverage on AB-1215 between February (the introduction of the spot bill) and October 2019 (the signing of AB-1215 into act by the California governor). Although not exhaustive, these sources provided us with a comprehensive and diverse collection of stakeholders and their contributions to the conversation around AB-1215. They allowed us to examine the stakeholders' primary participation in the discourse through their own publications and statements as well as what was picked up by the media.

News reports were collected and combined from three databases – Access World News, LexisNexis and Factiva – using the search phrase: ['ab 1215' OR 'ab1215' OR 'ab-1215' OR 'a.b. 1215' OR California AND ('face recognition' OR 'facial recognition') AND ('police' OR 'law enforcement')]. Only articles published in the United States between 1 February 2019 and 31 October 2019, mentioning these terms in the headline or lead paragraphs were included in the corpus. Next, legislative documents were downloaded from the official California Legislative Information website. This included: the multiple versions of the bill from 21 February to 8 October (7 documents), bill and floor analyses at each stage (7 documents), and vote information. Finally, we manually downloaded any communications material mentioning 'AB-1215' from the websites of the civil society organisations mentioned as supporting and opposing the legislation. If there was no mention of the bill, we chose any materials that discussed 'facial recognition' within our target dates. This yielded a total of 38 documents for analysis. After discarding duplicate and irrelevant articles from the media coverage, we had a total of 148 documents (96 news articles, 14 legislative documents and 38 outreach materials).

Following Mayring (2004), analytical criteria were determined based on our research question as: arguments made in support of facial recognition use by law enforcement; arguments made against facial recognition use in body cameras by law enforcement; actors making these arguments; and social and political values present. These documents were then read line-by-line to develop inductive codes. In the first cycle of coding, we used the 'in-vivo' or 'literal' coding technique (Saldaña 2009) wherein exact phrases from the actual language of the documents can be used as code. According to Charmaz (2006), this technique helps preserve the meaning of actors' views and actions within the code, making it easier to analyse while and after coding. Using this technique, we developed a set of codes that were arguments for and against the use of facial recognition on police body cameras. We also noted the actors

making these arguments. In the second cycle of coding, similar in-vivo codes were grouped together under a summative 'value' to discern patterns in the data. Taebi et al. (2014), in their study of public values in technology and innovation, cite Talbot (2011) to suggest that the term refers to the public view of what may be considered valuable or worth striving for.

AB-1215 as a Social Discourse

The coding, categorization and organisation of data revealed that the main arguments for and against the use of facial recognition in police body cameras in California could be filed under three themes: (1) privacy, surveillance and liberty; (2) public safety; and (3) discrimination as technological artefact. Interestingly, both the pro- and anti-FR factions were able to use the concepts underlying these three themes to their advantage in their arguments. These themes are discussed below.

Media discussions about the bill were dominated by civil society organisations (especially ACLU), and lawmakers (especially Assembly member Phil Ting, the author AB-1215), followed by representatives of police unions and organisations. To a much lesser extent, comments and statements from technology manufacturers such as Amazon, Axon and Microsoft were also featured in news coverage. Most media coverage was local, with city and county news organisations contributing the bulk of the articles in the corpus. A lot of the coverage was based almost entirely on press statements from Assembly member Ting, congressional and senate floor analyses and press statements from organisations with involvement in the political process such as ACLU, Electronic Frontier Foundation, Fight for the Future, or, from the other side, the Riverside Sheriffs' Association, the California Peace Officers' Association and the California Police Chiefs Association. As a result, arguments tended to repeat themselves across the dataset. But as our corpus was multi-pronged, this enabled us to understand what arguments were more likely to be picked up from legislative debates and outreach materials and become popularised in the media – and what arguments were not. In addition, we were able to track how the discourse changed over time.

Privacy, Surveillance and Liberty

Fears that arming police body cameras with FR systems would lead to mass surveillance and intrusions of privacy were prominent in statements from civil rights groups and lawmakers supporting AB-1215, especially in the weeks and months following the introduction of the bill. For instance, Ting, the assembly member who wrote the bill, said on 9 May, when the bill went to the Senate: 'Without my bill, face recognition technology can subject law-abiding citizens to perpetual police line-ups, as their every movement is

tracked without consent. Its use, if left unchecked, undermines public trust in government institutions and unduly intrudes on one's constitutional right to privacy' (Office of Assembly member Ting, 2019a). Similarly, ACLU representatives repeatedly brought up the 'invasive' nature of FR. Matt Cagle, an attorney for ACLU of Northern California, said, 'AB-1215 helps ensure Californians don't become test subjects for an invasive and dangerous tracking technology that undermines our most fundamental civil liberties and human rights' (ACLU 2019a).

The idea that FR would violate not some arbitrary notion of privacy but the constitutionally ordained rights of American citizens was important in these claims. The Fourth Amendment to the US Constitution protects citizens from unreasonable search and seizures. In effect, this allows citizens to be present in public without having to show any form of identification to authorities (Body Camera Accountability Act: Hearing 2019a). This protection would vanish with the widespread adoption and use of FR by law enforcement to scan civilians on the street. As described in the text of AB-1215, this is the 'functional equivalent of requiring every person to show a personal photo identification card at all times in violation of recognized constitutional rights'. (Body Camera Accountability Act: Hearing 2019a, 3).

In addition, critics warned of the technology's potential to 'chill' free speech in public spaces. The case of China was brought up as an example, such as in this news report:

> If there is one cautionary tale that surfaces in discussions of this technology, it is the case of China's policing through an array of cameras equipped with facial recognition software of the Uighurs, a largely Muslim minority in the western part of the country. (della Cava, 2019)

Pushback against these arguments was divided. Some proponents of FR in law enforcement, such as Ron Lawrence of the California Police Chiefs Association, asserted that privacy would be respected and technology won't be misused. 'Let me be clear, law enforcement respects and understands the importance of protecting a person's right to privacy', he said. 'We believe a person's privacy should not be violated unless that person is a threat to themselves or to others. We stand by this and will continue to do so in the future' (Lawrence 2019). But others, such as the Riverside Sheriffs' Association in its official opposing argument on the legislative floor, argued that citizens *did not* have a reasonable expectation of privacy in public. They also questioned why 'civil libertarians' were only concerned about privacy now and did not speak up for the privacy of law enforcement officers when an earlier law mandated the public disclosure of body camera video (Body Camera Accountability Act: Hearing 2019b).

The contradictory claims of FR proponents serve to justify the fears of FR's critics. Even if one takes the assurances of officers such as Lawrence at face value, there would be others in law enforcement who simply do not respect or

recognise people's privacy – nor the basic rights guaranteed to them in the law they are claiming to enforce.

Public Safety

The most common argument proffered by proponents of FR in policing was that it would improve public safety. To do so, they cited the 'success' of FR in other states and countries. Two frequently quoted examples included: the reported use of FR to capture the perpetrator of the *Capital Gazette* shooting in Maryland in 2018 (della Cava 2019); and a supposed 60% reduction in carjackings in Detroit after the installation of a citywide FR system (Lawrence 2019).

In addition, law enforcement groups such as the Riverside Sheriffs' Association frequently brought up California's plans to host mega events – including the annual Coachella Arts and Music Festival and the 2028 Summer Olympics – and the need to ensure public safety at these events. A ban on FR, they claimed, would signal the state's inability to protect participants and visitors and could potentially mean the events would move elsewhere. This argument, first made in official comments to the legislature in opposition of AB-1215 (Body Camera Accountability Act: Hearing 2019b, 10), was picked up and amplified by media coverage.

But critics of FR turned the argument on its head. They claimed that the technology would undermine rather than improve public safety – especially for minorities. If law enforcement officers were to police these communities while wearing FR-enhanced body cameras, members of the public would likely hesitate to interact with officers, even as victims or witnesses of crimes, for fear of having their faces caught on camera and stored in a database in perpetuity. This would make the job of law enforcement more difficult and also put these communities in greater danger (Body Camera Accountability Act 2019a). They also argued that public safety can be undermined by law enforcement officers suspecting or arresting innocent civilians.

Discrimination as Technological Artefact

Indeed, the disproportionately negative impact of FR on marginalised communities – minorities and immigrants – was a theme that surfaced in many different ways. Lawmakers and civil rights groups made this a key part of their argument against FR from the outset. In his 9 May statement when the bill went to the Senate, for instance, Ting noted that 'AB-1215 is an important civil rights measure that will prevent exploitation of vulnerable communities' (Office of Assembly member Ting 2019a).

Albert Fox Cahn, who founded The Surveillance Technology Oversight Project at the New York-based Urban Justice Center, wrote in a commentary,

'There's something unbearable about thinking that our country's largest investment ever in police accountability' – referring to body cameras – 'would be turned into a weapon against the very communities of colour that it was supposed to protect' (Cahn, 2019). Similarly, the Electronic Frontier Foundation noted that FR 'exacerbates historical biases born of, and contributing to, overpolicing in Black and Latinx neighbourhoods' (EFF 2019).

Curiously, a few months into the bill's passage, there was an important shift in this argument. Social discrimination moved from being a consequence of 'historical biases' to being a function of the inaccuracy of the FR technology itself. The shift was especially evident after Rekognition failed the ACLU's test on 13 August 2019, during which it misidentified 26 California lawmakers as criminals – many of them, including Ting, being people of colour. This technological failure quickly became central to Ting's statements against FR and justification for AB-1215. As he said after the experiment: 'This experiment reinforces the fact that facial recognition software is not ready for prime time – let alone for use in body cameras worn by law enforcement' (ACLU Northern California 2019).

To be sure, FR's inaccuracy when it came to identifying people of colour, had been known for long. ACLU had run a similar test a year earlier, with similar results. However, as the new experiment, and Ting's mobilisation of its results to back his push for the bill, began to dominate the news, the discourse about social discrimination shifted subtly yet recognisably. Journalists increasingly began to associate FR's potential for exacerbating marginalisation to what Ting called its 'dangerous inaccuracies' as a technological artefact (Office of Assembly member Ting 2019b).

Ironically, scientific support for the arguments of FR's detractors soon became the bill's undoing – at least in terms of what it was initially intended to be. Amazon disputed the findings of the experiment, claiming the researchers purposefully used a lower confidence threshold than recommended (they used the default settings in the Rekognition software) (Gardiner 2019a). The Information Technology and Innovation Foundation argued FR was, in fact, more accurate than human recognition and so was an improvement to existing techniques in terms of accuracy (Information Technology and Innovation Foundation 2019). Even those who acknowledged that the ACLU experiment revealed a significant problem added that the problem, being technological, could – and eventually would – be resolved. Both Axon, one of the largest manufacturers of body cameras in the US, and Microsoft, which has its own FR software, agreed that FR had an accuracy and fairness problem – in its *current* state. As an Axon report said, 'At the least, face recognition technology should not be deployed until the technology performs with far greater accuracy and performs equally well across races, ethnicities, genders and other identity groups' (Sumagaysay 2019). The idea that the main problem with FR was a technological flaw that needed to be and *could be* fixed given time became dominant.

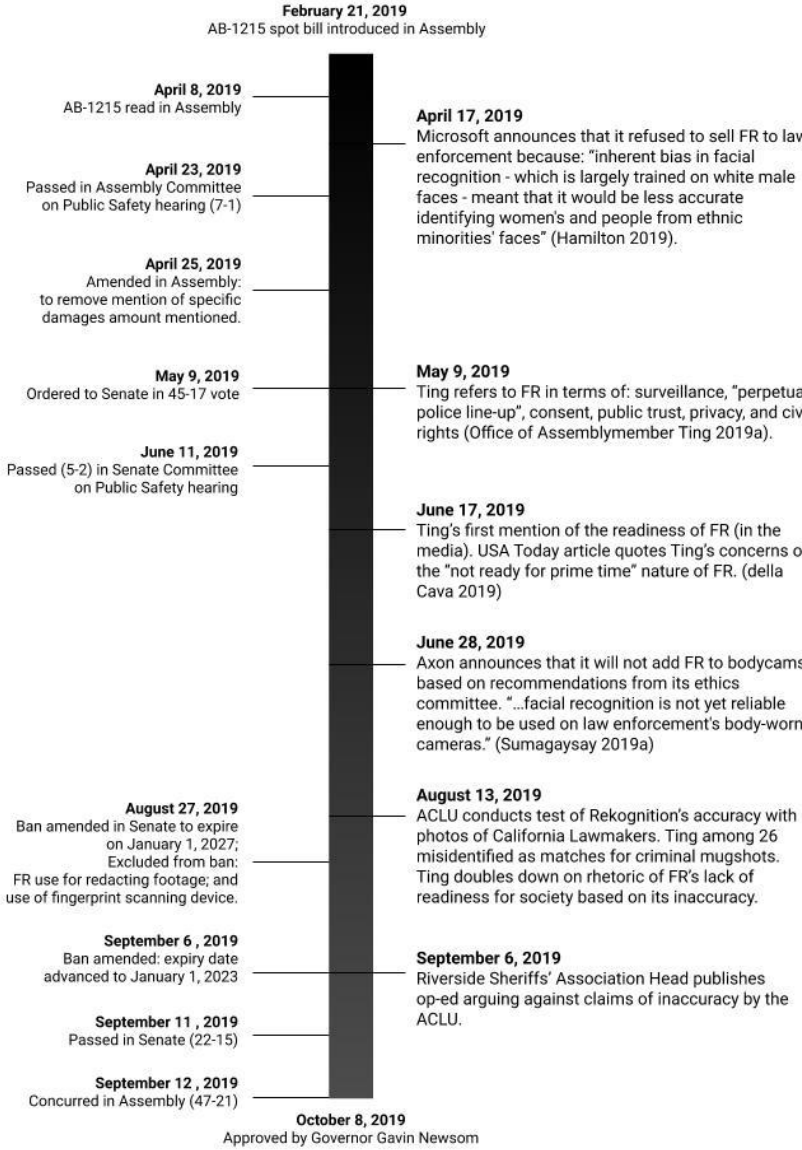

Figure 13.1: A Brief History of AB-1215.

This discursive shift came as a shot in the arm for FR proponents on the assembly floor, such as the Riverside Sheriffs' Association. They had already been arguing that, as a world leader in technology development, California was not the kind of state that would ban a technology simply because it wasn't perfect. 'Could any of us imagine a statutory ban on Microsoft Office or Apple's

iOS until the software was able to be certified as 100% flawless?' they had asked in their official comments to the legislature in June (Body Camera Accountability Act: Hearing 2019b, 10).

Two weeks after the 13 August experiment, the bill was amended to include a sunset clause (see Figure 13.1). The indefinite ban on FR was now set to expire after seven years. On 6 September, the expiry was further reduced to three years. Days after these amendments, the bill was passed in both houses – and signed by Governor Newsom a month later.

Conclusion: Why Biopolitics Matters

Our study has examined AB-1215 as a social discourse, adopting a *law and society* approach that views legislations as socially negotiated and focuses, among other things, on the 'construction of meanings' of the law and how it influences the legal/regulatory process (Ewick and Silbey 1998). Our analysis leads us to two broad conclusions to guide future social research and social action vis-à-vis FR and algorithmic governance in general. Firstly, the discourse comprised both 'particular' and 'universal' features – arguments and claims that were rooted in the political and cultural contexts in which they were made but also drew on ideas and concepts that transcended those contexts. For instance, the concern with FR systems infringing upon citizens' privacy is universal, but in this discourse, this concern was hybridised with guarantees of personal freedom that were specific to the US constitution. Meanwhile, FR's proponents, while echoing universalist claims about enhancing public safety with the aid of technology, also mobilised California's identity as a forward-looking and technology-friendly state – home to the Silicon Valley – to delegitimise calls for banning FR from law enforcement. A *hybrid* analytical lens that is sensitive to both universal and particular characteristics of FR as a social discourse is therefore vital for producing a nuanced picture.

Secondly, while FR's negative externalities – or the harms to 'third parties' it can cause (Petit 2017) – were initially perceived as *systemic*, they later came to be constructed in more *discrete* terms, albeit with certain systemic elements. Specifically, social discrimination – which FR was expected to reinforce – was discussed in terms of 'historical biases' at first but eventually became a function of machine learning-related inaccuracies of the technology itself. Crucially, this shift also implied that the harms it caused would be individualised and random, rather than affecting large sections of the populace in a predictable manner. The extent of the harm was still deemed unsustainable – wrongly identified individuals could end up being in prison or worse – and thus regulation was still warranted. But because the shortcoming was now perceived as technological, the need for regulation was supposed to be temporary: technology would, after all, improve – as technology is always expected to – and inaccuracies would reduce and eventually go away.

In theoretical terms, we witnessed the discourse transforming from *biopolitical* to technological determinist – foregrounding technology as the cause underlying social phenomena, good or bad, and de-emphasising concerns about institutionalised racism and mass surveillance (see also, Gangadharan and Niklas 2019). Ironically, this shift was precipitated by a well-intentioned scientific experiment carried out by a civil rights group. It is possible the ACLU experiment helped the bill get past the assembly floor by making the harms FR can cause appear more concrete and measurable. However, it simultaneously reduced the concerns with FR to the level of the technology itself – an example of what Selbst et al. (2019) have called a failure of abstraction. The social discourse lost its biopolitical magnitude – and so did the bill.

Shahin (2019) has drawn attention to the theoretical significance of 'critical junctures' – emergent conditions in which a social discourse takes on a new direction without the principal stakeholders intending it to – in shaping regulations about technology. The ACLU experiment in August 2019 was such a critical juncture in the discourse about AB-1215. Even though ACLU's own position on why FR had no place in law enforcement did not shift after the experiment – its press releases, for instance, retained their focus on systemic issues – both Assembly member Ting and the media latched on to Rekognition's manifest failure as a piece of technology. This quickly undermined the original intent of the bill. To be sure, other factors may have also played a role in the willingness shown by Ting and other supporters of the bill to accept a sunset clause – twice – and change the character of the bill from a permanent ban on the use of FR in police body cameras to a three-year moratorium. But the noticeable change in the social discourse following the experiment, coinciding with changes in the bill itself, does indicate that the experiment weakened the bill even as it became instrumental for its passage.

Our analysis is significant not only for future research on FR but also for future efforts to check algorithmic governance, legislative or otherwise, in the US and around the world. Firstly, it underlines the significance of a biopolitical approach to understanding – and resisting – algorithmic governance. That does not mean technological flaws are not important to point out. But those flaws are themselves the consequence rather than the cause of institutionalised racism: they don't produce but serve to *re*-produce discrimination and marginalisation.

Research and resistance therefore need to press forward with an agenda in which FR and other forms of artificial intelligence are viewed as sociotechnical artefacts interpellated in relations of power – produced by them even as they serve to reproduce those relations. Moreover, a technologically determinist outlook, where activists focus only on the machine, its algorithms, input and output, and not as much on the social contexts of its design and use, is not only a failure of abstraction (Selbst et al. 2019) – it is also a failure of strategy. Blaming the technology alone might appear attractive in the short-term but, as our analysis indicates, it does not aid the long-term goal of regulating algorithmic governance as a means of achieving social justice.

Secondly, these relations of power increasingly have both universal and particular – or global and local – dimensions. Being sensitive to such hybridity is important for research on and resistance to algorithmic governance in countries like the US, as our analysis indicates, but even more so in the Global South. That is partly because sociotechnical artefacts such as FR are constructed in North America and Western Europe, along with certain norms and practices of governance, and are often then 'localized' (Zaugg 2019) in the Global South amidst different forms of social hierarchy.

Understanding these hybridised dynamics opens new avenues for research and resistance aimed at exposing and destabilising such hierarchies. For instance, how are algorithms trained to discriminate against people of colour implicated in the biopolitics of societies where the entire population is 'of colour'? If algorithms for governance are coded and trained from scratch rather than copied and pasted, what kinds of power and norms of control do they reflect and reproduce? Comparative research across countries, and empirical research focusing on specific countries and contexts of design and use of these technologies are important. As new legal instruments are developed to regulate this technology, collaborations between legal scholars and scholars of social science are key in understanding how we can negotiate these technologies as a society.

In conclusion, we emphasise the main argument of our study. Scholars and activists have long been aware of the role of algorithms and artificial intelligence in marginalising minorities and immigrants and reinforcing relations of power (boyd and Crawford 2012; Eubanks 2018). Biometric technologies, such as FR, are particularly insidious examples as they can act at the level of both the 'body' and the 'body politic'. They reduce human beings into data points that may be stored, manipulated and controlled en masse (Browne 2010; Hood 2020). At the same time, they enable forms of domination that are *systemic* in nature: they have a long history and they are institutionalised in a variety of social practices. Indeed, algorithms themselves represent one such social practice. The prejudices they exhibit are a consequence of the systemic bias they are produced by – even as the algorithms help re-enact and reproduce that bias. Research on and resistance to algorithmic governance should, therefore, avoid the trap of technological determinism and not lose sight of the systemic nature of their subject matter – the biopolitics of discrimination and domination.

References

ACLU. 2019a. California Senate Votes to Block Face Recognition on Police Body Cameras [PRESS RELEASE], 11 September 2019. https://www.aclunc.org/news/california-senate-votes-block-face-recognition-police-body-cameras.

ACLU. 2019b. California Governor Signs Landmark Bill Halting Facial Recognition on Police Body Cams [PRESS RELEASE], 8 October 2019. https://

www.aclunc.org/news/california-governor-signs-landmark-bill-halting-facial-recognition-police-body-cams.

ACLU Northern California. 2019. Facial Recognition Technology Falsely Identifies 26 California Legislators with Mugshots [PRESS RELEASE], 13 August 2019. https://www.aclunc.org/news/facial-recognition-technology-falsely-identifies-26-california-legislators-mugshots.

Anderson, B. 2019. California Considers Plan to Ban Facial Recognition Technology. *The Fresno Bee*. 16 May 2019. https://www.fresnobee.com/news/california/article230437789.html.

Andrejevic, M. 2019. Automating Surveillance. *Surveillance & Society* 17 (1/2): 7–13.

Barbrook, R. and Cameron, A. 1996. The Californian Ideology. *Science as Culture* 6 (1): 44–72. DOI: https://doi.org/10.1080/09505439609526455.

Barik, S. 2020. Microsoft Will Not Sell Facial Recognition Tech to Police in the US without Federal Law. *Medianama*, 12 June 2020. https://www.medianama.com/2020/06/223-microsoft-facial-recognition

Bedoya, A.M. 2019. The Color of Surveillance. *Slate*. 18 January. https://slate.com/technology/2016/01/what-the-fbis-surveillance-of-martin-luther-king-says-about-modern-spying.html.

Benjamin, R. 2016. Catching Our Breath: Critical Race STS and the Carceral Imagination. *Engaging Science, Technology, and Society* 2: 145–156.

Body Camera Accountability Act, A.B. 1215. 2019a. Amended in Assembly. 8 April 2019.

Body Camera Accountability Act, A.B. 1215. 2019b. Amended in Assembly. 25 April 2019.

Body Camera Accountability Act, A.B. 1215. 2019c. Amended in Senate. 27 August 2019.

Body Camera Accountability Act, A.B. 1215. 2019d. Amended in Senate. 6 September 2019.

Body Camera Accountability Act, A.B. 1215. 2019e. Enrolled, 8 October 2019.

Body Camera Accountability Act: Hearing on A.B. 1215 Before the Sen. Comm. on Public Safety. 2019a. (statement of the American Civil Liberties Union). http://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=201920200AB1215

Body Camera Accountability Act: Hearing on A.B. 1215 Before the Sen. Comm. on Public Safety. 2019b. (statement of the Riverside Sherriffs Association). http://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=201920200AB1215

boyd, d. and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15 (5): 662–679.

Browne, S. 2010. Digital Epidermalization: Race, Identity and Biometrics. *Critical Sociology* 36 (1): 131–150.

Buolamwini, J. and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability and Transparency*, 77–91.

Cagle, M. 2020. California Just Blocked Police Body Cam Use of Face Recognition. *ACLU Free Future* (blog). 11 October 2020. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/california-just-blocked-police-body-cam-use-face.

Cahn, A.F. 2019. Facial Recognition Tech Is a Blatant Misuse of Police Bodycams. *The Daily Beast*. 17 October 2019. https://www.thedailybeast.com/facial-recognition-tech-is-a-blatant-misuse-of-police-bodycams.

Cameron, D. 2020. Detroit Police Chief Admits Face Recognition Doesn't Work '95–97% of the Time'. *Gizmodo*, 29 June 2020. https://gizmodo.com/detroit-police-chief-admits-face-recognition-doesnt-wor-1844209113.

Chabria, A. 2019. Facial Recognition Software Mistook 1 in 5 California Lawmakers for Criminals, Says ACLU. *Los Angeles Times*, 13 August 2019. https://www.latimes.com/california/story/2019-08-12/facial-recognition-software-mistook-1-in-5-california-lawmakers-for-criminals-says-aclu.

Chappell, B. 2019. ICE Uses Facial Recognition to Sift State Driver's License Records, Researchers Say. *NPR*, 8 July 2019. https://www.npr.org/2019/07/08/739491857/ice-uses-facial-recognition-to-sift-state-drivers-license-records-researchers-sa.

Charmaz, K. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Thousand Oaks, CA: Sage.

Cohen, J.E. 2012. What Privacy Is For. *Harv. L. Rev*. 126: 1904.

Colaner, S. 2020. Detroit's Fight over Policing and Facial Recognition Is a Microcosm of the Nation. *VentureBeat*, 25 June 2020. https://venturebeat.com/2020/06/25/detroits-fight-over-policing-and-facial-recognition-is-a-microcosm-of-the-nation

della Cava, M. 2019. Face-Recognition Technology Has Heads Spinning in Calif. *USA Today*, 17 June 2019. https://www.pressreader.com/usa/usa-today-us-edition/20190617/281500752769807

Demush, R. 2019. A Brief History of Computer Vision (and Convolutional Neural Networks). *HackerNoon*, 26 February 2019. https://hackernoon.com/a-brief-history-of-computer-vision-and-convolutional-neural-networks-8fe8aacc79f3.

Devereaux, R. 2019. Secret Terrorism Watchlist Found Unconstitutional in Historic Decision. *The Intercept*, 6 September 2019. https://theintercept.com/2019/09/06/terrorism-watchlist-lawsuit-ruling

EFF. 2019. Hearing Tuesday: EFF Will Voice Support for California Bill Reining in Law Enforcement Use of Facial Recognition [PRESS RELEASE], 10 June 2019. https://www.eff.org/press/releases/hearing-tuesday-eff-will-voice-support-california-bill-reining-law-enforcement-use.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Ewick, P. and Silbey, S.S. 1998. *The Common Place of Law: Stories from Everyday Life*. Chicago, IL: University of Chicago Press.

Ferretti, C. 2020. Residents Urge City Council to Reject Proposed Facial Recognition Contract Extension. *The Detroit News*, 16 June 2020. https://www.detroitnews.com/story/news/local/detroit-city/2020/06/16/residents-urge-city-council-reject-proposed-facial-recognition-contract/3197917001/.

Foucault, M. 2003. *Society Must Be Defended: Lectures at the Collège de France 1975–1976*, trans. David Macey. New York: Picador, 242.

Gandy Jr, O.H. 1993. *The Panoptic Sort: A Political Economy of Personal Information.* Critical Studies in Communication and in the Cultural Industries. Boulder, CO: Westview Press/ERIC.

Gangadharan, S.P. and Niklas, J. 2019. Decentering Technology in Discourse on Discrimination. *Information, Communication & Society* 22 (7): 882–899.

Gardiner, D. 2019a. Facial ID Misses Mark, Test by ACLU Reveals. *San Francisco Chronicle*, 13 August 2019. https://www.sfchronicle.com/politics/article/Facial-recognition-misidentified-26-California-14301190.php.

Gardiner, D. 2019b. Lawmakers OK Ban on Police Use of Facial Recognition. *San Francisco Chronicle*, 13 September 2019. https://www.sfchronicle.com/politics/article/California-lawmakers-vote-for-3-year-ban-on-14436245.php.

Gardiner, D. 2019c. California Blocks Police from Using Facial Recognition in Body Cameras. *San Francisco Chronicle*, 7 October 2019. https://www.sfchronicle.com/politics/article/California-blocks-police-from-using-facial-14502547.php.

Garvie, C., Bedoya, A. and Frankle, J. 2016. The Perpetual Line-Up. Unregulated Police Face Recognition in America. Georgetown Law Center on Privacy & Technology, October. https://www.perpetuallineup.org.

Gates, K.A. 2011. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. Vol. 2. New York: New York University Press.

Goldstein, A.J., Harmon, L.D. and Lesk, A.B. 1971. Identification of Human Faces. *Proceedings of the IEEE* 59 (5): 748–760. DOI: https://doi.org/10.1109/PROC.1971.8254.

Gomez, J. and Rosenberg, L. 2019. In the Hands of Police, Facial Recognition Software Risks Violating Civil Liberties. *USA Today*, 18 October 2019. https://www.usatoday.com/story/opinion/policing/2019/10/18/hands-police-facial-recognition-tech-violates-civil-liberties/3904469002

Greenwald, G., MacAskill, E. and Poitras, L. 2013. Edward Snowden: The Whistleblower behind the NSA Surveillance Revelations. *The Guardian*, 11 June 2013. https://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance.

Guariglia, M. 2019. Victory! California Governor Signs A.B. 1215. *Electronic Frontier Foundation* (blog). 9 October 2019. https://www.eff.org/deeplinks/2019/10/victory-california-governor-signs-ab-1215.

Hamilton, I.A. 2019. Microsoft Took an Ethical Stand on Facial Recognition Just Days after Being Blasted for a Sinister AI Project in China. *Business Insider*, 17 April 2019. https://www.businessinsider.in/tech/microsoft-took-an-ethical-stand-on-facial-recognition-just-days-after-being-blasted-for-a-sinister-ai-project-in-china/articleshow/68922085.cms.

Hill, K. 2020. The Secretive Company That Might End Privacy as We Know It. *New York Times*, 18 January 2020. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Hood, J. 2020. Making the Body Electric: The Politics of Body-Worn Cameras and Facial Recognition in the United States. *Surveillance & Society* 18 (2): 157–169.

Huang, G.B., Mattar, M., Berg, T. and Learned-Miller, E. 2008. Labeled Faces in the Wild: A Database Forstudying Face Recognition in Unconstrained Environments. In:*Workshop on Faces in 'Real-Life' Images: Detect ion, Alignment, and Recognition*. https://hal.inria.fr/inria-00321923/file/Huang_long_eccv2008-lfw.pdf.

Information Technology and Innovation Foundation. 2019. Information Technology & Innovation Foundation Issues Statement on Facial Recognition Technology for Law Enforcement [PRESS RELEASE], 11 September 2019.

Kak, A. 2020. *Regulating Biometrics: Global Approaches and Urgent Questions*. AI Now. https://ainowinstitute.org/regulatingbiometrics.pdf.

Lawrence, R. 2019. Commentary: Why Law Enforcement Should Use Facial Recognition. *San Diego Union-Tribune*, 6 September 2019. https://www.sandiegouniontribune.com/opinion/story/2019-09-06/commentary-why-law-enforcement-should-use-facial-recognition.

Mann, M. and Smith, M. 2017. Automated Facial Recognition Technology: Recent Developments and Approaches to Oversight. *UNSWLJ* 40: 121.

Mayring, P. 2004. Qualitative Content Analysis. *A Companion to Qualitative Research* 1 (2004): 159–176.

Metz, R. 2019. Beyond San Francisco, More Cities Are Saying No to Facial Recognition. *CNN Wire*, 17 July 2019. https://edition.cnn.com/2019/07/17/tech/cities-ban-facial-recognition/index.html.

Morozov, E. 2012. *The Net Delusion: The Dark Side of Internet Freedom*. New York: Public Affairs.

Naughton, J. 2020. Quick, Cheap to Make and Loved by Police – Facial Recognition Apps Are on the Rise. *The Guardian*, 25 January 2020. https://www.theguardian.com/technology/commentisfree/2020/jan/25/facial-recognition-apps-are-on-the-rise.

Ng, A. 2020a. Microsoft Pushed Its Facial Recognition to Federal Agencies, Emails Show. *CNET*, 17 June 2020. https://www.cnet.com/news/microsoft-pushed-its-facial-recognition-to-federal-agencies-emails-show/.

Ng, A. 2020b. Lawmakers Propose Indefinite Nationwide Ban on Police Use of Facial Recognition. *CNET*, 25 June 2020. https://www.cnet.com/news/lawmakers-propose-indefinite-nationwide-ban-on-police-use-of-facial-recognition

NPR. 2020. Tech Companies Are Limiting Police Use of Facial Recognition. Here's Why. *Short Wave*, 23 June 2020. https://www.npr.org/2020/06/22/881845711/tech-companies-are-limiting-police-use-of-facial-recognition-heres-why.

Office of Assembly member Ting. 2019a. Ting Proposal Banning Facial Recognition Technology in Body Cams Approved by State Assembly [PRESS RELEASE], 9 May 2019. https://a19.asmdc.org/press-releases/20190509-ting-proposal-banning-facial-recognition-technology-body-cams-approved-state

Office of Assembly member Ting. 2019b. Facial Recognition Technology Falsely Identifies 26 California Legislators, Including Ting, with Mugshots [PRESS RELEASE], 13 August 2019. https://a19.asmdc.org/press-releases/20190813-facial-recognition-technology-falsely-identifies-26-california-legislators.

Oliver, D. 2019. Facial Recognition Scanners Are Already at Some US Airports. Here's What to Know. *USA Today*, 16 August 2019. https://www.usatoday.com/story/travel/airline-news/2019/08/16/biometric-airport-screening-facial-recognition-everything-you-need-know/1998749001/.

Pato, J.N. and Millett, L.I. 2010. National Research Council (US) Whither Biometrics Committee. In: *Biometric Recognition: Challenges and Opportunities*. National Academies Press (US).

Petit, N. 2017. Law and Regulation of Artificial Intelligence and Robots-Conceptual Framework and Normative Implications SSRN, 9 March. DOI: http://dx.doi.org/10.2139/ssrn.2931339

Pope, C. 2018. Biometric Data Collection in an Unprotected World: Exploring the Need for Federal Legislation Protecting Biometric Data. *JL & Pol'y* 26: 769.

Saldaña, J. 2009. *The Coding Manual for Qualitative Researchers*. New York: Sage.

Selbst, A.D., boyd, d., Friedler, S.A, Venkatasubramanian, S. and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68. New York: ACM.

Seron, C. and Silbey, S. 2004. Profession, Science, and Culture: An Emergent Canon of Law and Society Research. In Sarat, A (Ed.)*The Blackwell Companion to Law and Society,* pp. 30–59. Oxford: Blackwell Publishers

Shahin, S. 2019. Facing up to Facebook: How Digital Activism, Independent Regulation, and Mass Media Foiled a Neoliberal Threat to Net Neutrality. *Information, Communication & Society* 22 (1): 1–17.

Shwayder, M. 2020. Police Facial Recognition Tech Could Misidentify People at Protests, Experts Say. *Digital Trends*, 2 June 2020. https://www.digitaltrends.com/news/police-protests-facial-recognition-misidentification

Simonite, T. 2019. The Best Algorithms Struggle to Recognize Black Faces Equally. *Wired*, 22 August 2019. https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally

Snow, J. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots. *ACLU* (blog). 26 July 2018. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

Sumagaysay, L. 2019. No More Facial Recognition in Cop Body Cams – Axon Announces New Policy as Concern Grows over Reliability. *The Mercury News*, 28 June 2019. https://www.mercurynews.com/2019/06/27/no-facial-recognition-in-police-body-cams-their-biggest-maker-vows

Taddonio, P. 2019. How China's Government Is Using AI on Its Uighur Muslim Population. *Frontline PBS*, 21 November 2019. https://www.pbs.org/wgbh/frontline/article/how-chinas-government-is-using-ai-on-its-uighur-muslim-population

Taebi, B., Correlje, A., Cuppen, E., Dignum, M. and Pesch, U. 2014. Responsible Innovation as an Endorsement of Public Values: The Need for Interdisciplinary Research. *Journal of Responsible Innovation* 1 (1): 118–124.

Talbot, C. 2011. Paradoxes and Prospects of 'Public Value'. *Public Money & Management* 31 (1): 27–34.

Thebault, R. 2019. California Could Become the Largest State to Ban Facial Recognition in Body Cameras. *The Washington Post*, 12 September 2019. https://www.washingtonpost.com/technology/2019/09/12/california-could-become-largest-state-ban-facial-recognition-body-cameras

van der Ploeg, I. 2002. Biometrics and the Body as Information: Normative Issues of the Socio-Technical Coding of the Body. *Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination*, pp. 57–73. London: Routledge.

Zaugg, J. 2019. India is Trying to Build the World's Biggest Facial Recognition System. *CNN*. 18 October 2019. https://www.cnn.com/2019/10/17/tech/india-facial-recognition-intl-hnk/index.html

CHAPTER 14

Beyond Mechanical Turk: The Work of Brazilians on Global AI Platforms

Rafael Grohmann and Willian Fernandes Araújo

Introduction

'Artificial Artificial Intelligence', the slogan of Amazon Mechanical Turk (AMT), a global AI platform, is as ironic as the history of the 19th century Mechanical Turk itself. But it is less ironic than perverse to see that part of the profit of one of the owners of Cloud Empire (Couldry and Mejias 2019) is related to the crowd work of millions of workers around the world.

The slogan also reveals the simplistic nature of the debates about the future of work and artificial intelligence (AI) that carry persistent representations of a 'general artificial intelligence' driven by Hollywood imagery. These discourses point to visions of AI playing a key role in the full automation of work, whether in positive or negative frameworks (Dyer-Witheford, Kjosen and Steinhoff 2019).

These narratives make inequalities and ghost work (Gray and Suri 2019) invisible, even though crowdsourced labour performed by humans, in fact, supports AI. Thus, there are tasks that might, in theory, be performed by AI, but are cheaper and/or quicker to simply outsource to human workers' (Woodcock and Graham 2019, 58). However, AMT workers are not the only ones supporting AI. Companies like Appen, Lionbridge, Mighty AI, Clickworker and Spare5 also play a key role as data trainers for AI, with a variety of work activities on their platforms, including data training for self-driving cars. The

How to cite this book chapter:
Grohmann, R. and Araújo, W. F. 2021. Beyond Mechanical Turk: The Work of Brazilians on Global AI Platforms. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 247–266. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.n.

discourses of these platforms propagate meanings of future and progress. Appen's slogan, for example, is: 'confidence to deploy AI with world-class training data – artificial intelligence will improve the world'. This helps to consolidate images of digital labour, AI and data among workers (Soriano and Cabanes 2019; Beer 2019).

Work behind artificial intelligence is called 'ghost work' (Gray and Suri 2019), 'clickwork' (Casilli 2019) and 'micro-work' (Tubaro and Casilli 2019; Lehdonvirta 2016). These metaphors are attempts to name work activities on AI platforms. They are not definitive notions, but just illustrations. For instance, the fact that this work consists of individual, compartmentalised 'tasks' lasting perhaps only seconds or minutes, does not make it 'micro'. In a similar way, these workers do more than just click on ads. On the one hand, the multiplicity of tasks involves audio transcriptions and translations, describing images, recording videos and photos, and so on. On the other hand, we understand that work activities, whatever they may be, involve the entire (material) body of workers (Huws 2014).

Whatever the name, these people work for global artificial intelligence platforms. The global character of these systems points towards an important factor in the complexification of the human work behind AI: the platforms' geopolitical dimension. Fuchs and Sandoval (2014) point to a new international division of labour (NIDL). However, work on global AI platforms is different from the circuit of labour involving the iPhone, for example (Qiu, Gregg and Crawford 2014). On the global AI platforms there is no division between lithium battery production in one place and software production in another, although they depend, in a sense, on this circuit of labour and these digital infrastructures. In this case, there are a few companies from the Global North managing and controlling a crowd of workers from many countries in the world, mainly from the so-called Global South.

Crawford and Joler (2018) show connections among human labour, data and planetary resources, including the production of data for AI as a circuit of digital labour (Qiu, Gregg and Crawford 2014): physical labour of mine workers, labour in distribution centres, crowdsourced labour on global AI platforms, and so on (Crawford and Joler 2018). This means highlighting the material dimensions of the work behind AI, from workers' 'human intelligence' to media materialities, digital infrastructures, with diverse impacts, including geological (Parikka 2015; Milan 2018; Murdock 2018).

We agree with the notion of the 'planetary labour market' (Graham and Anwar 2019). Global AI platforms do not eliminate physical spaces and are dependent on material infrastructures. The planetary scale of these platforms means that 'labour markets help clients operate unboundedly and transspatially, and allow them to reconfigure the geography of their production networks for almost zero cost' (Graham and Anwar 2019, 28). The workers, although they 'can sell their labour power globally, [they] are tethered to the

locales in which they go to bed every night' (Graham and Anwar 2019, 29). According to the Online Labour Index (OLI) of the University of Oxford, the country of origin of the largest number of online freelance tasks is the United States, and the country with the largest number of online workers is India.

Thus, understanding the inequalities in work for global AI platforms means going beyond AMT and workers from the Global North. In a digital economy, there are inequalities involving local workers and global platforms. There is no 'digital labour universalism' or a homogeneous and unique notion of global workforce. There are, in fact, diverse work and AI scenarios around the world.

The aim of this chapter is to analyse the work of Brazilians on global AI platforms, mainly Appen and Lionbridge, from the workers' point of view, understanding platform labour and the platformization of labour as a secret ingredient in automation and contextualising platform labour in terms of the Global South (Grohmann and Qiu 2020). The Latin American scenario is less well known in academic research on labour and AI in the Global South in relation to countries like the Philippines and India. The focus of this research in Brazil is not intended to argue a singularity of this country, but to show the existence of other realities beyond Global North that are also invisible academically from critical research on AI.

In the following section, we discuss platformization of labour and AI. Then, we present the methodology, with interviews and a survey with workers, together with observations of Facebook and WhatsApp groups. The topics of analysis are: communication among workers, difficulties of work, lack of infrastructural reliability, workers' strategies, definition of work and understanding of AI. The results reveal the complexity of working on global AI platforms and AI imaginaries.

Platformization of Labour and Artificial Intelligence

There is no AI without 'ghost work' (Gray and Suri 2019). Conversely, the work that supports AI is only possible with the existence of digital platforms. 'Platform' is the term that refers to the sociotechnical infrastructures of digital conglomerates and that connect different parts – State, companies, consumers, workers, and so on (Srnicek 2016). In the literature on digital media, it highlights the performative character of these structures (Introna 2016). Online platforms are considered mediators that not only enable or facilitate certain practices, but also actively shape, transform, and distort content, relationships, understandings, etc. (Gillespie 2014). Simultaneously, that contemporary sociotechnical model for organising practices inserts processes driven by algorithms and digital data into different contexts (Van Dijck, Poell and De Waal 2018). The algorithm-driven logic that underpins these platforms is the confluence of different factors – business models, user data, algorithms, data centres,

servers, etc. Thus, digital platforms are infrastructures that depend on data and algorithms and which present values and standards inherent in their designs.

Van Doorn and Badger (2020), for instance, highlight the role of data assets and meta-platforms in platform labour. However, as Srnicek (2016) argues, there is not just one type of platform. The role of algorithms and data in platform labour depends on its mechanisms and ways of extracting value (Sadowski 2020). In the case of AI, platforms are the places for many workers to produce and circulate data to automate processes, that is, towards AI (Beer 2016, 2019). Ghost work behind AI is a way of mining workers' process for data (Neff, McGrath and Prakash 2020). Thus, working for global AI platforms is a type of data labour for automation.

There is no platform without moderation, as state Gillespie (2018) and Roberts (2019). More recently, Gillespie (2020) highlights the promise of 'the promise of AI' and the question of scale regarding data. 'The claim that moderation at scale requires AI is a discursive justification for putting certain specific articulations into place – like hiring more human moderators, so as to produce training data, so as to later replace those moderators with AI' (Gillespie 2020, 2). This means understanding the heteromation of labour (Ekbia and Nardi 2017) on global AI platforms.

The platformization of labour, in line with Casilli and Posada (2019) and Van Dijck, Poell and De Waal (2018), is linked to the growing dependence on digital platforms to get and/or maintain work activities. This process is a social synthesis of others: datafication, financialisation and neoliberal rationality (Sadowski 2020, Dardot and Laval 2013). Platform labour requires large-scale data extraction and collection for its surveillance and algorithmic management mechanisms to be successful. Sadowski (2020) states that there are three key mechanisms of rentier capitalism platforms: data extraction, digital enclosure and capital convergence. Moreover, the data is a form of capital for platform companies, expropriating and colonising workers' resources, especially in the Global South (Couldry and Mejias 2019).

The platformization of labour does not materialise in the same way for different actors and social institutions. There is a heterogeneity of workers with relation to gender, race, class and geography issues. This heterogeneity means not only differences, but also inequalities that structure the platform labour (Van Doorn 2017, Grohmann and Qiu 2020). According to Abilio (2020), platform labour is generalisation and appropriation of the livelihoods of peripheral populations by the platform companies.

In Brazil – the focus of this chapter – gig work is not an exception, but historically the way of life of most workers, an ordinary state. The change is that gigs are now platformized. According to Grohmann and Qiu (2020, 5), 'analysing platform labour in the South means that patterns in the North are often erroneously assumed to have also existed in Latin America, Africa and Asia's developing regions, as if labour precarity is a novel phenomenon'. Thus, the

theme of regulation in platform labour is very different in a Latin American context, as workers generally do not want to become regular employees (Abilio 2020). Platform couriers and drivers, for instance, want to have the feeling of autonomy, flexibility and self-management of their own work, which makes it a challenge to think about the regulation of platform labour. This does not mean, however, that they are not organising themselves into unions, associations and strikes (Grohmann et al. 2020; Grohmann and Alves 2020).

The platformization of labour intersects social and geographical contexts with issues around platform materialities and design. Global AI platforms work differently in relation to companies like Uber or Deliveroo, although they have similar mechanisms, such as algorithmic management and surveillance (Woodcock and Graham 2019). First, companies and workers can be located in different parts of the world, and workers often work from their own homes. But this does not necessarily mean that the tasks performed by workers are global. Sometimes they are located in the worker's neighbourhood, city or country, such as advertising analysis or text translation tasks. Second, payment methods vary. AMT, for example, only pays US and Indian workers in cash. In Brazil, workers receive Amazon store credit. Other companies, like Appen and Lionbridge, pay workers in dollars, which makes workers see themselves as part of 'world-class work' (Soriano and Cabanes 2019). This means that there is no homogeneity in practices across global AI platforms. According to Tubaro, Casilli and Coville (2020, 2), 'some platforms such as Mechanical Turk cater to a diverse range of corporate needs, while others specialise in AI services'. Thirdly, there are different worker perceptions regarding the platforms' objectives. At Deliveroo, for example, the courier knows that he/she will deliver food from a restaurant to someone. At AMT or Appen, as our analysis shows, there is production and circulation of ideas about what it means to train algorithms and 'work for AI'. In other words, this can mean alienation from the circuit of labour in global chains.

Global AI platforms accelerate the platformization of labour from the process of 'taskification of labour'. The distribution of micro tasks to the crowd workers materialises from the data labour. According to Casilli and Posada (2019, 10), 'the standardization and the fragmentation of previously complex and specialised processes are essential to run a platform ecosystem where the activities of users fit in and are synchronised with others'. There is data circulation (especially the so-called 'good data') only with the circulation of labour on global AI platforms (Beer 2016). According to Tubaro, Casilli and Coville (2020, 4), 'AI companies depend heavily on data resources, including not only raw data but also annotations that add extra meaning by associating each data point, such as an image, with relevant attribute tags.'

Dyer-Witheford, Kjosen and Steinhoff (2019) state that AI is the current general condition of production, configuring an AI capitalism. However, this means neither general AI nor full automation. Platform labour is the secret ingredient

of automation (Casilli 2019). The work on global AI platforms symbolises the 'heteromation of labour', according to Ekbia and Nardi (2017). Heteromation means keeping human beings in the system, because capitalism needs human beings, extracting value in invisible ways, with new logics of wealth accumulation. According to Ekbia and Nardi (2017), humans are doing a lot of work and machines are getting credit. The role of human beings at work is invisible, although the companies' discourse is towards valuing human beings: 'platforms tell clients that human contribution has value, but not who these humans are and in what conditions they work' (Tubaro, Casilli and Coville 2020, 10).

The intersection of platform labour and heteromation is the synthesis that the future of work will not be exactly automation, but the growing taskification of labour. According to Tubaro, Casilli and Coville (2020, 2), 'automation is still in the making and has not yet been deployed at large scale, [but] its demand for micro-tasks is already transforming the daily practices, experiences and career trajectories of thousands of workers worldwide'. This occurs, according to the authors, in processes of AI preparation, AI verification and AI impersonation. However, it is necessary to highlight that the 'taskification of labour' would not be something original or new, since the 'salary per piece' was already a reality for Marx (1894).

Research leaded by Antonio Casilli and his group has shown that AI platforms can have multiple configurations, from local start-ups to global companies – and this is the emphasis of this chapter. Tubaro, Casilli and Coville (2020) point out that there are platforms like AMT and Clickworker whose workers perform tasks in the most diverse areas and there are other platforms more specialised, ones such as Spare5 and Mighty AI (now owned by Uber), focused on data training for self-driving cars. Thus, the multiplicity of possible tasks on AI platforms shows the flexible nature of their workforce.

Most research on global AI platforms focuses on the Global North, such as Irani (2015), Milland (2017), Gray and Suri (2019) and Ludec, Tubaro and Casilli (2019) highlighting countries such as the United States and France, with a centrality of AMT. However, research by Ludec, Tubaro and Casilli (2019) reveals interesting data for research on global AI platforms outside United States. In France, there are about 500,000 workers registered in AMT. This is a much smaller number than other platforms like ClixSense (7,000,000 workers), Microworkers (1,215,829), Clickworker (1,200,000) and Appen (1,000,000). This reveals the impossibility of generalising the localised experience of AMT workers in countries like the United States. As we stated, there is no digital labour universalism.

One of the few studies that veer away from that Global North trend is Graham, Hjorth and Lehdonvirta (2017) which focuses on Africa, but was written by Global North authors and does not focus exclusively on global AI platforms. Schmidt (2019), in his research on workers that train AI for self-driving cars, finds that most of the workers are from Venezuela. There is even a Brazilian in his research sample. Schmidt's work, nonetheless, does not go further into Latin America itself.

In Brazil, Kalil (2019), and Moreschi, Pereira and Cozman (2020) researched Brazilians working at AMT. Kalil (2019) interviewed 52 people, usually single and graduate men about 30 years old. The alleged reason for work in these platforms is the need for additional income. Kalil (2019) also investigated workers from the United States (sample = 685) and India (sample = 125). Some of the workers' statements are: 'too much work, too little pay, too much exploitation' (Kalil 2019, 189) and 'I can't earn the equivalent of a minimum wage even if I work more than eight hours a day' (Kalil 2019, 189).

The research by Moreschi, Pereira and Cozman (2020) presents survey results with 149 Brazilians working at AMT and observation in a WhatsApp group. The profile is similar to the findings of Kalil (2019): White men and 29 years old. They have been formally unemployed for a long time. In addition, they cite other global AI platforms for which they work, such as Clickworker, Appen and Figure Eight. However, the research focuses only on AMT.

The authors consider the working conditions of Brazilians worse than that of countries like India due to factors such as 'the role of AMT in Brazilian turkers' economic lives, the consequences of the lack of direct payment and the importance of WhatsApp for organising' (Moreschi, Pereira and Cozman 2020, 61). This survey reveals that work of turkers is closely intertwined with the historical informality of labour in the country, a gig economy that existed prior to digital labour itself. 'As Amazon does not make a transfer to their bank account, like turkers in some other countries can [sic], the turkers in Brazil find themselves at the bottom of an unregulated market' Moreschi, Pereira and Cozman (2020, 61).

Research on Brazilians in AMT reinforces that communication also supports the organisation of workers, although still in informal solidarities (Soriano and Cabanes 2019). In the WhatsApp group, workers help each other and present a 'rhetoric that blends entrepreneurship with elements of religiosity and self-help' (Moreschi, Pereira and Cozman 2020, 53). One of the research statements reveals that workers do not want to be a 'ghost': 'I exist and I want you and others to know that' (Moreschi, Pereira and Cozman 2020, 47).

In general, the literature review shows that, on the one hand, there are connections between the tasks of workers for AI platforms in many parts of the world, with the potential for circulation of workers' struggles (Dyer-Witheford 2015; Englert, Woodcock and Cant 2020). On the other hand, there are specificities – of the Global South and, in this case, of Brazil – in relation to payment, task supply, working conditions, difficulties in accessing the platform and problems with language. The background of workers is also a central difference due to issues such as training and the legacy of the informal economy in the country.

Thus, this chapter aims to analyse something not yet explored by research: how Brazilians work on other global AI platforms such as Appen and Lionbridge, in order to highlight other inequalities involving AI and labour. From the class composition perspective (Woodcock 2019; Cant 2019), we focus here on the technical composition of the workers. This does not mean to disregard

the social and political dimensions of class struggles, but to focus the discussion on work activities.

Methodology

From February to April of 2020, we conducted exploratory research across groups where Brazilians discuss their work for global AI platforms. We understand exploratory research as a set of methodological practices developed to offer 'new and innovative ways to analyse reality' (Reiter 2017, 131). To this purpose, we provide transparent guidelines about the research process, aiming to demonstrate its reliability and validity (Reiter 2017). Based on this methodological framework, we began with a preliminary online search for content on the subject, from Brazilian blogs and YouTube channels on this topic. Regarding this content, and considering the previous studies in the literature about digital labour in Brazil (Moreschi, Pereira and Cozman 2020; Abilio 2020; Grohmann and Qiu 2020), we were able to create the following list of AI platforms beyond Mechanical Turk: Appen, Lionbridge, Clickworker, MightyAI, Clixsense, Pactera, iSoftStone and Streetbees.

Starting from this point, we investigated these platforms on LinkedIn, a social networking platform focused on business and employment. This service was chosen because of its public data about the relation between workers and platforms, making possible a segmentation by country. In order to construct a professional self-presentation, many Brazilians list themselves on LinkedIn as 'employees' of these companies. This facilitates finding worker profiles linked with these global AI platforms. Our investigation found a significant number of these workers on the LinkedIn profiles of Appen and Lionbridge. On Appen's LinkedIn profile, Brazilians were the second largest nationality group (776 workers, at the time of writing), just behind Americans and followed by Filipinos, Indonesians and Indians. On Lionbridge's LinkedIn profile, 137 Brazilians were listed, representing the fourteenth largest nationality group. From these two lists of hundreds of Brazilians, we selected 63 workers that we were able to contact through a message on LinkedIn, which limits contacts to members that have connections in common.

This initial observation that Appen and Lionbridge had the highest numbers for Brazilian workers was corroborated by a search for online groups that we conducted on Facebook. We found Appen and Lionbridge workers' groups with a significant number of members: the two biggest groups had respectively more than 4,000 and 1,000 workers. In these groups, we found an intense cycle of conversations with dozens of daily posts. Most of the debates had as their subject some specificities of the common experiences of workers at these companies.

Regarding the objective of analysing work of Brazilians on global AI platforms from the workers' point of view, we conducted participant observation during the research period on the two biggest groups on this subject on

Facebook (Hewson 2014). During this period, we interacted with the groups' members, aiming to map the themes and work practices that they discuss. We also had contact with WhatsApp groups advertised on the Facebook groups that we analysed.

Then, we conducted semi-structured interviews with 16 workers through messaging applications (Brinkmann 2014). Moreover, we contacted workers listed as 'employees' in the LinkedIn profiles of Appen and Lionbridge and the members of the online groups that we interacted with that manifested interest in giving interviews. Finally, we surveyed an additional 15 workers through a questionnaire with 15 questions about working conditions. Regarding these research efforts, the sample was composed of notes and excerpts of participant observation, 16 semi-structured interviews and another survey with 15 responses.

Acknowledging the diversity of the corpus, composed of notes about participant observation, survey answers and online interviews, we conducted an exploratory analysis aiming to map recurrent issues of this environment. In this initial overview, it was possible to identify a list of notable subjects that emerged more frequently. We began by describing general aspects of these workers' points of view, and subsequently we divided the findings into five categories regarding their relation to workers' perception: hiring processes; time tracking and difficulties proving hours worked; lack of infrastructural reliability; work strategies; and, finally, their definition of work and understanding of AI. All the categories and their types are intertwined and some of them overlap. The categorisation proposed in this chapter is intended as an exploratory effort that can help in understanding the specificities of these global value chains.

Analysis

It is immediately possible to note a significant range of different jobs that are performed by these workers. Organised into different '*projects*', these '*tasks*', as these specific work activities are called, can be very different types of data production: rating advertisements, correcting intelligent assistants' responses, correcting map information, producing personal data, analysing Facebook pages' relation with real businesses, categorising images, responding to surveys, transcribing, translating, subtitling and recording audio or video, etc.

During the interviews we analyse how these workers understand advantages and disadvantages of this type of work. The main advantages observed are related to flexibility, which was the topic most verbalised. Interviewees usually presented these activities as a job that they can do in different places at different times: 'I can do it while I'm watching TV, while I'm watching some TV series. I prefer it because it is time that I can turn off my brain and still make money', said Daiana, a 25-year-old odontology undergraduate student living with her parents. Laura, 41 years old, argued that this flexibility helps her 'work without

leaving home so I can take care of my 6-year-old son and do the housework'. Another frequently mentioned advantage is the payment in dollars, given that the exchange rate (at the time of writing) shows an aggressively favourable trend, boosting their income.

The most frequently mentioned disadvantage is the lack of job security, an aspect that shapes workers' overall understanding of their activities. As we will see in next sections, this lack of stability encompasses multiple factors such as the possibility of sudden termination or non-payment of wages. Geraldo, a 50-year-old IT worker, said: 'There are some tasks that I can't understand and if my answers are classified as wrong, I will lose 50% of my wages or lose my account. So, it's nice money, in dollars, but I don't have security. The rules change a lot and we are not always informed.' This sense of continuing instability, combined with the low pay, are factors that are linked with the evidence that the majority of the workers that we interacted with consider this just a side job. Although many of them are unemployed, the compensation that they receive from these platforms generally is not enough to cover the workers' financial needs.

One important issue in the communication between Brazilian workers in Facebook and WhatsApp groups is the selection process that workers must pass before being chosen for projects on these platforms. Generally, the selection processes consist of submitting a résumé in English and passing tests about the projects' guidelines, also in English. They are seen as tough obstacles to be overcome and are perceived as somewhat mysterious: 'It's impossible that I am the only one that is being rejected all the time', says a man in a Facebook group. Another member of the group responded: 'Dude, I don't know the criterion used or if the choice is just random, but I was accepted at the Mechanical Turk when I used a new email from Outlook.'

Time Tracking and Difficulties Proving Hours Worked: 'It's Annoying to be Accused All the Time'

In the online conversations and in the interviews that we conducted, the payment process is a subject that seems to be connected with many of the workers' practices regarding work for global AI platforms, in line with research on AMT in Brazil (Moreschi, Pereira and Cozman 2020). Some platforms, such as Appen, require workers to self-report hours worked. Although it would be quite reasonable to assume that these platforms have the technological capability to monitor work, they delegate at least part of this control to the workers: each one needs to present how many hours she or he has spent doing the tasks monthly. It's common to encounter workers' narratives about rejections and disagreement about the worked hours. Daiana said '[The platforms] are very disorganized. They dispute hours of work. They say "no, no, no, according to our system you didn't work those hours." And then it's very difficult for you to prove the opposite. Sometimes, I just shut up and accept it, because

otherwise, you won't get anything.' Regarding this, workers share experiences about how to verify the activities completed during the worked hours. In a Facebook group, a woman complained: 'Seems silly, but I do everything to meet the quota of five to seven working days a week. And for some reason, their crazy system just seems to be picking on me. Because it's no exaggeration when I say that I receive at least two [system warnings] per month, and they are NEVER true. It sucks to be accused all the time.'

As the first quote of this section illustrates, there is in these conversations and narratives a sense of inevitability, that any dispute concerning these alleged false accusations is pointless and can also harm workers' relations with platforms. On a Facebook group, a man says: 'a colleague went through a similar situation in [project name]. He sent several screenshots that proved his job completion and they never accepted them, unfortunately. Finally, he was removed from [project name] at a certain time.' Another male worker summarises this understanding: 'Whether you complain or not, it doesn't matter if you have your hours recorded, in the end you'll have to accept what they want.'

Work management has historically been based on the control of working time as a fundamental resource for work organisation by companies (Wajcman 2015). In line with other research on platform labour (Woodcock and Graham 2019; Van Doorn 2017; Abilio 2020), this reveals an algorithmic management of labour that produces meanings as impartial, inevitable and unattainable, like a data gaze (Beer 2019). However, this does not mean that workers do not communicate – and organise – among them about tactics around working on global AI platforms.

Lack of Infrastructural Reliability: 'A Lot of Bugs'

Frequent lack of infrastructural reliability shapes the work processes of Brazilians in the global AI platforms. It is necessary to recognise that these technical difficulties can vary from one platform to another and between projects on the same platform. During the research process, however, we encountered different narratives about problems such as platforms' 'bugs', dysfunctional apps, connection losses and difficulties regarding payment. The first layer of instability may be observed in the structures of the systems that workers use to accomplish their tasks. In the interviews, it was possible to understand that these difficulties shape workers' practices and affect their work capacity. Tarcila, a 27-year-old law student, relates: 'I worked on a Facebook project whose application had a lot of errors, a lot of bugs… [...] I had to analyse 30 ads, and sometimes I opened the application and I had only 15. [...] And sometimes, I sent the screenshot to Appen, and they didn't accept it because I didn't have the screenshot time, bla bla bla… A lot confusion.'

In Facebook and WhatsApp groups the narratives about platforms' 'bugs' produce a range of different perceptions about work. As Tarcila described, it

is common to see stories about loss of worked hours. A female worker in a Facebook group asks if she can be removed from a project for not working due to a bug that blocked access to the classification tool: 'I haven't been able to classify ads for days because of the bug. Hence today I received the following email: "Hello, You are receiving this message because our records show that during the week of [a week], you have not completed the requirements for the [name] project". These discussions are marked by expressions of feelings such as frustration and sadness.

Linked with the narratives about lost work hours due to platform bugs is the development of tactics to overcome these technical difficulties. These tactics represent knowledge of the technical logic of these labour processes, strategies based on the workers' experiences and their communication in online groups. In the same post mentioned in the last paragraph, another worker responded to the question, giving a suggestion about how to avoid system failures: 'Girl, you have to uninstall the Facebook app from your cell and install this version that I will put here. I do this every day (uninstall and install again) because my phone updates and there is no way to prevent the update.' As this quote shows, in this context many of the conversations about bugs are based on the workers' knowledge and may reference platforms' manuals and guidelines, but also cross into the grey areas in the everyday reality of workers which are not covered by these documents.

These narratives about the frequent inefficiency of the platforms' systems move this topic away from the *high-tech* image that is generally associated with the AI debate. This can be even more evident if we consider the problems of structure that are an inherent part of many Brazilian platform workers' everyday reality, in line with the research of Moreschi, Pereira and Cozman (2020). Problems such as loss of internet connection and glitches in mobile devices are frequent and shape the workers' practices and communications. These situations are relatively common in the groups which we interacted with.

Lastly, another lack of structural reliability for this planetary labour force (Graham and Anwar 2019) is represented by the payment systems. As processes may differ significantly from one platform to another, payment constrains the workers' routines and their relations with global AI platforms. In the communication analysed in our research, these processes are depicted as complex and costly, as a dynamic that involves a variety of platforms and financial methods. Therefore, different strategies are elaborated and shared with the objective of simplifying the receipt of payment and avoiding financial losses. Daiana explains her methods for maximising gains: 'At Appen, you get paid through another platform called Payoneer that charges three dollars per withdrawal. If I don't have any bills to pay, I gather a significant amount of money and minimize withdrawals. On the Lion[bridge], I can economize using the Husky platform [...]. The traditional Brazilian banks charge high taxes to receive foreign money. This platform charges only 3.5 per cent of the amount that you will receive. So, for me this is good because banks can charge up to 25 dollars,

a significant amount considering the value of my wages.' In online groups of Brazilian workers, the discussions about these strategies are frequent. In a post where a worker asks for help avoiding the high fees of a Brazilian bank, a male worker said: 'Guys, for God's sake, do not transfer directly to the bank, no! They will rob you blind. I have been robbed by all of them [banks]. Best option is the Husky [platform].'

Workers' Strategies: 'I Don't Speak German, but, Like, I Roll With It'

As was presented in the previous sections, the production and sharing of labour strategies is a practice that shapes the dynamics of these online groups. Many of the strategies that are discussed in these groups concern increasing or maintaining workers' earnings. Given the economic crisis, work on these platforms may represent an important factor in the economic survival of these workers. To exemplify this, one of the WhatsApp groups followed during the research period, with more than 150 members, has as its title the sentence 'Online Income in the Crisis'.

In our conversations with workers' different strategic approaches were mentioned as important ways to increase earnings, such as applying to different platforms for different projects. Renata, a 46-year-old translator, explains her method: 'My strategy to increase the remuneration is trying to combine the fulfilment of several small and simple jobs with big jobs of bigger values.'

On this subject, it was possible to observe a tension concerning workers who weren't qualified to do the jobs they had applied for. Tarcila explains that she applies for many projects, and then she tries to deal with the specific knowledge necessary to accomplish the tasks: 'There is a project that is for those who speak German. I don't speak German, but, like, I roll with it... I can try to do this project at home.' In contrast, Marta, a 27-year-old PhD student, complains about those who don't speak English but apply for projects: 'You have to know at least basic English, and most of the people don't. They only see an opportunity for easy money and their work isn't of good quality.' Daiana explains: 'People lie a lot on their résumés.' Marta highlights online translation as a tool for workers that do what she calls a low-quality job: 'People think that the online translator is good enough, so they deliver unsatisfactory work.' In the online groups that we followed, this tension is still more evident. It was possible to find a significant number of posts where language knowledge emerges as a topic in dispute. In a post where a worker asks whether he can do tasks in English just using Google Translate, another worker responds: 'You can use it, but be careful because if you want to continue on the project and be renewed, it is better not to use it or always try to correct it. If not, in six months they will not renew your contract.' In another post with the same theme, another worker contests this version: 'I always used big bro Google, and it's cool, contract renewed. This is an urban legend.' In other discussions, workers complain about what they

perceive as a sense of superiority and hypocrisy from those that critique people who don't know English. One worker protests: 'I thought it was a support group about questions and not a courtroom… Asking for help doesn't mean that I don't know English or that I'm stupid. You don't know enough about people's lives to speculate like this. There are people who don't lose opportunities to make others feel like garbage, right?' Similarly, in another post a worker says: 'Let's be sincere, a large part of the people who are here don't know how to speak English. So, don't come and say that if the person doesn't know this or that she/he won't be able to pass this test, because I guarantee that everyone has used or still uses some form of machine translation.'

On the one hand, the statements above show the tensions identified by Soriano and Cabanes (2019) between 'world class work' and 'proletarianised labour' involving digital labour imaginaries in the Global South. The inequalities and struggles of human labour behind AI have a strong geopolitical dimension. To understand AI, it is necessary to think about spaces of labour. On the other hand, this reveals the material conditions of data production. Why do global IA platforms need Venezuelans to train data for self-driving cars (Schmidt 2019)? Why are Filipinos required for content moderation labour (Roberts 2019)? AI, as a techno political and economic project, is based on these inequalities in the most diverse layers of mediation.

The circulation of workers' struggles, as stated by Dyer-Witheford (2015), does not happen in the same way in all parts of the world. Platformized forms of colonisation are not just in terms of data, but how they are produced and circulated by human beings (Couldry and Mejias 2019). This means understanding AI colonialism, or how resources are expropriated from people in countries of the Global South to endorse platforms based in the Global North and their mechanisms of value extraction.

The tension concerning the usage of online translators in the AI platform workers' practices brings out the subject of data production more evidently in our analysis. The greatest part of the public discourse about AI emphasises its computational potential, which is generally portrayed as a force that can reshape society. While the role of data is positioned as a key element in the AI infrastructure, the debate about its production is still a secondary topic, and these platforms advertise their training datasets as 'reliable sources' produced by 'skilled annotators'.

However, the analysis of workers' narratives and communications presents a more complex scenario, marked by many levels of *translations/mediations*, in the philosophical sense that the Actor-Network Theory gives to these terms (Latour 2005). This point is reinforced by authors such as Beer (2019) and Couldry and Mejias (2019). Computationally, data always represents abstraction of real phenomena (Wirth 1985). Critically observed, data used in the AI industry cannot be considered natural matter that is captured 'from the world in neutral and objective ways' (Kitchin 2014, 6). This training data is produced by complex

processes that involve many different groups of workers, both high and low technologies and various socio-economic conditions. In other words, based on our analysis we argue that data produced by AI global industry platforms is shaped by these work conditions, as a reality composed of many layers of mediation.

Definition of Work and Understanding of AI: 'I Say That I Work Improving Top Secret Artificial Intelligence'

One of our specific objectives in the exploratory analysis of workers' communications was to understand how they perceive their labour practices in the context of the global AI platforms industry. Although this didn't represent a central topic in workers' conversations in the online groups that we interacted with, it was possible to encounter some discussions about it. In one post, a worker asks other members of the group: 'Guys, just out of curiosity: When someone asks you what your work is, what do you answer? I always say that I work for [platform] which is an online company, but I never go deeper... and you?' Another worker responds: 'I say that I work improving top secret artificial intelligence.'

The notions of data, algorithm and AI appeared only sporadically and in the context of other discussions, as the last quote exemplifies. We consider it possible to sustain that many workers in online groups do not see themselves as part of the AI industry, viewing their work in a more practical sense, for example, as categorising, evaluating, segmenting or correcting information, behaviour, content or ads. In our interviews, the situation was similar, but we were able to deepen the discussions. Asked whether she understands the relation between her work and AI technologies, Daiana says: 'For me, it was always very clear. I always knew I was doing it to train the companies' algorithms. They say that. They say that our work is essential to improve search engines, to train their algorithms.' In the survey responses, just one respondent spontaneously correlated his/her work with AI: 'I think I'm helping artificial intelligence systems to assimilate cultural aspects, determining funnels that help to show ads and contents to specific user profiles.' When they were questioned about how they think that their work helps to create or train AI systems, the responses focus on the idea of improving algorithms and helping users.

Conclusion

In this chapter, we described exploratory research conducted through online workers' groups on Facebook and WhatsApp with the intent to deepen and diversify the empirical analyses of the work of Brazilians on global AI platforms. This research involved a diverse corpus comprised of our notes from

participant observation and workers' survey answers, as well as online semi-structured interviews. Although the study covers a limited number of workers, the results are significant and point to some consistent trends. Based on this initial effort, we conclude that the labour dynamics of Brazilians engaged in work for global AI platforms are complex and evidence several specificities. This initial finding corroborates the hypothesis of a *taskification of labour*, as Casilli (2019) states, and we added the geopolitical dimension, in order to affirm that there is no digital labour universalism. Platform labour behind global AI companies reveals something deeper in relation to working conditions in countries like Brazil where gig is the norm and whose economy is based on informality.

As the analysis revealed, the workers' online communication represents an important practice that shapes the way they understand their work activities and the way they orient themselves in their interactions with these platforms. In other words, the knowledge that is produced and negotiated in these online environments shapes the workers' activities as tactics and strategies. The communication between workers represents a historical phenomenon associated with the labour practices. However, in the context of the platformization of labour (and consequently the isolation of the workers) this communication represents a key element of what Abilio (2020) termed 'management of survival', proving that workers aren't *unorganisable*. Platforms can be considered as means of communication and production (Williams 2005). Thus, communication helps both in the organisation and control of work and in the organisation and strategies of workers.

Finally, we consider that our research reinforces the idea that the datasets that fuel AI models need to be understood in the context of complex global chains of digital labour. The fact that computational data is an abstraction that embodies its conditions of production should prompt us to consider the various layers of mediation (some of them presented in our analysis) that these data production systems encompass, especially because these systems ultimately shape AI decision-making processes. Approaching the potential and agency of AI without considering the conditions of data labour, we sustain, represents replicate uncritical understandings that depict AI as objective, high-tech computation produced by the Global North. We assert that critical AI studies have to consider the Global South perspectives, acknowledging that data production for automation is not a homogeneous process, neither in relation to workforce composition nor to platform structure specificities. The labour dimension, we sustain, is a vital component in approaches that aim to be critical about AI. As our exploratory analysis shows, there is more negotiation, conflict and low-tech in the AI industry's Global South workforce than is presented in the global AI platforms' discourse. There is no digital labour universalism nor a homogenous workforce regarding heteromation of labour on AI platforms. Rather, there is an AI colonialism reinforcing North–South inequalities from a platform labour perspective.

References

Abilio, L. 2020. Digital Platforms and Uberization: Towards the Globalization of an Administrated South? *Contracampo*, 39 (1), 1–15.

Beer, D. 2016. *Metric Power*. London: Palgrave.

Beer, D. 2019. *The Data Gaze*: *Capitalism, Power and Perception*. London: Sage.

Brinkmann, S. 2014. Unstructured and Semi-structured. *The Oxford Handbook of Qualitative Research*, pp. 277–299. New York: Oxford University Press.

Cant, C. 2019. *Riding for Deliveroo: Resistance in the New Economy*. London: Polity.

Casilli, A. 2019. *En Attendant les Robots: Enquête sur le travail du clic*. Paris: Seuil.

Casilli, A. and Posada, J. 2019. The Platformization of Labor and Society. In: M. Graham and W. H. Dutton (Eds.), *Society and the Internet. How Networks of Information and Communication are Changing Our Lives* (2nd ed.), pp. 293–306. Oxford: Oxford University Press.

Couldry, N. and Mejias, U. 2019. *The Costs of Connection*. Palo Alto, CA: Stanford University Press.

Crawford, K. and Joler, V. 2018. Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources. New York: AI Now Institute and Share Lab (7 September). Available at: https://anatomyof.ai

Dardot, P. and Laval, C. 2013. *The New Way of the World: On Neoliberal Society*. London: Verso.

Dyer-Witheford, N. 2015. *Cyber-Proletariat: Global Labour in the Digital Vortex*. London: Pluto Press.

Dyer-Witheford, N., Kjosen, A. and Steinhoff, J. 2019. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. London: Pluto Press.

Ekbia, H. and Nardi, B. 2017. *Heteromation, and Other Stories of Computing and Capitalism*. Cambridge, MA: MIT Press.

Englert, S., Woodcock, J. and Cant, C. 2020. Digital Workerism: Technology, Platforms, and the Circulation of Workers' Struggles'. *tripleC: Communication, Capitalism & Critique*, 18 (1), 132–145.

Fuchs, C. and Sandoval, M. 2014. Digital Workers of the World Unite! A Framework to Critically Theorising and Analysing Digital Labour. *tripleC*, 22 (1), 1–20.

Gillespie, T. 2014. The Relevance of Algorithms. In T. Gillespie, P. Boczkowski and K. Foot (Eds.), *Media Technologies,* pp. 167–194. Cambridge, MA: MIT Press.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven, CT: Yale University Press.

Gillespie, T. 2020. Content Moderation, AI, and the Question of Scale. *Big Data & Society*, 7 (2), 1–5.

Graham, M. and Anwar, M. 2019. The Global Gig Economy: Towards a Planetary Labour Market? *First Monday*, 24 (4), 1–15.

Graham, M., Hjorth, I. and Lehdonvirta, V. 2017. Digital Labour and Development: Impacts of Global Digital Labour Platforms and the Gig Economy on Worker Livelihoods. *Transfer*, 23 (2), 1–15.

Gray, M. and Suri, S. 2019. *Ghost Work*. Boston, MA: Houghton Mifflin Harcourt.

Grohmann, R. and Alves, P. 2020. Unions and Associations of App-drivers in Brazil: the Meanings in Circulation of Platform Workers' Struggles. *Proceedings of the 21st Annual Conference of the Association of Internet Researchers (AoIR)*.

Grohmann, R. and Qiu, J. 2020. Contextualizing Platform Labor. *Contracampo*, 39 (1), 1–10.

Grohmann, R., Carelli, R., Abs, D., Salvagni, J., Howson, K., Ustek-Spilda, F. and Graham, M. 2020. The Uprising of Brazilian Food Delivery Riders. *Fairwork Website*, 10 August 2020. https://fair.work/the-uprising-of-brazilian-food-delivery-riders

Hewson, C. 2014. Qualitative Approaches in Internet-mediated Research: Opportunities, Issues, Possibilities. *The Oxford Handbook of Qualitative Research*, , pp. 423, 451. New York: Oxford University Press.

Huws, U. 2014. *Labor in the Global Digital Economy*. New York: Monthly Review Press.

Introna, L. 2016. Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science, Technology, & Human Values*, 41 (1), 17–49.

Irani, L. 2015. The Cultural Work of Microwork. *New Media & Society*, 17 (5), 1–15.

Kalil, R. 2019. *Capitalismo de plataforma e direito do trabalho: Crowdwork e trabalho sob demanda por meio de aplicativos*. São Paulo: USP.

Kitchin, R. 2014. Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society*, 1 (1), 1–12.

Latour, B. 2005. *Reassembling the Social*: *An Introduction to Social Life*. Oxford: Oxford University Press.

Lehdonvirta, V. 2016. Algorithms that Divide and Unite: Delocalisation, Identity and Collective Action in 'Microwork'. In J. Flecker (Ed.), *Space, Place and Global Digital Work*. Dynamics of Virtual Work. London: Palgrave.

Ludec, C., Tubaro, P. and Casilli, A. 2019. How Many People Microwork in France? Estimating the Size of a New Labor Force. *arXiv:1901.03889v1 [econ.GN]*. 12 January 2019.

Marx, K. 1894. *The Capital – Vol I*. London: Penguin.

Milan, S. 2018. Cloud Communities and the Materiality of the Digital. In: R. Bauböck (Ed.), *Debating Transformations of National Citizenship*. IMISCOE Research Series. Cham: Springer.

Milland, K. 2017. Slave to the Keyboard: The Broken Promises of the Gig Economy. *Transfer*, 23 (2), 1–15.

Moreschi, B., Pereira, G. and Cozman, F. G. 2020. The Brazilian Workers in Amazon Mechanical Turk: Dreams and Realities of Ghost Workers. *Contracampo*, 39 (1), 1–15.

Murdock, G. 2018. Media Materialities: For a Moral Economy of Machines. *Journal of Communication*, 68 (2), 359–368.

Neff, G., McGrath, M. and Prakash, N. 2020. *AI @ Work: Overcoming Structural Challenges to Ensure Successful Implementation of AI in the Workplace*. Future Says_ Report, 13 August.

Parikka, J. 2015. *A Geology of Media*. Minneapolis; London: University of Minnesota Press.

Qiu, J., Gregg, M. and Crawford, K. 2014. Circuits of Labour: A Labour Theory of the iPhone Era. *TripleC*, 12 (2), 1–15.

Reiter, B. 2017. Theory and Methodology of Exploratory Social Science Research. *International Journal of Science and Research Methodology*, 5 (4), 129.

Roberts, S. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Sadowski, J. 2020. The Internet of the Landlords: Digital Platforms and New Mechanisms of Rentier Capitalism. *Antipode*, 52 (2).

Schmidt, F. 2019. *Crowdsourced Production of AI Training Data: How Human Workers Teach Self-Driving Cars How to See*. Working Paper Forschungsförderung, No. 155, Hans-Böckler-Stiftung, Düsseldorf.

Soriano, C. and Cabanes, J. 2019. Between 'World Class Work' and 'Proletarianised Labor': Digital Labor Imaginaries in the Global South. In: E. Polson, L. Schofield-Clarke and R. Gajjala (Eds.), *Routledge Companion to Media and Class*. New York: Routledge.

Srnicek, N. 2016. *Platform Capitalism*. Cambridge: Polity Press.

Tubaro, P. and Casilli, A. 2019. Micro-work, Artificial Intelligence and the Automotive Industry. *Journal of Industrial and Business Economics*, 46, 333–345.

Tubaro, P., Casilli, A. and Coville, M. 2020. The Trainer, the Verifier, the Imitator: Three Ways in which Human Platform Workers Support Artificial Intelligence. *Big Data & Society*, 7(1). Online first.

Van Dijck, J., Poell, T. and De Waal, M. 2018. *The Platform Society: Public Values in a Connective World*. Oxford: Oxford University Press.

Van Doorn, N. 2017. Platform Labor: On the Gendered and Racialized Exploitation of Low-income Service Work in the 'On-demand' Economy. *Information, Communication & Society*, 20 (6), 898–914.

Van Doorn, N. and Badger, A. 2020. Platform Capitalism's Hidden Abode: Producing Data Assets in the Gig Economy. *Antipode: A Radical Journal of Geography*, 52 (5), 1475–1495.

Wajcman, J. 2015. *Pressed for Time: The Acceleration of Life in Digital Capitalism*. Chicago, IL: University of Chicago Press.
Williams, R. 2005. *Culture and Materialism*. London: Verso Books.
Wirth, N. 1985. *Algorithms and Data Structures*. Upper Saddle River, NJ: Prentice Hall.
Woodcock, J. 2019. *Marx at the Arcade: Consoles, Controllers, and Class Struggle*. London: Haymarket Books.
Woodcock, J. and Graham, M. 2019. *The Gig Economy: A Critical Introduction*. Cambridge: Polity.

CHAPTER 15

Towards Data Justice Unionism? A Labour Perspective on AI Governance

Lina Dencik

Introduction

The advent of data-centric technologies has in recent years culminated in the hype surrounding 'Artificial Intelligence' (AI), widely seen to propel transformations across areas of science, government, business and civil society. These transformations are often simultaneously touted as enhancing forms of efficiency and better decision-making at the same time as presenting significant societal challenges. The question of jobs and the changing nature of work has been one prominent area where AI is said to have dramatic implications. Workers are subjected to evermore data collection about not just their activities at work, but beyond factors related to work. At the same time, machine learning systems are using this data to transform how work is being allocated, assessed and completed. Often it is these two components – data collection and machine learning – that is referred to under the banner of AI (Sánchez-Monedero and Dencik 2019). This has a profound impact on workers' lives, the nature of jobs and the economy. Moreover, the position of labour in relation to AI brings to light the social stratifications embedded within and created by the advancement of AI across social life. AI extends long-standing debates on modes of capitalism that significantly shape the circumstances of working people whilst limiting their ability to influence decisions that govern their lives.

How to cite this book chapter:
Dencik, L. 2021. Towards Data Justice Unionism? A Labour Perspective on AI Governance. In: Verdegem, P. (ed.) *AI for Everyone? Critical Perspectives.* Pp. 267–284. London: University of Westminster Press. DOI: https://doi.org/10.16997/book55.o.

Yet, in advancing governance frameworks that may contend with the challenges of data infrastructures and AI, there has been a notable absence of worker voices, unions and labour perspectives. Labour concerns have predominantly focused on the immediate threat of job losses and changing forms of work, but these have rarely been translated into AI governance debates more broadly. Instead, we have seen governance frameworks emerge that centre on questions of individual rights, data protection, ethics and fairness, privileging citizen and consumer rights over workers' rights. These frameworks tend to engage at a level of abstract principles that centre on the nature of technology rather than the conditions of injustice in which technology is situated, and struggle to account for AI as an outcome of power dynamics and interests that serve to shape social relations. The absence of labour perspectives within these debates, and voices that extrapolate from workplace struggles to a wider engagement with AI in the context of advanced capitalism, is therefore a significant gap in the context of what we might refer to as data justice; an understanding of datafication in relation to social justice (Dencik, Hintz and Cabel 2016).

In this chapter, I argue for the need for 'data justice unionism', a form of social justice unionism that engages with data-centric technologies as firmly situated within a workers' rights agenda and that approaches AI governance as informed by the labour movement in solidarity with other social movements. I start by briefly outlining how AI relates to issues of labour before going on to discuss a range of dominant frameworks for AI governance. These have tended to exclude broader labour concerns and often frame what is at stake with AI in terms of trade-offs between economic gains and individual rights that bypass an engagement with collective rights and more fundamental questions about the political economy of AI. I then go on to discuss key issues in AI that a labour perspective foregrounds, in the workplace and beyond, before situating these in relation to data justice unionism. As a component of social justice unionism that argues for unions working in coalition with other social movements to advance a more just society, data justice unionism makes an explicit connection between digital rights and socio-economic rights and contends with AI in the context of labour relations under capitalism. This needs to inform much more of current mobilisation efforts around data justice in order to, on the one hand, elevate the relevance of the labour movement, and, on the other, for AI governance debates to better account for lived experiences and actual social struggles.

Labour and AI

The implications of the advent of AI for labour and labour relations has garnered much attention in recent years, building on long-standing debates on the transformative potentials of emerging technologies. Whilst some have argued that the rapid development of data-centric technologies signal a fundamental

shift in the operations of capitalism, others have focused on how these technologies entrench or extend particular features of capitalism that significantly impact on the nature and experiences of work. It is not the aim to discuss these different perspectives in detail here, but it is worth briefly outlining some of the ways in which labour concerns manifest in understandings of AI in order to understand their significance for AI governance debates. At one level, these concerns are often focused on the changing nature of the workplace itself and how the implementation of AI systems impact on employment relations and working conditions. Algorithmic management of the kind associated with the gig economy, for example, is rapidly becoming embedded within larger parts of the labour market, stretching across different kinds of workplaces (Kellogg et al. 2020). Devices and tools such as phones, laptops and emails are subject to monitoring, whilst data extracted from social networks, shared calendars and collaborative working tools are being integrated to gain insights into not only professional activities, but also who workers are or what they might do in the future. More and more, this is being complemented by sensory and recognition tools such as chips and wearables that dictate tasks and assess emotional and physical states as part of work performance (for an overview of data-driven technology in the workplace see Sánchez-Monedero and Dencik 2019). Based on this perpetual generation of data within and beyond the workplace, AI systems promise to automate key aspects of the labour process and management techniques. For some, these developments continue a trajectory of automation that has long been seen as a threat to labour in different ways, moving beyond the production process of the industrial era to also include information processing and decision-making (Andrejevic 2019). Key concerns have been raised about how the use of AI technologies in the workplace might displace jobs, impact on workers' rights and undermine labour power (Moore, Upchurch and Whittaker 2017).

Moreover, engagement with the AI-labour nexus has provided impetus for the on-going debate on the implications of emerging technologies for transformations in capitalism. Early accounts of the advent of information and communication technologies (ICTs), for example, indicated a significant shift in the relationship between capital and labour and emphasised value extraction outside of production as the principal location of the process of valorisation. Notions such as 'immaterial labour' and 'cognitive capitalism' (Moulier Boutang 2011) point to an elevated significance of knowledge, information and intellectual property over labour as traditionally understood in operations of capital that, for some, promises new visions of 'post-capitalism' (Mason 2016) and the possibilities for a 'world without work' (Srnicek and Williams 2016). Whilst these accounts have been criticised for lacking sound empirical basis and often underestimating the continued centrality of production and extraction of value from labour in supply chains (Thompson and Briken 2017), they point to particular processes in capitalism that have found resonance in more recent accounts of the value of data and the political economy of data-centric technologies such as AI.

In Zuboff's (2015, 2019) notion of surveillance capitalism, for example, value generation relies not on a division of labour, but a division of learning: between those who are able to learn and make decisions based on global data flows, and those who are (often unknowingly) subject to such analyses and decisions. In this model, capital moves from a concern with incorporating labour into the market as it did under previous forms of capitalism, to a concern with incorporating private experiences into the market in the form of behavioural data. This is an accumulation logic driven by data that aims to predict and modify human behaviour as a means to produce revenue and market control. Yet in thinking about the value of data and the social relations that emerge from its extraction, Sadowski (2019) argues that we need to understand it not just as a commodity but as capital that propels new ways of doing business and governance in what he sees as the 'political-economic regime' of datafication. Data collection is driven by the perpetual cycle of (data) capital accumulation, which in turn drives capital to construct and rely upon a universe in which everything is made of data. The digital platform is central for this transformation in that social practices are reconfigured in such a way that enables the extraction of data (Couldry and Mejias 2018). Data in this context serves to sustain an economic process that relies on capturing value through expanding the capacity for gaining information rather than creating value through production. This process does not break with how we might understand capitalism, but rather positions datafication as the extension of financialisation and the drive to turn everything into a financial asset. The aim is to latch onto circuits of capital and consumption for the purposes of rent extraction, whether in monetary form or as data (Srnicek 2017; Sadowski 2020).

The role of labour, under these conditions, is characterised by what Van Doorn and Badger (2020) call 'dual value production': the monetary value produced by the service provided is augmented by the use and speculative value of the data produced before, during, and after service provision. As they go on to explain, the value of data derives in part from its expected or actual practical utility (achieving functional goals and systems optimisation) but also from the expectation of data-rich companies to achieve competitive advantage and thereby attracting venture capital and higher financial valuations. AI is part of a suite of complex technologies that have been designed to extend and empower capital's abilities of assetisation, extraction and enclosure (Sadowski 2020) and that is rooted in the positive feedback loop between a data-producing labour process and algorithmic systems that self-optimise as they analyse this data (Van Doorn and Badger 2020). For Wark (2019), this constitutes a power shift from the owners of the means of production to the owners of the vectors along which information is gathered and used, what Wark describes as the 'vectorialist class'. This class controls the patents, the brands, the trademarks, the copyrights, and most importantly the logistics of the information vector. As such, Wark contends, whilst a capitalist class owns the means of production, the means of organising labour, a vectorialist

class owns the means of organising the means of production. Importantly, this does not necessarily make away with the exploitation of labour in value chains. As Srnicek (2020) has pointed out, AI systems rely not just on vast amounts of data, but on significant computational power and control over labour to drive monopolisation. We have a growing economy based on what Gray and Suri (2019, ix) refer to as 'ghost work': a new digital assembly line that aggregates the collective input of distributed workers, ships pieces of projects rather than products, and operates continuously across a host of economic sectors in order for AI systems to function.

The implications of AI for labour therefore extend from the workplace to the reorganisation of employment through to the operations of capital upon which AI depends and advances. The use of AI in automated hiring systems, performance assessment tools, scheduling, and other forms of algorithmic management in the workplace (platforms or otherwise) intersect with broader transformations in the economy and dynamics of capitalism in which developments in AI are embedded. These different concerns highlight the many complex and intricate ways AI impacts on the experiences of working people, the way their work is organised and how it is valued, and their ability to influence decisions that govern their lives. Yet, as I will go on to outline below, workers' voices and union perspectives have been notably absent from AI governance debates that have instead overwhelmingly championed liberal frameworks based on citizen and consumer rights. If we are to contend with AI in relation to the advancement of a more just society, then such frameworks are insufficient.

Governing AI

The advent of AI has sustained much discussion about what is actually at stake with the growing datafication of social life. Whilst there is widespread recognition that the rapid development and deployment of data-centric technologies has significant transformative implications, the question as to what these are and how they should be addressed is still up for grabs. Gangadharan (2019) has provided a useful overview of different frameworks for data governance that highlights some of the dominant ways in which AI and data-driven systems in general have been approached in governance debates, including privacy policy, data protection, ethics, fairness-in-design and human rights. Elaborating on this overview, I argue in this section that mobilisation around the governance of emerging technology, particularly AI, has thus far been situated in a digital rights and technology-driven agenda that has foregrounded individual rights and focused on the nature of the technology itself. Lacking from this agenda has been a more substantial engagement with collective rights and the actual conditions of injustice and lived experiences of struggle within which technology is embedded. The labour movement has an important role to play bringing such a perspective forward within the AI governance space.

Initial concerns over the mass generation and analysis of data collection have tended to highlight issues of surveillance and privacy, prominent in public debate particularly in the immediate aftermath of the Snowden leaks first published in 2013 (Hintz, Dencik and Wahl-Jorgensen 2018). In part, these events made clear the limitations of existing legislation around privacy, and the need for more oversight in the handling and processing of data by different actors. This saw the flourishing of a range of technology and policy initiatives aimed at restricting data gathering, such as the development of privacy-enhancing tools, mainstreaming the use of encryption and lobbying around anti-surveillance issues (Dencik, Hintz and Cable 2016). These have advanced important repertoires for resistance that directly challenge the power relations of data-driven surveillance and have provided avenues for individuals to manage aspects of their digital engagement. However, the advancement of technological self-defence as a governance frame is also limited by the onus on the individual user to protect their own privacy. As Ruppert, Isin and Bigo (2017) describe it, many accounts of data politics are premised on an ontology of 'hyper-individualism' that nurtures a suggestion that 'ultimately it is up to you to change your behaviour to protect yourself from the dark forces of the internet'.

In translating some of the concerns of anti-surveillance resistance into regulation, the protection of personal data has been a particularly noteworthy frame for governance, such as the approach to the General Data Protection Regulation (GDPR) adopted by the EU in 2018. The premise is that individuals should be able to claim some rights with regards to information collected about their person, and that collecting such information requires some form of consent. In this sense, it privileges the individual data subject and understands the protection of personal data as distinct from, but complementary to, individual privacy. The GDPR is relatively broad in scope but it is worth noting that issues pertaining particularly to the workplace and the processing of data on workers were excluded from this regulation in its final stages (Colclough 2020). Rather, the GDPR predominantly favours an understanding of data subjects as individual citizens and consumers that are afforded certain rights about their ability to access, challenge and limit data collected about their person by private companies and parts of the public sector.

Although the GDPR has paved the way for engaging with data-centric technology in a broader sense, questions remain about both its scope and enforceability. Perhaps in part as a response, much attention and resources have been dedicated to advancing 'data ethics' and 'AI ethics' in recent years as alternative and complimentary governance frameworks. This field has engaged a range of different streams of thought and practice, some of which continue a long-standing tradition of computer ethics while changing the level of abstraction of ethical enquiries from an information-centric to a data-centric one (Floridi and Taddeo 2016). That is, the focus shifts from a concern with how to treat information as an input and output of computing to a focus on how people access,

analyse and manage data in particular, not necessarily engaging any specific technology but what digital technology manipulates (Taylor and Dencik 2020). Often this has privileged concerns with the responsible handling of data that considers risks to privacy, forms of discrimination and abuse, ensuring transparency and accountability. In translating this into practice, we have seen the proliferation of various initiatives across industry, government and civil society framed under 'ethics' that set out different guidelines and procedures that attend to the development, handling and deployment of data-centric technologies, particularly AI. Government initiatives such as the UK's Centre for Data Ethics and Innovation and the establishment of high-level expert groups on ethics within the EU have advanced some avenues for outlining ethical concerns in relation to technology, whereas civil society actors have turned to data ethics as a way to advance data developments 'for good' across a range of contexts. Of particular note has been the active engagement by the technology sector itself in this governance frame, swiftly setting up associations, creating guidelines and codes for the responsible handling of technological innovation. An early offering came in the form of the Partnership on Artificial Intelligence to Benefit People and Society set up by Amazon, Google, Facebook, IBM and Microsoft in 2016 as a non-profit organisation to advance 'best practices and public understanding'. Most of these companies have also subsequently attempted to set up their own ethics boards, sometimes in partnership with academics, with varying degrees of success (Naughton 2019).

While a focus on data and AI ethics has foregrounded some prominent concerns about data collection and use in a way that shifts the onus of responsibility onto developers and the data controller, it is not clear that these initiatives have resulted in any real intervention. Government entities have predominantly been set up as nominal oversight bodies without any real teeth to interfere, leaving civil society actors having to levy at the abstract level of principles and rely on the goodwill of the industry to uphold them. Corporate data ethics initiatives, meanwhile, have focused on 'micro-ethics', an orientation around the individual practitioner, and an emphasis on compliance that avoids any fundamental engagement with the bottom line or premise (Taylor and Dencik 2020). In some instances, this has led to accusations of 'ethics-washing' (Wagner 2018), allowing for technology companies to engage with public concerns about their activities, while continuing to avoid regulation or any major challenge to the business models that sustain them. Moreover, by actively capturing the ethics space, the very players who are creating, developing and directly profiting from these technologies have also been the ones dictating the terms upon which we are to understand both the nature of problems and what might be suitable responses. Unsurprisingly, therefore, the application of ethical frameworks within the technology sector has tended to concern itself with the actual data-sets or algorithms themselves, positing that the causes of harms that emerge from AI can be traced to 'errors' or 'bias' in the design and

application; causes that essentially have technological solutions, preferably through further data collection and algorithmic sophistication. We see this for example with the growing industry that now concerns itself with 'fairness' in the design of systems, creating more inclusive data-sets and algorithms that can account for more diverse experiences, or the development of 'bias mitigation' tools (Zelevansky 2019). Such projects have drawn attention to some of the contentious assumptions that are embedded in the design of technological systems, but have also been accused of advancing technical fixes that serve to legitimise the industry (Gangadharan and Niklas 2019).

The growing debate surrounding ethical challenges and the bias of algorithmic processes has helped spur on an engagement with data-driven technologies as socio-technical systems that have an impact on people's lives. Some of this is evident in emerging forms of regulation on AI, for example, the emphasis on 'Trustworthy AI' and a risk-based approach to minimising harms in AI systems at EU level (Niklas and Dencik 2020). It has also been prevalent in discussions on the future of work, for example, by attending to the ways in which hiring systems or other parts of automation in human resources might discriminate against particular groups (Ajunwa 2019; Graham et al. 2020). However, concerns about ethics washing and the tendency towards technical fixes have led to calls to centre rights, and particularly human rights, more firmly within these discussions. Drawing on human rights legislation in AI governance debates goes beyond issues of privacy and the protection of personal data whilst providing a sturdier point of reference than abstract principles of ethics and fairness. Using international human rights as a frame in relation to the governance of AI details the specificity of potential harms by linking them to particular rights, such as the right to freedom of association or the right to a fair trial, that can apply to different parts of social life (HRBDT 2020). These assertions of rights can help inform impact assessments, for example, when new AI systems are being developed or deployed (Jørgensen et al. 2019; Jansen 2020). By relying on universal terms of reference, a human rights framework is also effective for advocacy as an internationally recognised agreement, however much this may not play out in practice. A recent court case brought forward by NGOs in the Netherlands, for example, to challenge the use of data-centric technology in the welfare sector won on the basis that it was considered an infringement on human rights and supported on-going efforts by the human rights community to demand assessments of AI systems beyond the required initial data protection impact assessment (Toh 2020).

Governing AI from a human rights perspective can therefore provide an avenue for a more holistic engagement with data-driven systems that considers a broad range of rights that pertain to people's lives. However, the notion of international human rights has historically struggled to translate into successful concrete action, often seen to be at the whims of geopolitical concerns and international relations. Moreover, as a framework, it has traditionally centred on the individual and civil and political rights in a way that has struggled to

account for collective rights and that has tended to neglect social and economic rights (Alston 2005). In general, in line with how governance debates on AI have predominantly been approached, there is a lack of political mobilisation that can contend with the power relations that are inherent in the advancement of AI and that engages with datafication as a political economic regime. The absence of labour concerns is an important aspect in this respect. In part, this is a result of a deliberate exclusion of worker voices in governance debates, both in how they have been organised as well as in terms of institutional structures surrounding AI governance. It is noteworthy how even discussions on AI in relation to the future of work have been advanced around industry and citizen concerns, but with an absence of unions or other worker associations. At the same time, the lack of labour perspectives in relation to AI governance is also indicative of a labour movement that has been slow to engage with questions of data and data-centric technologies on a societal scale. As I will go on to argue, unions and labour activists have predominantly (understandably) focused on immediate concerns regarding the changing nature of work and the workplace, particularly with the advent of the gig economy and automation. This focus has brought to light some significant issues with regards to the power of technology companies in setting the terms of work and workers' worth, but it has not translated into mobilisation efforts around AI governance, the way in which AI positions labour in relation to capital, and how this informs the advancement of social justice. Instead, as I will go on to discuss below, other actors and communities that could benefit from alliances with the labour movement have driven mobilisation around this kind of data justice.

Data Justice and Labour

Privileging a concern with social justice in relation to datafication, a framework of data justice, is part of a notable shift in the framing and understanding of what is at stake with the growing development and use of data-centric technologies such as AI. In part a response to the rather limited interpretations that have informed governance debates thus far, data justice advances a research agenda that seeks to change the terms of the debate, situating data in relation to structural inequalities and histories of domination (Dencik, Jansen and Metcalfe 2018). This has, in some interpretations, led to articulations of principles to underpin data governance that can better account for such inequalities (Heeks 2017; Taylor 2017), or practices in the handling of data that make asymmetries in the representation and power of data explicit (Johnson 2018). In other interpretations, conceptions of justice have been foregrounded in the development of design, calling for more participatory processes that involve communities to build alternative infrastructures that empower rather than oppress marginalised groups (Costanza-Chock 2018). This is in line with a more general recognition of the need to shift what voices are centred in any understanding

of what is at stake with datafication and challenge the current constitution of the decision-making table. Gangadharan and Niklas (2019), for example, have made the case for 'decentering' technology in data justice debates and instead situate technology within systemic forms of oppression, meaning that harms that emerge from data-driven systems need to be articulated by those who are predominantly impacted and understand the history of such oppression.

We have seen some of these tenets of data justice debates translate into different forms of activism and campaigning. The Center for Media Justice in the United States, for example, have created a Data Justice Lab dedicated to thinking through ways to bridge research, data, and movement work relating to issues like surveillance, carceral tools, internet rights and censorship. The Detroit Digital Justice Coalition has worked with local residents to identify harms that emerge through the collection of data by public institutions, situating these in the context of on-going criminalisation and surveillance of low-income communities, people of colour and other targeted groups. In some instances, these activities have foregrounded a politics of refusal (Gangadharan 2019) that advance an abolitionist agenda as articulated by groups such as the StopLAPD Spying Coalition and the Data for Black Lives initiative. Here, the focus is not to make technologies more efficient, but rather to recognise how technology has meaning and impact in relation to the inequalities manifested in capitalist exploitation and a history of state violence. The call is to divest resources into oppressive data systems and to 'abolish big data' that is used to measure and profile people, and instead reinvest in communities (Benjamin 2019; Crooks 2019). In the context of environmentalism, the Environmental Data & Governance Initiative (EDGI) has preserved vulnerable scientific data in the aftermath of the US election of Trump in 2016, and in the process developed an 'environmental data justice' framework that considers the politics, generation, ownership and uses of environmental data. Similarly, in the context of municipalities, efforts to engage citizens in the control over urban public data have been central to the 'Roadmap to Technological Sovereignty' advanced in cities such as Barcelona, outlining ways to challenge the monopolisation of data by a few corporate platforms. These efforts tend to focus on forms of governance that include formats such as data trusts or data commons and that allow for platforms to be managed by the city itself (Tieman 2017; Fuster 2017).

Concerns with data justice therefore translate into a range of different debates and practices that find expression across areas of society. Yet whilst these activities speak to shared interests towards addressing inequalities, redistribution and conditions of injustice, labour concerns have often been on the margins of these efforts or have been pursued in siloes. There has been a considerable effort to address the issue of potential job losses in the face of automation, for example, with unions pushing for more avenues to pursue reskilling within jobs, changing union structures to accommodate for non-trade or non-sector specific memberships, and advocating for more support for transitioning

within workplaces (Colclough 2020). Concerns with job losses have also mobilised greater support for some form of universal basic income or other kinds of safety nets for workers who are displaced by automation (Standing 2019). More recently, there has been a growing focus on how technologies such as AI impact not just displacement, but the quality of work. This includes efforts to apply the GDPR in the workplace as a way to address issues of labour protection (Aloisi and Gramano 2019) and the potential for collective rights to form a greater part of the AI and data governance debate. De Stefano (2018), for example, has argued for the need for a 'human-in-command' approach that would involve collective regulation and social partners in governing automation and the impact of technology at the workplace. Similarly, some unions are pushing for 'new technology agreements' (NTAs) to form part of collective bargaining agreements in workplaces that have union representation. Under the terms of these, new technology will only be introduced with the agreement of the trade union, and if the employer agrees to reinvest any cost savings to provide new jobs elsewhere in the organisation (Cole 2019).

Furthermore, spearheaded by smaller and independent unions, the labour movement has been increasingly active in the area of platform labour and the gig economy (not all of which deploy AI), focusing particularly on the nature of these platforms as employers. Unions such as the Independent Workers of Great Britain (IWGB) and what is now the App Drivers and Couriers Union (ADCU) have successfully challenged the status of gig workers, such as those driving for Uber, as self-employed rather than as employees. Out of these struggles, there has also been a growing engagement with the collection and use of data by these platforms to manage or direct workers employed by them. Worker Info Exchange, for example, a non-governmental organisation that grew out of organising Uber drivers within the IWGB, explicitly concerns itself with 'data rights' and the ability for workers to access data collected about them as a way to increase transparency about their management. In parallel to this, a growing mobilisation effort has formed around alternative models of platform labour that draws inspiration from the cooperativist movement. Platform cooperativism as an idea and practice has, in the space of a few years, grown globally as a response to the dominance of platform capitalism. Under this model, platforms are generally based on decentralised forms of governance in which workers themselves own the platform and/or set the terms for how it should be run. Importantly, these have sometimes been established with the direct support of labour unions and have been an avenue through which to engage the labour movement more directly in data debates. As Scholz (2017) has argued, platform cooperativism should not be considered a technological solution but a 'mindset' that includes technological, cultural, political and social changes, bringing together different actors and stakeholders.

The growing arena of data justice therefore has much scope to incorporate labour concerns in how to articulate both what is at stake with datafication

as well as possible responses. In terms of mobilisation, the challenge remains how to integrate issues of data-centric technologies like AI into a broader understanding of the place of technology in advancing social justice for all. As I go on to argue below, this needs greater cooperation and solidarity between the labour movement and other social movements in order to strengthen and advance the kind of political engagement with AI governance that a data justice agenda demands.

Towards Data Justice Unionism

As I have argued so far, dominant governance frameworks relating to data and AI have overwhelmingly responded to concerns with implications for people as citizens and consumers over and above people as workers. Moreover, they have predominantly followed a liberal orientation that has centred on individual rights and ethics. This is perhaps unsurprising considering that much of the mobilisation efforts engaged in questions of data and AI have been those stemming from digital rights and civil liberties concerns. Whilst this has mobilised a number of key issues that have been translated into important legislation, particularly in Europe, such as the GDPR and more recently, the White Paper on EU Strategy for AI, such frameworks have some important limitations for engaging with the broader implications of the turn to data infrastructures across social life. The growing activities surrounding data justice have broadened and shifted the terms of engagement in ways that seek to address some of these limitations. Yet labour concerns regarding AI have often been pursued separately from these activities. This is a challenge for broad political mobilisation as the labour movement has historically played a significant role in connecting transformations in work to broader questions of society that have relevance for the governance of data and AI. In this final section, I therefore make the case for data justice unionism to be considered as a part of social justice unionism focused on engaging labour perspectives in the debate on AI governance, including a concern with the interests driving datafication, the forms of social and economic organisation that enable them, and how they might be challenged.

Social justice unionism has become an increasingly popular approach within the labour movement and advocates for unions to collaborate with social movements in order to work towards wider goals and the resolution of workplace issues. The argument is that unions should accept the reality that there are multiple forms of oppression and that they should work with groups in coalitions to challenge them (Healy et al. 2004; Dencik and Wilkin 2015). This often means an emphasis on more networked and informal relationships between individuals, groups and organisations that combine to undertake forms of collective action. A prominent example is the protests in Seattle in 1999 that brought together a diverse array of social movement groups with trade unions to protest against the specifics of the WTO proposals but also against the

growth of corporate power and the destruction of democracy, a much broader theme uniting these movements and providing grounds for a coalition to be built on labour-related interests (Wilkin 2000). More recently, the Occupy movement brought together union organisers with a range of different social movement groups to draw attention to uneven wealth distribution and income inequality that formed grounds for demands of a living wage and job security in precarious sectors historically neglected by mainstream trade unions (Dencik and Wilkin 2015). These kinds of mobilisations focus on improving the lives of working people through engaging with class-wide or social justice demands, which include traditional 'bread and butter' issues, but are not limited to them (Behrent 2015).

Social justice unionism therefore resonates with the argument that unions need to present a picture of a good society that can be built through cooperation, solidarity and mutual aid alongside other progressive social movements. This understanding of unionism has gained particular relevance in light of declining labour power coupled with the nature of the global challenges confronting working people. Climate change, for example, presents a complex challenge for the labour movement that cannot be fought along traditional lines. The recent push for a Green New Deal and a Just Transition is in part an attempt to foster new alliances between movements concerned with labour and the environment. This includes unions and campaigners teaming up to advance concrete alternatives to a fossil-fuel-based economy, while advocating that the government take action. Indeed, Bergfeld (2019) has argued that what is needed is a kind of 'climate-justice unionism' to address the intertwined social and ecological crises in a holistic way. Such an approach would use the organisational and institutional leverage of unions to rebuild workers' power at the workplace and at company level to regulate from below, whilst at sectoral level use collective agreements to refit companies, with the goal of reducing carbon emissions and enhancing labour standards. Importantly, climate-justice unionism would involve organising 'the whole worker' (McAlevey 2016) in which issues are not only rooted in workplaces but also in communities and society, such as the disproportionate impacts of exposure to and taxation of CO_2 emissions.

A concern with data justice in this context provides a further component that needs to be part of these efforts to address issues confronting working people within and beyond the workplace, privileging a view of unions as working in solidarity with other groups. The engagement with questions of data from a social justice perspective cannot be confined to digital rights, civil liberties or technologists, but requires a coalition of individuals, groups and movements. Unions have an important role to play in this respect, not just by explicitly connecting digital rights to social and economic rights, but perhaps more importantly by articulating concerns that are rooted in people's lived experiences of AI. This can help mobilise around actual and on-going social struggles informed by those who are the most impacted as a key component of current data justice debates. Furthermore, unions can leverage power within

the workplace to address the deployment of AI systems that can inform governance debates around AI more broadly. McQuillan (2020), for example, has advocated for workers and people's councils to advance situated knowledge as a form of interference in relation to AI, drawing on the social histories of workplaces and communities. Unions might also help to organise workers within technology companies as pursued by groups such as the Tech Workers Coalition that sees labour organising as a way of advancing solidarity between software engineers and social justice movements to undermine the development of harmful technologies. More broadly, data justice unionism provides an avenue for mobilising around AI that engages with the political economy upon which its advancement relies. By attending to the operations of capital in datafication and its positioning of labour, we are forced to move away from a focus on the responsible handling of data or to turn to the realm of moral conscience or market solutions as governance responses. Instead, we need to contend with the actual conditions of injustice that shape contemporary social relations, how AI shifts dynamics of power and approach questions of technology as part of alternative visions for how society should be organised. This requires coordinated efforts between the labour movement and other social movements.

Conclusion

The advent of datafication has culminated in recent discussions on AI, bringing to light the significant ways in which data-centric technologies are intersecting with various aspects of social life. A particular area of concern is the way labour relations are transforming with the growing development of AI. This has often focused on the risk of job losses to automation and the changing nature of the workplace, both in standard and non-standard employment. It has also incorporated an analysis of the way labour, often side-lined or made invisible, is central to sustaining AI systems at the same time as the mode of capital advanced by AI undermines labour power by extending and empowering capital's abilities of assetisation, extraction and enclosure. Yet in mobilising governance frames to contend with datafication and AI, there has been a noticeable absence of workers' voices and labour concerns. Instead, dominant frameworks of AI governance have tended to focus on citizen and consumer rights that have centred on the individual and on the ethical considerations that need to inform design and deployment. Labour concerns, meanwhile, have been pursued in separate arenas that have tended to focus on specific aspects of work and the workplace, but that have often not connected with broader debates on data. As AI comes to have increasing significance for how society is organised, there is a need to foster greater cooperation between different movements and groups to engage with data justice in a meaningful way. Unions can benefit from a more holistic form of organising that extrapolates workplace issues into society in order to gain relevance and advance the interests of their

membership. Engaging with data issues needs to be part of that organising. At the same time, unions bring particular leverage to existing efforts to advance social justice concerns in the context of AI by privileging lived experiences and foregrounding collective social and economic rights. Data justice unionism, therefore, is a way of pointing to the potential for a broader political mobilisation around the role of AI in society that involves the efforts and voices of actual working people. Such a mobilisation is urgently needed if we are to contend with the shifting power dynamics that are being advanced by the growing reliance on AI in our lives.

Acknowledgement

The research for this chapter was made possible with funding from a European Research Council Starting Grant for the project DATAJUSTICE (grant no. 759903).

References

Ajunwa I. 2019. Platforms at Work: Automated Hiring Platforms and Other New Intermediaries in the Organization of Work. In Greene, D., Steve P. S. P. Vallas and A. Kovalainen (Eds.), *Work and Labor in the Digital Age* (Research in the Sociology of Work, Vol. 33 , pp. 61–91). Bingley: Emerald Publishing.

Aloisi, A. and Gramano, E. 2019. Artificial Intelligence is Watching You at Work. Digital Surveillance, Employee Monitoring and Regulatory Issues in the EU Context. *Comparative Labor Law & Policy Journal,* 41(1).

Alston, P. 2005. Assessing the Strengths and Weaknesses of the European Social Charter's Supervisory System. In G. de Búrca, B. de Witte and L. Ogertschnig (Eds.), *Social Rights in Europe*. Oxford: Oxford University Press.

Andrejevic, M. 2019. *Automated Media*. New York and London: Routledge.

Behrent, M. 2015. The Meaning of Social Justice Unionism. *SocialistWorker.org*. Available at: https://socialistworker.org/2015/05/18/the-meaning-of-social-justice-unionism

Benjamin, R. 2019. *Race After Technology*. Cambridge: Polity.

Bergfeld, M. 2019. From Fridays for Future to a Global Climate-Justice Unionism? *Social Europe*. Available at: https://www.socialeurope.eu/from-fridays-for-future-to-a-global-climate-justice-unionism

Colclough, C. 2020. Empowering Workers in the Digital Future. Podcast. *Exponential View with Azeem Azhar*. Available at: https://www.exponentialview.co/podcast

Cole, P. 2019. Working with Unions to Introduce New Technology. *People Management*. Available at: https://www.peoplemanagement.co.uk/experts/legal/working-with-unions-to-introduce-new-technology

Costanza-Chock, S. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science* 3(5).

Couldry, N. and Mejias, U. 2018. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television and New Media*, 20(4), 338.

Crooks, R. 2019. What we Mean When we Say #AbolishBigData2019. *Medium*, 22 March. Available at: https://medium.com/@rncrooks/what-we-mean-when-we-say-abolishbigdata2019-d030799ab22e

Dencik, L., Hintz, A. and Cable, J. 2016. Towards Data Justice? The Ambiguity of Anti-surveillance Resistance in Political Activism. *Big Data & Society*, 3(2): 1–12. DOI: https://doi.org/10.1177/2053951716679678

Dencik, L., Jansen, F. and Metcalfe, P. 2018. A Conceptual Framework for Approaching Social Justice in an Age of Datafication. Working Paper, DATAJUSTICE project. Available at: https://datajusticeproject.net/2018/08/30/a-conceptual-framework-for-approaching-social-justice-in-an-age-of-datafication

Dencik, L. and Wilkin, P. 2015. *Worker Resistance and Media: Challenging Global Corporate Power in the 21st Century*. London and New York: Peter Lang.

De Stefano, V. 2018. 'Negotiating the Algorithm': Automation, Artificial Intelligence and Labour Protection. Employment, Working Paper No. 246, International Labour Office.

Floridi, L. and Taddeo, M. 2016. What is Data Ethics? *Philosophical Transactions of the Royal Society*, 374(2083). Available at: http://rsta.royalsocietypublishing.org/content/374/2083/20160360

Fuster, M. 2017. Analytical Framework of the Democratic and Procommons Qualities of Collaborative Economy Organizations. DECODE blog, 6 October 2017.

Gangadharan, S. P. 2019. What Do Just Data Governance Strategies Need in the 21st Century? Keynote at Data Power Conference, 12- 13 September, Bremen, Germany.

Gangadharan, S. P. and Niklas, J. 2019. Decentering Technology in Discourse on Discrimination. *Information, Communication & Society*, 22(7): 882–899.

Graham, L. et al. 2020. Artificial Intelligence in Hiring: Assessing Impacts on Equality. *Institute for the Future of Work*. Available at: https://static1.squarespace.com/static/5aa269bbd274cb0df1e696c8/t/5ea831fa76be55719d693076/1588081156980/IFOW+-+Assessing+impacts+on+equality.pdf

Gray, M. and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Underclass*. Boston, MA: Houghton Mifflin Harcourt.

Healy, G. et al. (Eds.). 2004. *The Future of Worker Representation*. Basingstoke, UK: Palgrave.

Heeks, R. 2017. A Structural Model and Manifesto for Data Justice for International Development. *Development Informatics Working Paper Series*, No. 69.

Hintz, A., Dencik, L. and Wahl-Jorgensen, K. 2018. *Digital Citizenship in a Datafied Society*. Cambridge: Polity.

HRBDT. 2020. The Human Rights, Big Data and Technology Project. Available at: https://www.hrbdt.ac.uk

Jansen, F. 2020. Consultation on the White Paper on AI – A European Approach. Submission. DATAJUSTICE project. Available at: https://datajusticeproject.net/wp-content/uploads/sites/30/2020/06/Submission-to-AI-WP-Fieke-Jansen.pdf

Johnson, J. A. 2018. *Toward Information Justice*. Cham, Switzerland: Springer.

Jørgensen, R. F. et al. 2019. Exploring the Role of HRIA in the Information and Communication Technologies Sector. In: N. Götzmann (Ed.), *Handbook on Human Rights Impact Assessment*. Cheltenham: Edward Elgar Publishing.

Kellogg, K. C., Valentine, M. A. and Christin, A. 2020. Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, 14(1): 366-410.

Mason, P. 2015. *Postcapitalism: A Guide to Our Future*. London: Allen Lane

McAlevey, J. 2016. *No Shortcuts: Organising for Power in the New Gilded Age*. Oxford: Oxford University Press.

McQuillan, D. 2020. AI After the Pandemic. *Medium*. Available at: https://medium.com/@danmcquillan/ai-after-the-pandemic-c603866a6ef8

Moore, P.V., Upchurch, M. and Whittaker, X. (eds) 2017. *Humans and Machines at Work: Monitoring, Surveillance and Automation in Contemporary Capitalism*. Basingstoke: Palgrave Macmillan.

Moulier Boutang, Y. 2011. *Cognitive Capitalism*. Cambridge: Polity Press.

Naughton, J. 2019. Are Big Tech's Efforts to Show it Cares About Data Ethics Another Diversion? *The Guardian*, 7 April. Available at: https://www.theguardian.com/commentisfree/2019/apr/07/big-tech-data-ethics-diversion-google-advisory-council

Niklas, J. and Dencik, L. 2020. European Artificial Intelligence Policy: Mapping the Institutional Landscape. Working Paper. DATAJUSTICE project. Available at: https://datajusticeproject.net/wp-content/uploads/sites/30/2020/07/WP_AI-Policy-in-Europe.pdf

Ruppert, E., Isin, E. and Bigo, D. 2017. Data Politics. *Big Data & Society*. July–December: 1–7. Available at: http://journals.sagepub.com/doi/abs/10.1177/2053951717717749

Sadowski, J. 2019. When Data is Capital: Datafication, Accumulation, and Extraction. *Big Data & Society*. January–June: 1–12.

Sadowski, J. 2020. The Internet of Landlords: Digital Platforms and New Mechanisms of Rentier Capitalism. *Antipode*, 52(2), 562–580.

Sánchez-Monedero, J. and Dencik, L. 2019. The Datafication of the Workplace. Working Paper. DATAJUSTICE project. Available at: https://datajusticeproject.net/wp-content/uploads/sites/30/2019/05/ReportThe-datafication-of-the-workplace.pdf

Scholz, T. (Ed.). 2013. *Digital Labor: The Internet as Playground and Factory*. New York and London: Routledge.

Scholz, T. 2017. *Uberworked and Underpaid: How Workers are Disrupting the Digital Economy*. Cambridge: Polity Press.

Srnicek, N. 2017. *Platform Capitalism*. Cambridge: Polity Press.

Srnicek, N. 2020. Data, Compute, Labour. *Ada Lovelace Institute*. Available at: https://www.adalovelaceinstitute.org/data-compute-labour

Srnicek, N. and Williams, A. 2015. I*nventing the Future: Postcapitalism and a World Without Work*. London: Verso Books.

Standing, G. 2017. *Basic Income: And How We Can Make It Happen*. London: Pelican Books.

Taylor, L. 2017. What is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally. *Big Data & Society, 4(2), 1–14.*

Taylor, L. and Dencik, L. 2020. Constructing Commercial Data Ethics. *Regulation & Technology*, 1-10.

Thompson, P. and Briken, K. 2017. Actually Existing Capitalism: Some Digital Delusions. In: K. Briken, S. Chillas, M. Krzywdzinski and A. Marks (Eds.), *The New Digital Workplace: How New Technologies Revolutionise Work*, pp. 13, 241–26. London: Macmillan Education.

Tieman, R. 2017. Barcelona: Smart City Revolution in Progress. *Financial Times*, 26 October.

Toh, A. 2020. Dutch Ruling a Victory for Rights of the Poor. *Human Rights Watch*. Available at: https://www.hrw.org/news/2020/02/06/dutch-ruling-victory-rights-poor

Van Doorn, N. and Badger, A. 2020. Where Data and Finance Meet: Dual Value Production in the Gig Economy. Working Paper. *Platform Labor*. Available at: https://platformlabor.net/output/dual-value-production-gig-economy

Wagner, B. 2018. Ethics as an Escape from Regulation: From 'Ethics-washing' to Ethics-shopping? In E. Bayamlioglu, I. Baraliuc, L. A. W. Janssens and M. Hildebrandt (Eds.), *Being Profiled, Cogitas Ergo Sum*, pp. 84–90. Amsterdam: Amsterdam University Press.

Wark, M. 2019. *Capital is Dead. Is this Something Worse?* New York: Verso.

Wilkin, P. 2000. Solidarity in a Global Age – Seattle and Beyond. *Journal of World-Systems Research*, 6(1), 19–64.

Zelevansky, N. 2019. The Big Business of Unconscious Bias. *New York Times*. Available at: https://www.nytimes.com/2019/11/20/style/diversity-consultants.html

Zuboff, S. 2015. Big Other: Surveillance Capitalism and the Prospect of an Information Civilization. *Journal of Information Technology*, 30(1): 75–89.

Zuboff, S. 2019. *The Age of Surveillance Capitalism*. London: Profile Books.

The Editor and Contributors

Editor

Pieter Verdegem is Senior Lecturer in Media Theory in the Westminster School of Media and Communication and a member of the Communication and Media Research Institute (CAMRI), University of Westminster, UK. His research investigates the political economy of digital media and the impact of digital technologies on society. He has published in journals such as *New Media & Society, Information, Communication & Society, European Journal of Communication*, *Telecommunications Policy*, *Government Information Quarterly*. He is a Senior Fellow of the Higher Education Academy.

Contributors

Alkim Almila Akdag Salah is an Assistant Professor at Utrecht University, Dept. of Information and Computing Sciences. Her research interests combine qualitative and quantitative methods to study mainly humanities and social data. After completing her PhD at the Art History Dept of UCLA, Almila joined University of Amsterdam, New Media Studies with an NWO VENI grant to analyse the online art platform deviantArt with computational means, combining computer vision, natural language analysis and social network analysis. She is currently the Visual Media and Interactivity WG leader at DARIAH

(Digital Research Infrastructure for the Arts and Humanities) and recently co-authored the article 'The Sound of Silence: Breathing Analysis for Finding Traces of Trauma and Depression In Oral History Archives', in the journal *Digital Scholarship in the Humanities*.

Willian Fernandes Araújo is PhD in Communication and Information (Federal University of Rio Grande do Sul, Brazil) and Professor of Digital Technology in the Communication Studies department at the University of Santa Cruz do Sul (UNISC), Brazil. He researches technological mediation in online platforms, mainly the debates concerning algorithms, governmentality and subject production.

Asvatha Babu is PhD candidate in the School of Communication at American University, Washington DC. Her research critiques the construction of emerging technologies as solutions to complex social problems. She is writing her dissertation on the use of facial recognition technology in policing in Chennai, India.

Benedetta Brevini is Associate Professor in Political Economy of Communication at the University of Sydney and Visiting Fellow of the Centre for Law, Justice and Journalism at City University, London. She is the author of *Public Service Broadcasting Online* (2013) and editor of *Beyond Wikileaks* (2013) and *Carbon Capitalism and Communication: Confronting Climate Crisis* (2017) and *Climate Change and the Media* (2018). *Amazon: Understanding a Global Communication Giant* (2020) and *Is AI Good for the Planet?* (2021) are her latest volumes.

Angela Daly is Reader in Law & Technology and Co-Director of the Strathclyde Centre for Internet Law & Policy in the University of Strathclyde (Scotland). She is a critical socio-legal scholar of the regulation of (digital) technologies. She is the author of *Socio-Legal Aspects of the 3D Printing Revolution* (Palgrave 2016) and *Private Power, Online Information Flows and EU Law: Mind the Gap* (Hart 2016). Along with S. Kate Devitt and Monique Mann, she is the co-editor of the open access collection *Good Data* (INC 2019). From October 2021 she will join the University of Dundee as Professor of Law & Technology.

Lina Dencik is Professor in Digital Communication and Society at Cardiff University's School of Journalism, Media and Culture and Co-Director of the Data Justice Lab. She has published widely on digital media, resistance and the politics of data and is currently Principal Investigator of the DATAJUSTICE project funded by an ERC Starting Grant. Her most recent publications include *Digital Citizenship in a Datafied Society* (with Arne Hintz and Karin Wahl-Jorgensen, 2018) and *The Media Manifesto* (with Natalie Fenton, Des Freedman and Justin Schlosberg, 2020).

S. Kate Devitt is Chief Scientist of Trusted Autonomous Systems (TAS) and adjunct Associate Professor Human-Computer Interaction, ITEE, University of Queensland (UQ), Australia. Kate is a transdisciplinary science leader, using expertise in epistemology, cognitive science and ethics to enable the development of human-autonomy teams that incorporate ethical, legal, and regulatory structures to achieve social license to operate and trusted adoption. Key publications include: *A Method for Ethical AI in Defence* (2021), 'The Ethics of Biosurveillance', in *Journal of Agricultural and Environmental Ethics* (2019) and the 'Trustworthiness of Autonomous Systems', in *Foundations of Trusted Autonomous Systems* (2017).

Rafael Grohmann is an Assistant Professor in Communication at the University of Vale do Rio dos Sinos (Unisinos University), Brazil. He is the Coordinator of DigiLabour Research Lab and Principal Investigator for the Fairwork project in Brazil and a member of Scholars Council, Center for Critical Internet Inquiry (C2i2), UCLA and a Founding Board Member of Labor Tech Research Network. His research interests include platform cooperativism and worker-owned platforms, work & AI, datafication, workers' organisation, platform labor, communication and work.

Wolfgang Hofkirchner is retired Professor of Technology Assessment at the TU Wien (Vienna University of Technology) and currently Director of the Institute for a Global Sustainable Information Society (GSIS), Vienna. His research explores self-organisation, information and the information society, that is, complexity thinking, science of information, and information and communication technologies (ICTs) in society. Recent publications include 'Promethean Shame Revisited. A Praxio-Onto-Epistemological Analysis of Cyber Futures' in Hofkirchner, W., Kreowski, H.-J. (eds.), *Transhumanism – The Proper Guide to a Posthuman Condition or a Dangerous Idea*? (2021).

Andreas Kaplan is Dean, ESCP Business School, Paris, France. His research focuses on analyzing the digital world, in particular the areas of artificial intelligence and social media. His work has been featured in various national and international press and media outlets such as the *California Management Review*, the *Financial Times,* the *Harvard Business Review France*, *La Tribune, La Repubblica*, *Süddeutsche Zeitung*, and *die Zeit*. He is the author of the book, *Higher Education at the Crossroads of Disruption: The University of the Twenty-First Century* (2021).

Monique Mann is a Senior Lecturer in Criminology and member of the Alfred Deakin Institute for Citizenship and Globalisation at Deakin University, Australia. Dr Mann is also Adjunct Researcher with the Law, Science, Technology and Society Research Centre at Vrije Universiteit Brussel. Mann's research expertise concerns three main interrelated lines of inquiry: new technology

for policing and surveillance, human rights and social justice and governance and regulation.

Dan McQuillan is a Lecturer in the Department of Computing at Goldsmiths, University of London. He has a PhD in Experimental Particle Physics and is currently writing a book on 'Anti-Fascist AI'.

Jenna Ng is currently Senior Lecturer in Film and Interactive Media at the University of York, UK. Jenna works primarily on digital visual culture, but also has research interests in the philosophy of technology, the posthuman, computational culture, and the digital humanities. She is the editor of U*nderstanding Machinima: Essays on Films in Virtual Worlds* (Bloomsbury, 2013) and author of *The Post-Screen Through Virtual Reality, Holograms and Light Projections: Where Screen Boundaries Lie* (forthcoming, Amsterdam University Press, 2021).

Carrie O'Connell is PhD candidate and NSF-IGERT Fellow in the Department of Communication at the University of Illinois at Chicago. Her current research is grounded in STS and focuses on communication with and memorialisation of the dead. Selected papers published include 'Breaking the "Black Box": Exploring the Philosophical Parallels Between Epic Theatre and Modern Digital Interface' (AoIR 2020).

Jernej A. Prodnik is researcher at the Social Communication Research Centre, Faculty of Social Sciences, University of Ljubljana (Slovenia) and Assistant Professor at the Department of Journalism. He is currently serving as the head of the department. His research is based in critical strands of media and communication studies, especially political economy of communication. Among his recent publications is '3C: Commodifying Communication in Capitalism' in C. Fuchs and V. Mosco (eds.), *Marx in the Age of Digital Capitalism* (2016).

Rainer Rehak is a researcher at the Weizenbaum Institute for the Networked Society, the German Internet Institute, which undertakes interdisciplinary research on the digital society in order to primarily understand but also shape the digital transformation for the common good. He has a background in computer science as well as in philosophy and his research focus comprises the areas of data protection, IT security, ascriptions and societal implications of artificial intelligence, (self critical) computer science and society combined with the philosophy of mind, language and science. See https://weizenbaum-institut.de/en/portrait/p/rainer-rehak.

Saif Shahin is an Assistant Professor in the School of Communication at American University, Washington DC, where he also serves as Faculty Fellow of the Internet Governance Lab. He studies data and technology as sociocultural

phenomena, power in online social networks, and the politics of digital identity construction. His research has been published in journals such as *New Media & Society*, *Information, Communication & Society*, *Social Science Computer Review*, *American Behavioral Scientist* and *Communication Methods & Measures.*

James Steinhoff is a Postdoctoral Fellow at the University of Toronto. His research focuses on the political economy of algorithmic media, media theory and philosophy of technology. He is author of *Automation and Autonomy: Labour, Capital and Machines in the Artificial Intelligence Industry* (2021) and co-author of *Inhuman Power: Artificial Intelligence and the Future of Capitalism* (2019). His current project is a study of non-profit AI institutes.

Chad Van de Wiele is PhD candidate and NSF-IGERT Fellow in the Department of Communication at the University of Illinois Chicago, with a graduate concentration in Black Studies. Their research examines the relationship between innovation, knowledge, power, and carcerality from an interdisciplinary perspective.

Index

B

C

S

T

AI FOR EVERYONE?

We are entering a new era of technological determinism and solutionism in which governments and business actors are seeking data-driven change, assuming that Artificial Intelligence is now inevitable and ubiquitous. But we have not even started asking the right questions, let alone developed an understanding of the consequences. Urgently needed is debate that asks and answers fundamental questions about power.

This book brings together critical interrogations of what constitutes AI, its impact and its inequalities in order to offer an analysis of what it means for AI to deliver benefits for everyone. The book is structured in three parts: Part 1, AI: Humans vs. Machines, presents critical perspectives on human-machine dualism. Part 2, Discourses and Myths about AI, excavates metaphors and policies to ask normative questions about what is 'desirable' AI and what conditions make this possible. Part 3, AI Power and Inequalities, discusses how the implementation of AI creates important challenges that urgently need to be addressed.

Bringing together scholars from diverse disciplinary backgrounds and regional contexts, this book offers a vital intervention on one of the most hyped concepts of our times.

ARTIFICIAL INTELLIGENCE | DIGITAL MEDIA STUDIES | INTERNET STUDIES

CDSMS CRITICAL DIGITAL AND SOCIAL MEDIA STUDIES

THE EDITOR

PIETER VERDEGEM is Senior Lecturer in Media Theory in the Westminster School of Media and Communication and a member of the Communication and Media Research Institute (CAMRI), University of Westminster, UK. His research investigates the political economy of digital media and the impact of digital technologies on society.

UNIVERSITY OF WESTMINSTER PRESS

uwestminsterpress.co.uk

ISBN 978-1-914386-16-9

9 781914 386169 90000