unesco

Mila

MISSING LINKS IN AI GOVERNANCE

EDITED BY
Benjamin Prud'homme
Catherine Régis
and Golnoosh Farnadi

MISSING LINKS IN AI GOVERNANCE

Published in 2023 by the United Nations Educational, Scientific and Cultural Organization (UNESCO), 7, place de Fontenoy, 75007 Paris, France; and by Mila – Québec Artificial Intelligence Institute, 6666 Rue Saint-Urbain, QC H2S 3H1, Montréal, Canada.

ISBN: 978-92-3-100579-4

Editors: Benjamin Prud'homme, Catherine Régis, Golnoosh Farnadi, Vanessa Dreier, Sasha Rubel (2021), Charline d'Oultremont.

Coordination: Amanda Leal de Lima Alves

Graphic design: Alphatek

Professional Translation: Martine Sénecal, Daly Dallaire Services

Cover design: Frédérick Gélinas

SHORT SUMMARY

18 selected submissions offering a pluralistic, informed and critical approach to AI Governance

NAVIGATING THE UNKNOWN: INSIGHTS INTO AI GOVERNANCE

Over the next decade, Artificial Intelligence ("AI") will continue to significantly impact societies. While these scientific and technological advances take place at an extraordinary pace, it is necessary that we simultaneously stimulate a global and inclusive conversation around their development and governance.

It is in this context that Mila and UNESCO join forces to steer a collective work to identify and understand missing links in AI governance. This publication is a compilation of 18 selected submissions from a global open call for proposals launched in 2021. The works featured cross disciplinary and geographical boundaries, and include the perspectives of academics, civil society representatives, and innovators to help shift the conversation on AI from what we do know and foresee to what we do not, the missing links. The topics covered are wide ranging, including AI and Indigenous rights, Deepfakes, Third-Party Audits of AI Systems, AI alignment with SDGs, and the centralization of decision-making power AI allows.

Policymakers and civil society members will benefit from the insightful perspectives brought forward to face the immense task they are presented with – which is to ensure the development of AI in a human-centred, responsible and ethical way, in accordance with human rights.

unesco

"Since wars begin in the minds of men and women it is in the minds of men and women that the defences of peace must be constructed"

ABOUT THE ORGANIZATIONS

As Artificial Intelligence (AI) applications continue to expand opportunities for the achievement of the Sustainable Development Goals, UNESCO is working to harness these opportunities in its fields of competence across education, the sciences, culture, communication and information. UNESCO is leading reflections around pressing concerns related to the rapid development of AI, from a Human Rights and ethics perspective.

These range from AI's role in the future of education, the omnipresent challenges of disinformation and hate speech online, harnessing AI for the sustainable development goals and to empower the global south, and to promoting gender equality in the AI Sector and combatting algorithmic bias.

Mila's mission is to be a global pole for scientific advances that inspires innovation and the progress of AI for the benefit of all. The Montreal-based institute rallies over 1,000 university researchers working on AI research across a wide range of subfields and application areas, led by our commitment to studying and developing frameworks that support the advancement and deployment of responsible AI.

As a global leader in the field, Mila contributes to national and international efforts to foster social dialogue and engagement on questions of ethics, bias and social justice related to AI's transformative social role, with a view of supporting the development and operationalization of responsible AI technologies and governance.

This publication was made possible thanks to the financial support of

Québec

ACKNOWLEDGEMENTS

The editorial team would like to thank the following people from UNESCO for shaping and refining this publication: Marielza Oliveira, Guy Berger, Prateek Sibal, Cedric Wachholz and Jacinth Chia.

The editorial team would like to thank the Quebec government – and in particular the *Ministère des relations internationales et de la Francophonie* – as well as the *Fonds de recherche du Québec* for their financial support to this publication.

The editorial team is also grateful to those who shared their time, knowledge and work in response to the open call for contributions to explore creative, novel and far-reaching approaches to artificial intelligence. Thanks go to all authors who took the time to share their insights, and whose voices are essential to make this publication possible.

The editorial team would also like to thank those who contributed to the formatting, translation, copyediting, and design of the publication: Martine Sénecal, Daly Dallaire Services, Marie Zumstein, Noah Oder, Laura Gagliano, Frédérick Gélinas, and Alphatek.

Finally, the editorial team would like to extend a warm thank you to Amanda Leal de Lima Alves, whose tireless coordination work has made this publication possible.

FOREWORD FROM UNESCO

UNESCO is the UN specialized agency mandated to build peace through international cooperation in education, the sciences, culture, information and communication. Within its mandate, the Organization works to deepen knowledge, foster global collaboration and offer policy advice on key issues related to digital innovation and transformation. Under this area of expertise, this publication is the latest offering in a line of cutting-edge knowledge resources as UNESCO continues to exercise its many functions, notably as a laboratory of ideas.

As the development and use of Artificial Intelligence (AI) expands, it continues to defy what we once thought was possible. Just 80 years ago, computer scientists were focused on enabling computers to do simple tasks, such as storing commands. Today, AI applications have extended to natural language processing, judicial processes, driverless vehicles and disease mapping, just to name a few. AI offers new opportunities to achieve the UN Sustainable Development Goals, including within UNESCO's fields of competence in education, natural and social and human sciences, culture, and communication and information. However, while it has significant potential for fostering sustainable development, the complexity and pace of AI development poses a challenge to not only the governance of AI, but also the protection and promotion of human rights.

With its leading role in international cooperation, UNESCO has guided global reflection around pressing concerns related to AI. Part of this work led to the adoption in 2021 by UNESCO's 193 Member States of the Recommendation on the Ethics of AI, the first global standard-setting tool in this area.

With the common understanding that AI is to be first and foremost human-centered, UNESCO also recognizes the challenges that many countries face in the development and governance of AI. As such, this publication, stemming from UNESCO's collaboration with Mila – Québec Artificial Intelligence Institute, is an important contribution to reflections on the governance challenges and the human and institutional capacity gaps that countries face to ensure the trustworthy and responsible use of AI.

As an example, UNESCO's Judges' Initiative has developed a comprehensive set of tools to train members of the judiciary on AI's application and impact in the administration of justice. Moreover, we are engaging with civil servants, who are important stakeholders in AI policy development, with the goal to develop an understanding of opportunities and challenges related to AI in their work. UNESCO has also produced research on issues related to gender and artificial intelligence, such as on the digital skills gap and gender bias in AI algorithms, or on the impact of AI on women in the world of work.

I am convinced that this publication will help provide policymakers and civil society members with the critical perspectives needed to ensure that the development of AI reaches its full potential in accordance with fundamental rights and freedoms. I hope that readers will find that the insights encompassed in the various chapters will surface both the answers that we seek, and the questions that we need to ask, to ensure that AI technologies leave no one behind.

It is our hope that this publication thereby will help reinforce the essential contribution that digital technologies and specifically AI can make to foster inclusive and peaceful societies, when applied with a human rights-based approach, and to establish trustworthy and responsible AI.

UNESCO thanks Mila and all the contributors for making this publication a reality. We wish you, the reader, an inspiring read, and look forward to engaging with you.

TAWFIK JELASSI
Assistant Director-General for Communication and Information, UNESCO

FOREWORD FROM MILA – QUÉBEC ARTIFICIAL INTELLIGENCE INSTITUTE

Since its inception, Mila has strived to achieve the highest levels of scientific leadership in artificial intelligence (AI) while holding the development of responsible and ethical AI at the very core of its mission. Our collaboration with UNESCO speaks to our commitment to democratize AI knowledge and global cooperation that serves the benefit of all.

Technological innovation is already impacting every sphere of life. AI-driven advances in areas such as healthcare, agriculture and climate science offer game-changing opportunities that were until recently unimaginable. However, AI also poses important risks, and tireless energy must be directed at developing responsible and beneficial AI systems. This means AI systems that uphold human rights, the rule of law, and the principles put forward in UNESCO's *Recommendation on the Ethics of AI.* It also means AI systems that support the implementation of the Sustainable Development Goals (SDGs). This is an ambitious agenda, but it's one we must collectively embrace. And while technical breakthroughs are essential, AI governance will play a pivotal role in determining how – and for whom – we harness the power of AI.

This commitment to contributing to the development of responsible AI is deeply embedded in the Mila community. Already in 2018, Mila co-led the *Montreal Declaration for a Responsible Development of Artificial Intelligence,* which aimed at steering the ethical development of AI by formulating key principles with strong democratic legitimacy. In 2020, as part of the process leading to the adoption of UNESCO's Recommendation on the Ethics of AI, Mila co-led the *Inclusive Dialogue on the Ethics of AI*. Multiple members of the Mila community also conduct research at the intersection of AI and sustainability, health, fairness, ethics, and governance. Finally, Mila leads applied projects in a vast array of domains to harness the power of AI to support achieving the SDGs. This includes projects to mitigate and adapt to the climate crisis, to support the prevention of human trafficking and modern slavery, to inform the development of inclusive AI policies, and to identify gender biases in written texts.

This new collective publication is another example of Mila's dedication in this regard. It offers a pluralistic, informed and critical approach to AI Governance. The perspectives of many actors across disciplinary, geographical and professional backgrounds converge to amplify the scope and relevance of these reflections. For example, chapters explore the implications of AI development for Indigenous communities and LGBTI people, the pressing need for gender equality in AI ecosystems, the reduction of inequalities through access to AI education and knowledge, as well as ways to ensure AI is put at the service of the *2030 Agenda for Sustainable Development.*

Finally, we are proud and thankful to present this work alongside UNESCO, which has played a leading role in AI Governance with the adoption of the first global standard-setting instrument on the Ethics of AI. We also thank all contributors for their commitment towards this publication, and hope that leaders, policymakers and civil society across the globe can now engage with it.

VALÉRIE PISANO
President and CEO, Mila – Quebec Artificial Intelligence Institute

TABLE OF CONTENT

INTRODUCTION

AI is now part of our daily lives. It is used in a wide array of fields such as health, transport, manufacturing, and cybersecurity, thereby impacting the way we communicate, work, and learn. Already, AI offers opportunities as well as poses risks that were unforeseen only decades ago, and its governance has become a priority for all actors of society, mobilizing academia, governments, civil society and international organizations alike. As AI development continues to accelerate, its impacts on societies will be even more profound in the years to come. In this context, global and inclusive conversations are essential to help us shed light on these challenges, and to ideate novel ways to comprehend and tackle them.

This is why Mila and UNESCO joined forces to foster a multistakeholder exchange on salient issues that must be addressed to support the responsible development of AI. To ensure the inclusion of diverse perspectives, we published a global call for contributions. This publication is a compilation of 18 selected chapters, which cross disciplinary, cultural, and geographical boundaries. They put forward the views of a wide range of actors of the AI ecosystem including academics, civil society representatives, innovators, and policy makers. The aim was to expand the conversation on AI and shift our focus from what we do know to what we do not – the missing links –, as well as to propose actionable changes towards more equitable and inclusive AI (eco)systems. In other words, the intention of this publication is to create a space for opposing, novel and nuanced views on AI governance as it is seen, lived and understood by the many actors that contribute to its development and deployment.

This book contains thoughts and propositions that cover a wide range of important topics. This includes chapters on the risks and opportunities of AI for Indigenous rights (Toki and Phillips), the potential impacts of these technological developments on LGBTI communities (Horner), the ways in which AI governance can ensure the consideration of diverse voices, including from the Global South (Png; Mbayo), and the tensions that can arise between ethical principles when AI is mobilized

in times of pandemics (Voarino and Régis). Other chapters offer thought-provoking accounts of AI's impacts on legislative discussions and policy-making (Clement-Jones; i4Policy), on peace and democratic stability through Deepfake manipulations (Muñoz and Marijan), on the opportunities of using AI to support sustainable development (Letouzé et al.; Shawe-Taylor et al.), and on inclusive innovation (Future Earth; Dieng). The contributions also invite us to explore the challenges and opportunities that fairness and ethics pose in the AI industry (Farnadi et al.), to harness the opportunities AI offers for education (Mishra et al.) and to integrate a feminist ethics of care in the conversation on AI ethics (Noiseau). Finally, some of the world-leading voices in AI offer their reflections on the ways in which AI ecosystems could better support innovation for socially beneficial purposes (Bengio et al.), the important and sometimes devastating impacts of bias in data (Crawford), the centralization of decision-making power AI enables (Ng and Brynjolfsson), and the need to rethink and reform third-party audits of AI systems (Algorithmic Justice League).

With this publication, we want to provide policymakers, innovators, academics, and civil society with fruitful perspectives to help us face the immense task we are presented with: shaping the development of AI so that no one is left behind. This means working towards AI systems that are human-centered, inclusive, ethical, sustainable, as well as upholding human rights and the rule of law. This publication is our humble contribution to this global effort. It is in no way exhaustive, and many other initiatives and dialogues will need to take place for the world to harness the opportunities AI offers while responding to the risks it poses. We hope the perspectives included herein will stimulate discussions on some of the most pressing challenges in AI governance, and provide novel ideas for our readers to consider as they advocate for the responsible development of AI.

CHANGE FROM THE OUTSIDE: TOWARDS CREDIBLE THIRD-PARTY AUDITS OF AI SYSTEMS

INIOLUWA DEBORAH RAJI

Doctoral student in Computer Science at UC Berkeley and Research Fellow at the Algorithmic Justice League.

SASHA COSTANZA-CHOCK

Director of Research & Design at the Algorithmic Justice League.

DR. JOY BUOLAMWINI

Founder and Executive Director at the Algorithmic Justice League.

for the Algorithmic Justice League

SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

CHANGE FROM THE OUTSIDE: TOWARDS CREDIBLE THIRD-PARTY AUDITS OF AI SYSTEMS

ABSTRACT

When artificial intelligence (AI) systems cause harm, it is important to identify the responsible stakeholders and hold them accountable. Recently, AI audits have become an increasingly popular proposed accountability mechanism, and a growing ecosystem of AI auditors has emerged. By AI audit, we mean a process through which an auditor evaluates an AI system or product according to a specific set of criteria and provides findings and recommendations. AI audits can help identify whether AI systems meet stated performance targets, or in terms of other concerns such as bias and harm, data protection and privacy, transparency and accountability, adherence to standards and regulatory requirements, or labor practices and ecological impacts. AI audits may be conducted by first-party (internal), second-party (contracted), or third-party (external and fully independent) auditors. Third-party auditors, such as independent researchers, investigative journalists, community advocates, law firms, and regulators, have conducted many of the most impactful audits of AI systems to date. However, despite the importance of third-party auditors to AI accountability, this group has been mostly overlooked in AI policy.

In this chapter, we propose seven key policy interventions to strengthen the ability of third-party auditors to scrutinize AI systems: legal protections for third-party AI auditor access; accreditation for AI auditors; standards development for AI products; AI harm incident reporting; mandatory public disclosure of AI systems use; a frame shift beyond AI bias to harms; and accountability mechanisms to ensure appropriate audit responses.

By identifying these missing links, we hope to help advance a regulatory landscape that enables, protects and supports the ability of “outsiders” such as third-party auditors and other external stakeholders to scrutinize AI systems. We believe that credible third-party audits will help protect the human rights of communities that are most likely to be harmed by the use of AI systems.

INTRODUCTION

Artificial intelligence (AI) systems are too often developed and used in ways that reproduce existing forms of systemic inequality and cause genuine harm, particularly for marginalized groups. The harms perpetuated by AI systems have become increasingly evident and are now well-documented in research literature and in popular culture, as well as in emerging policy conversations and proposals. However, despite the growing public awareness of potential and actual harm, these issues remain difficult to identify, assess and ultimately address.

We are writing this chapter as researchers operating on behalf of the Algorithmic Justice League (AJL). AJL is an organization whose mission is to raise awareness about the impacts of AI, equip advocates with empirical research, build the voice and choice of the most impacted communities, and galvanize researchers, policymakers and industry practitioners to mitigate AI harms and biases.

In this chapter, we focus on one accountability mechanism that we believe has been underspecified and underutilized to date: AI audits. Of the many possible mechanisms for AI accountability, audits remain exemplary in their ability to lead to broader public awareness, impactful product recalls, regulatory action and successful litigation. However, many of the external stakeholders who play the role of third-party auditors and who are invested in protecting communities against the threat of AI harms are themselves vulnerable to retaliation from powerful technology companies. This diverse group of auditors, who may be university or private-sector researchers, non-governmental organizations (NGOs), law firms, regulators or other public sector bodies, are often left to their own devices as they try to figure out what to do. They are left to design and execute audits without much guidance, support or protection from policymakers.

In the pages that follow, we therefore provide a map of some of the policy interventions necessary to enable and empower third-party auditing for AI systems. Currently, the role of third-party auditors in the broader AI accountability ecosystem garners limited public policy consideration. Policymakers have an important role to play in ensuring the continued protection and support of those that choose to play this essential role. In this chapter, we begin by defining background terms, and then outline seven critical policy interventions necessary to allow for effective third-party auditing. These include:

1) legal protections for third-party AI auditors;
2) accreditation for AI auditors;
3) standards development for AI products;
4) AI harm incident tracking;
5) mandatory public disclosure of AI systems use;
6) a frame shift beyond AI bias to harms; and
7) accountability mechanisms to ensure that audit outcomes produce change.

Our hope is that this work provides a starting point for a much-needed discussion about policy interventions to support third-party auditors within the larger context of AI accountability policy.

BACKGROUND

To begin, we briefly summarize several key concepts: we specify what we mean by AI audits, describe the emerging ecosystem of AI auditors, clarify the distinction between first-, second-, and third-party audits, and summarize a few of the unfolding policy initiatives that shape the current landscape. We highlight some of the productive directions that policymakers have explored so far, and then emphasize missing links in the AI accountability ecosystem.

In public discourse, as well as in policy circles, people tend to use the terms Artificial Intelligence (AI), automated decision systems (ADS), algorithmic systems, and machine learning (ML) somewhat interchangeably, in ways that can be frustratingly vague. In this chapter, we use "AI systems" as a broad umbrella term, with occasional reference to other terms where more specificity is required (Richardson, 2021). We use the term "AI systems" to refer to a range of sociotechnical systems that fully or partially automate processes involving information processing and pattern recognition. Most deployed AI systems are developed using ML techniques, and are thus heavily influenced by the training data that informs the system's output. Most AI systems are developed to mimic or automate some cognitive process, although practically, most are deployed to execute a specific task such as classification, ranking or identification. We acknowledge that the use of the term "AI system" is imprecise, but we have opted to match the language most commonly deployed in current policy discussions in order to anchor our recommendations to this context.

What do we mean by "AI audits"?

Outside of the context of the AI industry, auditing has gradually gained acceptance as a mainstream accountability mechanism in many domains. Unlike other forms of risk assurance, such as impact assessments or checklists, auditing tends to imply precision, where a system is evaluated against a known standard. As post-hoc system evaluations, audits can provide precise, explicit statements about limitations and risks. By determining whether or not an organization or product complies with requirements, audits can help determine whether a vendor is selected, and whether a product is ready for deployment or needs to be recalled. Audits are an essential instrument that affected populations can potentially use to critique and influence power holders who make decisions about AI systems that impact their circumstances (Wieringa, 2020).

By "AI audit," in this chapter, we mean a process through which an auditor evaluates an AI system or product according to a specific set of criteria and provides findings and recommendations to the auditee, to the public, or to another actor, such as to a regulatory agency or as evidence in a legal proceeding. AI audits can help identify whether AI systems meet or fall short of expectations, whether in terms of stated performance targets (such as prediction or classification accuracy) or in terms of other concerns such as bias and discrimination (disparate performance between various groups of people); data protection, privacy, safety and consent; transparency, explainability and accountability; adherence to standards, ethical principles and legal and regulatory requirements; or labor practices, energy use and ecological impacts.

Ideally, audits of AI systems, like audits in other domains, should be conducted by entities that are formally accredited by a recognized accreditation body (although we note that accreditation processes must be organized carefully in order to avoid capture by industry or the exclusion of independent researchers). Audits should also be conducted in reference to well-articulated expectations, typically in the form of clearly defined and widely recognized standards. Additionally, auditors should be empowered to disclose their findings (while taking care to protect personally identifiable information) and protected in various ways from hostile corporate reactions to audit results. However, as we discuss at length later in the chapter, the typical protection afforded to third-party auditors in other domains are not yet provided to AI auditors.

The emerging ecosystem of first-, second- and third-party AI auditors

As understanding of AI bias and harms becomes more widespread, AI audits have become more popular. AI auditors can be classified into three broad categories: first-party, second-party and third-party. First-party auditors are employees of the AI system developer, providing internal oversight of compliance with performance expectations defined by the organizational leadership. Many of the top technology companies have set up or are now in the process of setting up internal AI ethics teams, which effectively operate as first-party auditors. Examples include Facebook's Society and AI Lab (SAIL), Microsoft's Fairness, Accountability, Transparency, and Ethics (FATE) team, Twitter's ML Ethics, Transparency and Accountability (META), PayPal's Justice by Design group, Google's Ethical AI and Responsible Innovation teams, and many similar groups.

Second-party auditors are contracted consultants or research collaborators, typically hired by the audited company to outsource the audit task to more skilled practitioners or to provide a fresh perspective. Second-party auditors are visible in the growing industry of those providing AI audits as a service. This includes smaller startups and consulting companies (such as ORCAA and Parity) as well as teams within larger consultancies (such as Deloitte, McKinsey and Accenture) that offer ethical, legal, or technical reviews of other firms' AI products. Some AI auditing teams that begin as first-party auditors within the largest technology companies have also gone on to offer second-party AI auditing services to other firms, such as AI audit teams at Google (Simonite, 2020) and IBM (IBM, 2021). First- and second-party auditors both have a contractual relationship with the audit target.

Third-party auditors, on the other hand, have no contractual relationship with the audit target. As outsiders, they often have limited access to the audited system, and this constrains some auditing techniques. However, they maintain complete independence, and this can enable freedom to ask more difficult questions about system outcomes and to disclose negative audit findings.

Third-party auditors are completely separate from the audit target. They are completely external stakeholders scrutinizing the audited systems and institutions. Examples of third-party AI auditors include independent researchers, teams of investigative journalists (such as The Markup or the Associated Press's Tracked project), civil society organizations, law firms and, in some cases, regulators. They conduct independent external investigations of harms from AI systems, with no contractual obligation to the developer, vendor or operator of the AI system in question. Although the relationship between third-party auditors and audit targets tends to be adversarial, this is not a necessary condition.

We also note that a number of university-based centers and research groups have emerged to focus on the social, technical and legal aspects of algorithmic harms. In the United States, this includes institutions such as the AI Now Institute at New York University (NYU), the Center for Critical Internet Inquiry at the University of California, Los Angeles (UCLA) and others. In the UK, this includes the Institute for Ethics in AI at Oxford, the Ada Lovelace Institute and many more. Some of these research centers conduct formal audits of AI systems, and some do not. Some act as second-party auditors, by working with companies under a signed contract, while others act as third-party auditors.

The impact of third-party auditors

Auditor independence, credibility and integrity have long been concerns in other fields, including financial auditing (AICPA, 2017), environmental auditing (Gunningham, 1993), and food and safety audits (Lytton and McAllister, 2014). Both first- and second- party auditors are beholden to terms set by the audit targets. Only third-party auditors are free to act independently of requests by the audit target. As a result, third-party auditors, free from contractual obligations or conflicts of interest, can ask unique and challenging questions. They can operate against the preferences of the audit target, when necessary, in order to hold targets accountable.

Third-party AI auditors have conducted many of the most impactful AI audits to date. For example, investigative journalists at ProPublica demonstrated racial bias in recidivism risk assessment scores, prompting the involved vendor to respond, and spurring an ongoing reconsideration of the use of risk assessment tools within the broader judicial system (Angwin et al., 2016). ProPublica reporting also exposed how Facebook allowed discriminatory ad targeting by employers, landlords and lenders (Gillum and Tobin, 2019), a story which led to a $5 million settlement and changes to Facebook's ad systems (Spinks, 2019), although subsequent investigations by The Markup found that the problem continues (Keegan, 2021). The Markup also showed how NYC school admission algorithms reproduce racial segregation (Lecher and Varner, 2021) and how mortgage lenders deny people of color at twice the rate of white applicants (Martinez and Carollo, 2021). Independent researchers Dr. Joy Buolamwini, Inioluwa Deborah Raji and Timnit Gebru demonstrated gender and skin type performance accuracy disparities in facial analysis technology sold by the world's largest vendors (Buolamwini and Gebru, 2018), a result that led to IBM, Amazon and Microsoft declaring indefinite moratoriums on their sale of facial recognition to police (Raji and Buolamwini, 2019). This further informed an American Civil Liberties Union (ACLU) complaint against Detroit police after the false arrest of a Black man, Robert Williams, due to a false facial recognition match (Hill, 2020). Around the same time, the National Institute of Standards and Technology (NIST) audited 189 software algorithms from 99 developers to find that most systems performed drastically worse for people of color (Grother et al., 2019). In a similar fashion, the legal firm Foxglove was able to push the UK government to reverse its position on the use of algorithms to assign final A-level grades, informed by an analysis on the disparate impact of that algorithm on low-income students (Foxglove, 2020).

From these examples, it is clear that third-party audits are consequential accountability interventions. Third-party auditors may hold a variety of motives: they may be investigative journalists, independent academic researchers, civil society organizations, lawyers, or regulators. These and many other third-party audits of AI systems have focused sustained and growing attention on bias, harms, equity and accountability for AI systems. However, we believe that the current AI policy landscape has mostly ignored or under-specified the importance of third-party auditors to the equitable and accountable development of AI systems.

The unfolding policy landscape for AI audits

The policy landscape that governs and shapes the development of AI systems is rapidly evolving. While the Algorithmic Justice League does not lobby for particular bills, we closely monitor legislative developments, especially in the USA and in the EU. We are encouraged by the introduction of various regulatory proposals to rein in the unchecked power of AI systems, although we believe that not enough attention has been given to third-party audits.

Recently, regulatory bodies such as the UK Information Commissioner's Office (ICO), the country's primary data protection regulator, have entered the fray with guidelines on how companies and government agencies could audit their systems. In response to the General Data Protection Regulation (GDPR), ICO has provided documentation guidelines to help companies understand how to account for and communicate algorithmic and data management details (Kazim and Koshiyama, 2020). The ICO guidelines are a positive development, but in our view, they over-emphasize the data protection frame that dominates the EU policy conversation, since they frame the control of personal data as the primary control lever over AI systems, although this is not always the case. Assessing responsible data handling requires direct access to the audit target. As a result, UK and EU policy interventions tend to prioritize providing guidance for how internal or first-party auditors at private companies might influence engineering and product teams' decision-making around data collection, storage, and use. While internal first-party or second-party AI audits may be a useful tool, they cannot replace external third-party scrutiny. Although such regulatory interventions might lead to more responsible engineering practices, they do little to help outside observers bring forward their unique concerns. In related legislative efforts in the USA,

the 2019 *Algorithmic Accountability Act* emphasizes the need for internally developed Algorithmic Impact Assessments (AIAs), analogous in some ways to the Data Protection Impact Assessments (DPIA) mentioned in GDPR, but potentially with a broader remit since the focus is not limited to data protection practices (United States Congress, 2019). Again, this is a positive development, but unfortunately, the 2019 Algorithmic Accountability Act focuses on the internal development of AIAs rather than on any explicit requirement for third-party verification. Also, although the Algorithmic Accountability Act does propose that firms report system details to regulators, the 2019 version of the Act does not require public disclosure of the AIA (MacCarthy, 2019). Furthermore, AIAs currently serve more as internal tools for open-ended reflection rather than as strict compliance evaluations. While several AIA proposals espouse the need for broader community participation (Metcalf et al., 2021), in practice, AIA implementations are not typically inclusive of a broad range of views and effectively amount to an internal first-party audit (Selbst, 2021).

Other recent policy developments in the USA, such as the Algorithmic Justice and Online Platform Transparency Act of 2021 (Markey, 2021) and the Automated Decision Systems Accountability Act of 2021 (Cuevas, 2020), demonstrate an increased awareness of the role of federal agencies as third-party auditors, in particular the Federal Trade Commission (FTC). This is a positive development towards outside oversight of AI systems, yet the execution details remain ambiguous. In the EU, recent AI accountability policies focused on social media companies have also introduced the language of auditing. In particular, the Online Harms Bill in the UK and the Digital Services Act (DSA) from the European Commission explicitly mention the need for independent scrutiny of deployed AI systems (United Kingdom Government, 2020; European Commission, 2022). However, these interventions remain limited. For example, the "independent auditors" described in Article 28 of the DSA are really second-party auditors: paid consultants hired by the audit target to execute the mandated evaluation. Article 31 of the DSA comes closest to describing third-party auditor participation, but specifically narrows that definition to academic researchers and regulators, excluding by omission other potential third-party auditors such as investigative journalists, law firms or civil society organizations. The DSA does mention the need for auditors to be vetted but provides neither a clear accreditation mechanism for auditors nor clear standards against which AI system audits must be conducted.

So, despite encouraging recent policy developments in AI accountability, we are still a far cry from an AI policy ecosystem that enables the effective participation of third-party auditors. We do not yet have the standards and regulatory framework that we need to ensure that third-party auditors are accredited, protected and supported to play their part. To ensure equity and accountability in the deployment of AI systems, the communities that are most likely to be harmed by these systems must be better represented in the audit, assessment or evaluation process. Third-party auditors, who can play that role, need to be accredited and supported within a policy ecosystem that ensures their independence, integrity, and effectiveness. In the rest of this chapter, we articulate seven missing links in AI Policy that we believe are required to help third-party AI auditors do their work.

MISSING LINKS

Multiple policy interventions are necessary to ensure that third-party AI auditors can play their role. Despite recent developments, we have yet to see legislative action, even in proposal form, that satisfies our minimum criteria for effective oversight. Here we propose seven key interventions that we believe will strengthen the current policy discourse:

1. **Legal protection for third-party auditor access.** Once we have developed a robust ecosystem of accredited auditors, they need to be able to do their jobs. We need policy mechanisms that provide protected access for third-party auditors to the information that they need in order to conduct independent assessments of AI systems, through the lens of different priorities and concerns that they may bring to the table.

2. **Accreditation and training for auditors.** A formal accreditation process for first-, second- and third-party auditors is a prerequisite for providing auditor access, guaranteeing auditor integrity, and ensuring audit quality. AI auditors should be reviewed by an accreditation body that ensures they adhere to inclusive national or international expectations for conduct and competence. That said, we caution that an accreditation process must not be controlled by industry and must be carefully organized to avoid excluding independent researchers.

3. **Standards.** We need to see the development of clear and widely recognized standards for AI products that embody high-level expectations for AI systems and their use. Clearly defined standards, developed through a transparent process, are a prerequisite for meaningful AI audits.

4. **Harms incident tracking.** No AI system is perfect. We need AI harm incident reporting and tracking in order to ensure that those who are harmed by AI systems are able to share their experiences and concerns. Standardization of AI harm incident tracking supports a grounded understanding of problems, system improvements by vendors and operators, better oversight by regulators, legal action where necessary, and greater visibility of incidents in the press and in the public eye.

5. **Notice of use.** AI accountability policy should include mandatory public disclosure of AI systems use for any system with the potential to cause harm. Public agencies, in particular, must be required to notify the public when they procure, pilot and deploy AI systems. Public disclosure of use makes it possible for third-party auditors to identify audit targets and is a basic requirement upon which to build meaningful consent from those who will use and be impacted by the AI system.

6. **Frame shift beyond bias to harms.** AI policy should address a broad range of AI harms rather than focusing only on technical measures of accuracy and bias. This also requires meaningful definitions of all key terms and acknowledgement of multiple forms of harm.

7. **Post-audit accountability mechanisms.** Ultimately, third-party audits are only useful if there are multiple mechanisms to ensure that the issues they uncover are addressed. AI policy should include various enforcement tools to ensure that, in response to audit outcomes, firms disclose key audit findings, make improvements accordingly, seek compliance with standards and with the law, and redress harms.

In the next sections, we briefly expand upon each of our seven recommendations.

1. Access and protection for third-party auditors

Once we have developed a robust ecosystem of accredited auditors, they need to be able to do their jobs. We need policy mechanisms that provide protected access for third-party auditors to the data they need in order to conduct independent assessments of AI systems, through the lens of different priorities and concerns that they may bring to the table.

Third-party audits across the AI systems lifecycle are necessary accountability measures. Third-party auditors can shine a light on problems that are unforeseen, deprioritized, or ignored by those who develop, purchase, deploy, or maintain AI systems. Third-party audits may also be used to focus

attention on disparate impacts against various marginalized stakeholders who are too often excluded from consideration. As they have no contractual relationship with the audit target, third-party auditors are less likely to be influenced by the preferences, expectations or priorities of the audit target**.** Also, third-party auditors tend to represent a wider range of perspectives than internal stakeholders, and therefore cast a novel critical eye on key issues that may otherwise go unidentified or under-prioritized.

A third-party audit is an audit that is "performed by an audit organization independent of the customer-supplier relationship and is free of any conflict of interest" (ASQ, n.d.). Third-party audits play a unique role in algorithmic accountability.[1] First- and second-party audits, while potentially useful tools, are insufficient to ensure the development of equitable and accountable AI systems. These kinds of audits tend to be limited in certain ways. For example, first- or second-party audits are often reactive, occurring only after an issue is raised by regulators or by the public. First-party auditors rarely disclose their findings other than to other teams within their company, and second-party auditors are usually restricted from public disclosure of audit findings by contractual non-disclosure agreements (NDAs). Publication of first- or second-party AI audit findings is especially unlikely when audits reveal significant problems of AI bias or harm. Although these kinds of audits provide an opportunity for unlimited access to AI systems, and are therefore a useful tool for development or pre-deployment evaluation, they are not independent of being influenced by the target. This highlights the necessity of engaging third-party perspectives (Raji et al., 2020).

AI policy should help ensure that accredited third-party auditors are able to gather the data that they need to evaluate AI systems without fear of being blocked from system access, and especially without fear of legal retaliation from the audit target. For example, despite a recent legal decision that provides an exception for researchers conducting algorithmic discrimination studies (Weiger, 2020),[2] third-party auditors in the USA have shared in interviews with AJL that they fear prosecution under the Computer Fraud and Abuse Act (CFAA) when they risk breaching a product's terms of service in the course of their audit work (United States, 1984).

Therefore, we recommend the following policy interventions:

- **Provide legal protections for third-party auditors.** Currently, computer fraud, abuse and cybersecurity laws (such as CFAA in the USA) may make third-party or adversarial auditors vulnerable to lawsuits against data scraping that is important to their process. All such laws should provide exceptions for journalists and academic researchers.
- **Require third-party audits of AI systems** that are developed, purchased or deployed by any government agency or recipient of federal funds.
- **Provide data access to vetted third-party auditors.** Third-party auditors are not paid or contractually tied to the audited organization, and thus they also tend to only have consumer-level access to the audit target. This often manifests in a lack of access to data and code, as well as little access to documentation or discussion with system developers about the rationale for their decision-making. We have seen many policy interventions that include guidelines to internal auditing processes, which are potentially useful, but we also need to see requirements for vetted third-party auditors to receive the access they need to successfully scrutinize products both pre- and post- deployment.

1. Third-party audits may sometimes be referred to as "independent" or "external" audits. For us, the key is that third-party audits are performed by people or organizations that are completely independent of the audit target. It is important to distinguish between third-party audits and second-party audits. For instance, a consultant hired by an AI vendor or operator to execute an audit performs a second-party audit.

2. For example, see the Sandvig v. Barr case led by the ACLU. Details available online at: https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti.

- **Support and enable third-party audits of AI systems.** Rather than solely focusing on internal accountability guidelines, we propose that regulators consider their potential role in supporting and enabling third-party auditors. Examples of recent policies to support regulatory and academic auditors include Article 31 of the EU Digital Services Act (Ponce, 2020), the FTC review requirement in the Algorithmic Justice and Online Platform Transparency Act of 2021 (Markey, 2021), and the requirements for state agencies to produce automated decision system accountability reports for ADS vendors proposed in the California Automated Decision Systems Accountability Act (Chau, 2020).
- **Strengthen tech whistleblower protections** for employees of both private and public institutions who blow the whistle on algorithmic products, services and practices that violate standards, legal and regulatory requirements, civil rights and human rights law.

2. Accreditation for auditors

A formal accreditation process for first-, second- and third-party auditors would improve the quality and trustworthiness of AI audits. AI auditors should be reviewed by an accreditation body or bodies to ensure they adhere to international standards and are adequately trained and qualified. That said, we caution that an accreditation process must not be controlled by industry and must be carefully organized to avoid excluding independent researchers.

There is an emerging industry of AI auditors, but few mechanisms to vet such auditors to ensure that they are both qualified and truly independent from the audit targets. Current AI policy has failed to provide enforceable public oversight, accreditation or certification of these auditors. AI system vendors are therefore currently able to hire anyone they like to do work they call "auditing," then declare their systems audited. In particular, many firms now advertise their ability to take on lucrative second-party audit contracts with government agencies or with private sector firms. AJL is currently mapping the landscape of AI auditors, and we have identified 65 entities that claim to conduct first-, second- or third-party audits (Algorithmic Justice League, 2021).

If an individual wants to provide medical services, they must be accredited by a medical board; if they want to practice law, they must pass the bar exam in each place they practice. While there are multiple possibilities for how accreditation may be organized, we believe that accreditation for AI auditors is an important piece of the puzzle. Accreditation of auditors can provide a structured vetting process to help build trust in AI systems, increase access permissions for accredited auditors, and ensure standardized quality of audits.

We also note that auditor accreditation rests on clear definitions and standards for what counts as a valid audit. In the absence of accreditation, the door remains open to multiple scenarios where AI audits do not serve their intended function. AI auditors may provide excellent services, but they may also be more or less technically competent, or more or less aware of key social, historical, cultural, or other contextual factors. Many existing AI auditors focus only or primarily on the technical aspects of "algorithmic fairness," rather than on deeper goals of algorithmic equity or minimizing harms throughout the AI lifecycle. Disembodied AI audits that focus solely on model performance without also considering the contexts, products, operators and communities that interact with real-world systems cannot adequately address emerging harms and threats. What is more, the quality of audit data matters. If AI auditors only have access to data that does not adequately represent the target system, their findings may either be overly optimistic about or overly critical of the target.

Additionally, without independent accreditation, well-meaning, fraudulent or opportunistic actors may provide thin or weak algorithmic audits. There are powerful incentives for firms to seek auditors who will provide a stamp of approval, sometimes referred to as ethics-washing (Bietti, 2020), even when

AI systems result in real-world harm. There are also disincentives for monitoring the continued performance of a system once a stamp of approval is received; this is a problem because usage and context continually evolve and may produce new harms.

However, we also note that accreditation must be implemented with great care. The accreditation process must be transparent and must be safeguarded from corporate capture. In the worst-case scenario, accreditation can be used to unduly exclude the very civil society organizations that might be best positioned to represent communities most likely to be harmed by AI systems. The details of formal AI auditor accreditation need careful consideration to ensure that the process is inclusive of the variety of participants engaged in third-party auditing.

We thus recommend the following policy interventions:

- **Foster AI auditor education and training.** Considering the worldwide scale of AI systems deployment across every sector, and the freewheeling use of the term "AI auditor," we consider the actual community of individuals and organizations with practical AI audit experience to be quite small. There needs to be a much more concerted effort to train and coordinate the AI auditing community. We believe policymakers should consider establishing support for the education and training of AI auditors in general and third-party AI auditors specifically. This will help ensure the development of a robust ecosystem that includes localized organizations able to conduct audits and able to certify that AI systems meet agreed-upon standards, conform to local and national regulatory requirements, and do not violate the law (including human rights law). We also encourage the development of norms for auditors to be able to evaluate AI systems against both corporate principles (such as internal AI ethics principles), international technical standards, and community expectations or demands.
- **Invest in the development of AI audit tools, templates and procedures.** This can help standardize practices in the audit field and consolidate expectations. Standardization around particular tools and processes can also facilitate education, training and testing.
- **Assess auditors for independence.** Those participating in second-party or third-party audits could possibly be misrepresenting themselves as independent, disguising funding or affiliations to corporations (Abdalla and Abdalla, 2021), or failing to disclose conflicts of interest. Accreditation should require an inspection of the conflicts that might impede an auditor's ability to execute in each context or for a particular target.

Formal processes for training, vetting, and accrediting auditors will be a necessary component of future AI policy.[3] If organized well, recognized accreditation bodies that can evaluate whether auditors are capable of evaluating AI systems may become a key tool for advancing the development of equitable and accountable AI systems. AI policy that promotes auditor accreditation can help ensure that AI systems comply with industry standards, with legal and regulatory requirements at the local, national and international levels, and with the Universal Declaration of Human Rights (UDHR).

3. Standards

AI policy needs to support the development of clear and widely recognized standards for AI products and processes that embody high-level expectations for AI systems and their use. Clearly defined standards, developed through a transparent process, are a prerequisite for meaningful AI audits.

3. For an overview of the ISO approach to accreditation mechanisms, see this resource available online at: https://www.iso.org/conformity-assessment.html.

In order to execute any evaluation, a standard is required to compare the reality of the system's performance to a given expectation or ideal. For AI audits in particular, an auditor needs to be given clearly defined standards. The auditor can then explore how the system or product measures up to expectations and hold the involved stakeholders responsible if the system falls dangerously short of those expectations. Standards thus play a crucial role in establishing the performance norms and requirements necessary for reliable audit practice. At times, these standards are determined and enforced by government agencies and expressed through formal laws or regulations. Alternately, some standards are developed by industry actors, often negotiated through consensus processes.

Unfortunately, because AI is a field with little regulation and few opportunities for centralized industry coordination, widely agreed-upon standards for AI systems, products and processes remain largely rudimentary or non-existent (Mittelstadt, 2019). At the international policy level, AI standards are underdeveloped but are slowly emerging. AI engineering standards are in development within multiple national and international standards bodies, such as the National Institute of Standards and Technology (NIST) in the USA (Cochrane, 1966),[4] the International Organization for Standardization (ISO)[5] and the Institute of Electrical and Electronics Engineers (IEEE) internationally (Shahriari and Shahriari, 2017)[6] and other bodies. In the absence of consensus on standards, technology corporations have moved to develop publicly shared AI principles (Jobin et al., 2019) or internally shared deployment criteria (Raji et al., 2020) to convey their own understanding of ethical expectations about AI systems. However, at best these self-regulation measures tend to be high-level, difficult to operationalize, and voluntary (Bietti, 2020). Most importantly, such corporate principles statements do not necessarily emerge from a grounded understanding of harms and are rarely constructed in consultation or collaboration with the most impacted communities (Metcalf and Moss, 2019). Separately, academic researchers and civil society organizations have also developed frameworks to think through their concerns and expectations for sociotechnical systems, but for the most part these also remain high-level and difficult for engineering teams to operationalize (Krafft et al., 2021).

Some proposals for AI accountability have focused on the certification of AI products (IEEE Standards Association, 2019). In certain contexts, product certification may be useful, but we caution that there are clear limits to any approach that uses standards as a checklist for deployment. Standards compliance should be considered a baseline or starting point that demonstrates bare minimum product performance, not the end goal. Rather, standards and benchmarks can help set performance expectations, help map and specify potential concerns, and operate as expressions of the idealized form of AI systems.

We thus recommend the following interventions:

- **Set standards as guidelines, not deployment checklists.** Ideally, standards should be flexible enough to accommodate changes in public understanding and attitudes. They can help guide audits, impact assessments, incident reporting, and other forms of evaluation, but should not automatically dictate AI product deployment conditions.
- **Set standards for processes, not only for outcomes.** In addition to standards for outcomes (such as accuracy rates in prediction and classification), it is also crucial that there be process-focused standards that provide expectations for the AI product development process. Process standards

4. Updates on recent developments in Artificial Intelligence Measurement and Evaluation at the National Institute of Standards and Technology available online at: https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop
5. Updates on recent developments on ISO/IEC JTC 1/SC 42 standard available online at: https://www.iso.org/committee/6794475.html
6. Updates on recent developments available online at: https://www.researchgate.net/publication/320253529_IEEE_standard_review_-_Ethically_aligned_design_A_vision_for_prioritizing_human_wellbeing_with_artificial_intelligence_and_autonomous_systems

should articulate best engineering practices around consensual data collection and use, documentation requirements (Gebru et al., 2021; Mitchell et al., 2019; Raji and Yang, 2019), minimum deployment criteria, harms incident reporting and response processes, and other evaluation processes. For example, a facial recognition technology (FRT) product that scores well on the Face Recognition Vendor's Test (FRVT) may still not mitigate privacy harms in data collection (Learned-Miller et al., 2020). This type of harm can only be addressed by setting procedural standards around how data is collected, distributed and put to use.[7]

- **Set standards for legal compliance, not only technical benchmarks.** Some second-party AI auditors claim to be able to assess system performance both against technical benchmarks and for regulatory and legal compliance.[8] We believe that accredited AI auditors should be able to evaluate both technical standards compliance and adherence to local, national and international law, including the UDHR. Accredited AI auditors should also be able to evaluate AI systems against both corporate principles (such as internal AI ethics principles) and proposals surfaced by community and civil society advocates in response to potential or actual harm to vulnerable communities. Additionally, such claims should be subject to third-party review, and those who offer second-party services of this nature should be required to meet clear standards laid out by accreditation bodies.
- **Set standards as evolving documents, developed through broad consultation, including with those most likely to be harmed by AI systems.** Finally, standards development itself is a perpetually unfolding process that must constantly respond to evolving contexts. While engineering standards are typically maintained by standards bodies such as the ISO and IEEE, we do have concerns about powerful private-sector or nation-state actors capturing or watering down standards processes. We would like to see standards processes include consultation with organizations representing those most likely to be harmed by AI systems, and not just restricted to researchers with ties to industry (Veale, 2020). We thus recommend the development of independently produced, flexible standards for AI products, to set baseline expectations for both product and process requirements.

4. AI Harm Incident Tracking

No AI system is perfect. We need AI harm incident reporting and tracking in order to ensure that those who are harmed by AI systems are able to share their experiences and concerns. Standardization of AI harm incident tracking supports a grounded understanding of problems, system improvements by vendors and operators, better oversight by regulators, legal action where necessary, and greater visibility of incidents in the press and in the public eye.

Incident tracking is a well-developed practice in some sectors, such as information security (Kenway and François, 2021). A robust incident response includes several key activities, including:

- Discovery—in other words, learning that an incident has occurred
- Reporting and tracking—documenting and sharing information about the incident
- Verification—confirming that the incident is reproducible, or indeed caused by the system in question
- Escalation—flagging the incident in terms of severity and urgency
- Mitigation—changing the system so that the problem does not continue to cause harm, ideally through a root cause analysis rather than a superficial patch

7. ISO/IEC 19794-5 about biometric data exchanges standard, full details available online at: https://www.iso.org/standard/38749.html

8. See, for example, Parity AI (https://www.getparity.ai) or Credo AI (https://www.credo.ai).

- Redress—taking steps to ensure that anyone harmed by the problem feels that the harm they suffered has been recognized, addressed, and in some cases, compensated, and
- Disclosure—communicating about the problem to relevant stakeholders, including other industry actors, regulators, and the public.

If organized well, harm incident reporting guides those who have experienced harm from AI systems (or their advocates) to provide informative descriptions of their experience that can then be used to expose problematic systems, improve (or in some cases, shut down) those systems, and seek redress. Systematic collection of AI harm incident reports is a critical step towards gaining a better understanding of the risks associated with AI system deployment, and towards ensuring the minimization, mitigation and redress of harms. However, there are currently no existing policy proposals, requirements, norms, standards or functional systems for AI harm incident tracking.

We thus recommend the following policy interventions:

- **Develop AI harm incident tracking standards and databases.** The UN system, as well as national agencies and regulators, should collaborate to develop and maintain AI harm incident databases to document, track and share known cases where AI systems violate existing laws or otherwise harm people. Such incident tracking systems need to be tailored to meet the needs of each legal jurisdiction, but ideally would follow agreed-upon standards in terms of incident classification, evaluation of the level of severity, and more. An international AI incident database maintained by the UN would ideally set the standard and could be interoperable with national-level projects. We see this as similar to the evolution of shared industry-wide incident reporting in the other sectors, such as cybersecurity.
- **Require harm incident reporting and tracking.** In addition to laying out standards for harms incident reporting, AI policymakers should require both vendors and operators of AI systems to provide clear mechanisms to report harms, abuse, disparate impact, system failure and other incidents. Especially for high-risk systems, regulators should require vendors to regularly disclose summaries of incident reports, including the frequency and severity of incidents and the steps taken to mitigate the problem. Publicly accessible incident reporting and tracking mechanisms can increase accountability through a combination of pressure from regulators, journalists, class actions and the broader public, as well as through private-sector competition. A publicly maintained database or interoperable standards for incident tracking and reporting—or both—will help ensure that various actors in the ecosystem can address identified issues, share knowledge about common problems across sectors, and build trust through opening AI systems to greater external scrutiny.

5. Notice of use

AI accountability policy should include mandatory public disclosure of AI systems use for any system with the potential to cause harm. Public agencies, in particular, must be required to notify the public when they procure, pilot and deploy AI systems. Public disclosure of use makes it possible for third-party auditors to identify audit targets and is a basic requirement upon which to build meaningful consent from those who will use and be impacted by the AI system.

Often, the people who are directly harmed by AI systems don't know about the ways that a particular product or tool might have been used to hurt them. They know what harm looks like, what it feels like, and what it means for their daily life, but may struggle to identify the contribution of a specific product to their predicament. In some cases of AI harm, people learn about the system after they notice a vendor name or a user interface, or they are informed about the AI system by an institutional actor such as a law enforcement officer or a legal aid worker. To change this norm, we believe that AI policy must begin to mandate public disclosure of AI systems use.

Public disclosure is one of the weakest links in the current AI regulatory landscape. The public has a right to multiple forms of disclosure about AI systems. Institutions making use of an AI tool need to release information about the fact that the tool is in use, why and how it was procured, how the tool is performing, and whether the tool is known to have caused any harm. We need to develop norms and laws that ensure people are notified when AI systems are in use, and that people are given information about how to opt out, appeal decisions and report harms (Brennan Center for Justice, 2017). If the AI system requires specific guardrails for safe and effective use, this should also be disclosed.

Public disclosure of the use of AI systems improves people's ability to understand what is happening, and it also enables third-party auditors to identify audit targets. Of course, the level of disclosure and public notice varies depending on several factors, including the level of risk and severity of possible harm from the AI system. Disclosure and transparency requirements might be quite different for public versus private institutions.

We thus recommend the following policy interventions:

- **Set a notice of intent to develop or deploy AI systems.** When public agencies and institutions intend to develop or deploy AI systems, the public must be notified and consulted from the beginning of the process, with an urgency according to the level of risk and the severity of harms that might result. Additionally, sufficient information and access to allow meaningful assessment needs to be disclosed to regulators, to the public, and to accredited third-party auditors.
- **Set a notice of use.** In some cases, it is possible for regulators to require both public and private actors to disclose AI system use. For instance, in the court system, each criminal defendant subject to a recidivism risk assessment should be notified of use. In the private sector, in some contexts it is possible to require that each job applicant must be informed if their application is being processed by an AI screening tool, and existing law may contain provisions that allow some (or any) applicants to opt out, such as reasonable accommodation provisions under the Americans with Disabilities Act. Notice of use may also extend to notice of data collection similar to that recommended in the GDPR, such as consent notifications about information collection when visiting websites, or surveillance notices when in the presence of surveillance cameras.
- **Provide institutional explanation and justification of use.** People need to know when and how AI systems are being used in both public and private products and services. This information must address AI system capabilities and limitations in an easy-to-understand manner. Additionally, institutions need to justify why the AI system is being used.
- **Require consent.** In many contexts, both public and private, AI policy can directly require user consent, for data collection and use as well as for participation in automated decision-making processes. Regulators may require opt-in design over opt-out design in data collection; for example, Facebook faced a major class-action lawsuit by users whose biometric data was harvested without informed consent (Singer and Isaac, 2020). The collection of personally identifiable information, biometrics and other sensitive data, as well as the use of such data to develop downstream machine learning models, should require explicit consent.
- **Include mechanisms to ensure equitable and accountable AI systems in government procurement processes.** Any time a government agency opens a procurement process for an AI system, they should have detailed public disclosure and consultation requirements. In federal systems such as the USA, state and local governments that receive federal funding should also face these requirements. There should be a robust and transparent public process for developing procurement requirements. These might include public notice of intent to deploy an AI system, with a justification, a comment period, and hearings; performance reporting requirements such as Model Cards (Mitchell et al., 2019),

Algorithmic Impact Assessments[9] (McKelvey and MacDonald, 2019), or Datasheets (Gebru et al., 2021); the disclosure of key results from audits and impact assessments; and requirements for consent, opt-out procedures and appeals, as well as incident reporting and response.

- **Include robust equity and accountability requirements for publicly funded AI systems.** Public-private partnerships, private sector contracts, research grants to academic institutions, and state and local governments who receive federal funding to develop AI systems including (but not limited to) automated decision systems (ADS), AI-enabled products, or general-purpose models should all be subject to robust equity and accountability requirements. Similar to procurement, these might include documentation of model performance, disclosure requirements to inform relevant stakeholders, pre- and post-deployment impact assessments and more.
- **Require publicly funded AI research to collect data that enables disparate impact analysis.** Academic institutions and government funding agencies such as (in the US context) the National Science Foundation (NSF), the National Institutes of Health (NIH), the Defense Advanced Research Projects Agency (DARPA) and others should increase understanding of the risks, harms and limitations of AI systems by requiring publicly funded AI research to collect demographic and other categorical information relevant to disparate impact analysis, as well as document the sourcing, labeling and interpretation of data collected.
- **Develop and institute mechanisms to improve private-sector disclosure of AI systems use.** In general, while public agencies have disclosure requirements, often private companies do not. For example, while a public housing authority can be compelled to disclose that it uses a tenant screening AI system via a public records request, a private landlord can use such a system without telling anyone. Prospective tenants who are concerned that they were denied housing because of their gender, race or disability may never know that an AI system was involved in screening them.

6. Frame shift: Beyond AI bias to AI harms

AI policy should address a broad range of AI harms, rather than focus only or primarily on technical measures of accuracy and bias. This also requires meaningful definitions of all key terms, and acknowledgement of multiple forms of harm.

At the Algorithmic Justice League, we regularly meet with community-based organizations and individuals who are directly harmed by AI systems. They don't tend to talk to us about accuracy rates or algorithmic bias. They say that they worry about rent, wonder how they'll pay for groceries and whether their kid is doing well in school, and now on top of all that, they have to worry about how they would navigate being locked out of their building late at night by a faulty facial recognition system, installed at their front door by a landlord without their consent (Bellafante, 2019).

AI systems may be used to cause harm in multiple ways. Disparate accuracy rates in prediction and classification between various groups of people are harmful, but so are systems deployment without notification or consent, opacity in how the system makes determinations, a lack of opt-out or appeals process to contest AI decisions, dysfunctional systems that do not perform to advertised expectations, and AI systems that are developed and deployed in ways that violate people's privacy or security.

We thus recommend the following policy interventions:

9. See, for instance, Canada's algorithmic impact assessment: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

- **Shift AI policy frames from the narrow lens of bias to a more expansive discussion of AI harms.** A shift in framing from the narrow discussion of bias to a broader understanding of algorithmic harm is necessary to ensure that third-party audits address a broader range of concerns, based on the priorities of the community that auditors may represent. This broader framing also lends itself well to external validation techniques, allowing for auditors to probe and critique the system for issues beyond bias.
- **Focus AI accountability policy on impact, rather than solely on system accuracy.** A debiased AI system can still be an unfair system. The bias frame tends to emphasize mechanism over impact. Yet a system that has been debiased to meet some particular benchmark related to a specific task, but is disproportionately deployed on a particular population with a negative impact, is still harmful. While exploring the origin or root cause of harmful decisions by AI systems is essential, we need to understand the importance of analyzing downstream outcomes as well. Our goal is not only to de-bias a system according to technical standards, but to shift the lived realities of those who are impacted, according to their own standards. Any harm, even if it seems small, is worth communicating about. The situation does not need to escalate for individuals to feel that they have been negatively impacted. Anyone who experiences harm from an AI system should be given the opportunity to voice their concern.
- **Focus on harms to lay the groundwork for legal action.** Bias in models is not necessarily legally actionable, but documenting harms, in terms of downstream impacts and torts, can contribute to a more effective legal strategy. A focus on harms helps prepare the ground for legal actions, such as class-action suits, that are important redress mechanisms. Although anti-discrimination law has dominated discussions about legal liability for fairness issues, many of the formal definitions of fairness in AI systems remain incompatible with legal notions of discrimination (Xiang and Raji, 2019). In addition, we need to have more conversations about tort law, product liability, negligence, consumer protection and other legal concepts that can be leveraged to protect people from harmful outcomes. If third-party AI auditors focus beyond bias to include harms, they can contribute to legal action that serves to shift the lived experiences of those currently enduring negative circumstances brought on by unaccountable AI systems.

7. Post-audit accountability mechanisms

Ultimately, third-party audits are only useful if there are multiple mechanisms to ensure that the issues they uncover are addressed. AI policy should include various enforcement tools to ensure that, in response to audit outcomes, firms disclose key audit findings, make improvements accordingly, seek compliance with standards and with the law, and redress harms.

Accountability is key. We have to hold those who hold power over AI systems development, deployment and use accountable for the impact they have on vulnerable people and communities. By focusing on minimizing harm, we prioritize improving the lives of those who are directly impacted by unaccountable AI systems and place less emphasis on arbitrary goal posts and metrics that may not be directly relevant to people's lived experience. Yet accountability requires not only that audits be conducted, but that they be acted upon. Audit outcomes need to lead to material changes in the lives of those impacted, through the removal or redesign of the problematic AI product in question.

We thus recommend the following policy interventions:

- **Monitor deployed AI systems continuously**, especially for disparate impact on marginalized populations. Similar to the cybersecurity sector, where standards and norms have evolved around the concept of the Secure Development Lifecycle (SDL), we need to shift the AI field to ensure equity and accountability across the entire project lifecycle. That includes AI system conception, planning, data gathering, model development, testing, deployment and post-deployment. Single-point-in-time audits are not enough to ensure equity and accountability. Although some recent legislation establishes pre-deployment audit and impact assessment requirements, we believe it is important to develop mechanisms to ensure ongoing evaluation, including by third-party auditors. An AI system that passes

muster in the lab may still produce harm, including unlawful disparate impacts, once deployed. Disparate impacts or other harms may arise post-deployment for a variety of reasons, including complex systems that evolve over time and changes in the ways that system operators configure or use the AI tool.

- **Require audit response and harm mitigation plans.** With standardized auditing, continuous monitoring and incident reporting in place, we should also require AI system vendors and operators to develop and implement harm mitigation plans that govern their response to potential and actual harms that are revealed at any stage. In cybersecurity, it is now standard practice for companies to have incident response teams, as well as clear systems for reporting, verification, escalation and resolution. This needs to become the norm in AI systems as well.
- **Require public disclosure of key audit results.** While AI system vendors have some legitimate objections to public disclosure of information about their products, including both trade secret concerns and the desire to protect their users' personally identifiable information, we believe that AI policymakers should make public disclosure of key audit results mandatory. Information about performance against known standards and benchmarks, including the results of first-, second- and third-party audits, needs to be publicly available and accessible. Policy requirements for the public disclosure of key results would dramatically transform accountability.
- **Redress harms**. Finally, AI system vendors and operators must be held accountable to take action and address findings and recommendations from accredited auditors, harms incident reports and other revealed shortcomings or harms. AI systems vendors and operators should be required to make improvements accordingly, seek compliance with standards and with the law, and redress harms revealed by auditors.

CONCLUSION

In this chapter, we have described a series of missing links in the current AI policy discussion that we believe are necessary to enable credible and effective third-party audits of AI systems. As a key form of oversight, third-party audits should be supported, protected and encouraged through multiple policy interventions. Policymakers can take specific actions to enable the participation of external auditors in building more equitable and accountable AI across domains. We have proposed a series of interventions that we consider necessary, and we believe that these interventions should be brought in from the periphery of current policy discussions. Third-party auditing must become a central component of future AI policy proposals.

We described seven interventions that we see as critical to support third-party AI auditors. These interventions include: legal protections for third-party AI auditor access; accreditation for AI auditors; standards development for AI products; AI harm incident reporting; mandatory public disclosure of AI systems use; a frame shift beyond AI bias to harms, and accountability mechanisms to ensure appropriate actions when audits reveal that AI systems depart from standards or violate relevant local, national, and international law, including the Universal Declaration of Human Rights.

These initial areas of concern are just the beginning of a necessary reconsideration of the role of third-party auditors in broader AI accountability measures. We hope that our proposals help generate meaningful discussion and actions by policymakers as we work to ensure a world of more equitable and accountable AI systems.

REFERENCES

Abdalla, M. and Abdalla, M. 2021. The Grey Hoodie Project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 287–297. https://dl.acm.org/doi/abs/10.1145/3461702.3462563

AICPA. 2017. *Audit and Accounting Guide Depository and Lending Institutions: Banks and Savings Institutions, Credit Unions, Finance Companies, and Mortgage Companies*. Hoboken, NJ: John Wiley & Sons.

Algorithmic Justice League. 2021. *AI Audits Landscape Mapping 2021 (public)*. Google Docs. https://docs.google.com/spreadsheets/d/17MP8s0PxTluEt1YOV4kWeBz2SpEqk7VunyZVSWDGA54/edit?usp=embed_facebook.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. 2016. Machine bias. ProPublica. May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ASQ. n.d. *What is an Audit? Types of Audits & Auditing Certification*. ASQ. https://asq.org/quality-resources/auditing.

Bellafante, G., 2019. The landlord wants facial recognition in its rent-stabilized buildings. Why? *New York Times*. March 28. https://www.nytimes.com/2019/03/28/nyregion/rent-stabilized-buildings-facial-recognition.html

Bietti, E. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 210–219. https://dl.acm.org/doi/abs/10.1145/3351095.3372860

Brennan Center for Justice. 2017. *The Public Oversight of Surveillance Technology (POST) Act: A Resource Page*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/public-oversight-surveillance-technology-post-act-resource-page

Buolamwini, J. and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Chau. 2020. Public contracts: automated decision systems, California Legislature, United States, No. AB-13. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13.

Cochrane, R. C. 1966. *Measures for progress: A history of the National Bureau of Standards*, Vol. 13. National Bureau of Standards, Department of Commerce, US.

Cuevas, E. 2020. Chau introduces Automated Decision Systems Accountability Act of 2021. Ed Chau Assembly District 49, Press Release. 8 December 2020. https://web.archive.org/web/20210616125930/

European Commission. 2022. Digital services act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Foxglove. 2020. *We put a stop to the A Level grading algorithm!* London, Foxglove. Website news section announcement. August 17. https://www.foxglove.org.uk/2020/08/17/we-put-a-stop-to-the-a-level-grading-algorithm/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H. and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, Vol. 64, No. 12, pp. 86–92. https://arxiv.org/abs/1803.09010.

Gillum, J. and Tobin, A., 2019. Facebook won't let employers, landlords or lenders discriminate in ads anymore. Propublica, March 19. https://www.propublica.org/article/facebook-ads-discrimination-settlement-housing-employment-credit

Grother, P., Ngan, M. and Hanaoka, K. 2019. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. Gaithersburg, Md., National Institute of Standards and Technology.

Gunningham, N. 1993. Environmental auditing: Who audits the auditors? *Environmental and Planning Law Journal*, Vol. 10, pp. 229-238.

Hill, K. 2020. Wrongfully accused by an algorithm. *New York Times*, June 24. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

IBM. 2021. Trustworthy AI is human-centered. IBM website. https://www.ibm.com/watson/trustworthy-ai

IEEE. 2019. IEEE Standard for Safety Levels with Respect to Human Exposure to Electric, Magnetic, and Electromagnetic Fields, 0 Hz to 300 Ghz. *IEEE Access*, Vol. 7, pp. 171346–171356.

Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, No. 9, pp. 389–399. https://www.nature.com/articles/s42256-019-0088-2

Kazim, E. and Koshiyama, A. 2020. A review of the ICO's draft guidance on the AI Auditing Framework. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3599226

Keegan, J. 2021. Facebook got rid of racial ad categories. Or did it? The Markup, July 9. https://themarkup.org/citizen-browser/2021/07/09/facebook-got-rid-of-racial-ad-categories-or-did-it

Kenway, J., and François, C. 2021. *Bug Bounties for Algorithmic Harms? Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress*. Washington, DC: Algorithmic Justice League. https://www.ajl.org/bugs

Krafft, P. M., Young, M., Katell, M., Lee, J. E., Narayan, S., Epstein, M., Dailey, D., Herman, B., Tam, A., Guetler, V. and Bintz, C. 2021. An action-oriented AI policy toolkit for technology audits by community advocates and activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 772–781. https://dl.acm.org/doi/10.1145/3442188.3445938

Learned-Miller, E., Ordónez, V., Morgenster, J. and Buolamwini, J. 2020. *Facial recognition technologies in the wild: A call for a federal office*. Algorithmic Justice League. https://assets.website-files.com/5e027ca188c99e3515b404b7/5ed1145952bc185203f3d009_FRTsFederalOfficeMay2020.pdf

Lecher, C. and Varner, M. 2021. How we investigated NYC high school admissions. The Markup, May 26. https://themarkup.org/show-your-work/2021/05/26/how-we-investigated-nyc-high-school-admissions

Lytton, T. D. and McAllister, L.K. 2014. Oversight in private food safety auditing: Addressing auditor conflict of interest. *Wisconsin Law Review*, No. 6/2014, pp. 289–337.

MacCarthy, M. 2019. An examination of the *Algorithmic Accountability Act* of 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3615731

Markey, Ed. 2021. Senator Markey, Rep. Matsui introduce legislation to combat harmful algorithms and create new online transparency regime. Press release, May 27. https://www.markey.senate.gov/news/press-releases/senator-markey-rep-matsui-introduce-legislation-to-combat-harmful-algorithms-and-create-new-online-transparency-regime

Martinez, E. and Carollo, M. 2021. Dozens of mortgage lenders showed significant disparities. Here are the worst. The Markup, August 25. https://themarkup.org/denied/2021/08/25/dozens-of-mortgage-lenders-showed-significant-disparities-here-are-the-worst

McKelvey, F. and MacDonald, M. 2019. Artificial intelligence policy innovations at the Canadian Federal Government. *Canadian Journal of Communication*, Vol. 44, No, 2, pp. 43–50. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Metcalf, J. and Moss, E. 2019. Owning ethics: Corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research: An International Quarterly*, Vol. 86, No. 2, pp. 449–476.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R. and Elish, M. C. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 735–746. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3736261

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. https://dl.acm.org/doi/10.1145/3287560.3287596

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, Vol. 1, No. 11, pp. 501–507. https://www.nature.com/articles/s42256-019-0114-4

Open Government Partnership. 2021. *Algorithmic Accountability for the Public Sector*. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector

Ponce, A. 2020. The *Digital Services Act* Package: Reflections on the EU Commission's Policy Options. ETUI Research Paper-Policy Brief No. 12/2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3699389

Raji, I. D. and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435. https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/

Raji, I. D. and Yang, J. 2019. About ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. https://arxiv.org/pdf/1912.06166.pdf

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. https://arxiv.org/abs/2001.00973

Richardson, R. 2021. Defining and demystifying automated decision systems. *Maryland Law Review*, forthcoming. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3811708

Selbst, A. D. 2021. An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, Vol. 35, No. 10, pp. 117–191.

Shahriari, K. and Shahriari, M. 2017. IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pp. 197–201. https://ieeexplore.ieee.org/document/8058187

Simonite, T. 2020. Google offers to help others with the tricky ethics of AI. *Wired*, August 28. https://www.wired.com/story/google-help-others-tricky-ethics-ai

Singer, N. and Isaac, M. 2020. Facebook to pay $550 million to settle facial recognition suit. *New York Times*, January 29. https://www.nytimes.com/2020/01/29/technology/facebook-privacy-lawsuit-earnings.html

Spinks, C. N. 2019. Contemporary housing discrimination: Facebook, targeted advertising, and the *Fair Housing Act*. *Houston Law Review*, Vol. 57, No. 4. https://houstonlawreview.org/article/12762-contemporary-housing-discrimination-facebook-targeted-advertising-and-the-fair-housing-act

Team, A. P. 2021. *Artificial Intelligence Measurement and Evaluation at the National Institute of Standards and Technology*. Washington, D.C., National Institute of Standards and Technology. https://www.nist.gov/news-events/events/2021/06/ai-measurement-and-evaluation-workshop

United Kingdom Government. 2020. *Draft Online Safety Bill*. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/government/publications/draft-online-safety-bill

United States. 1984. *Computer Fraud and Abuse Act*, 18 U.S.C., § 1030. https://uscode.house.gov/view.xhtml?req=(title:18%20section:1030%20edition:prelim)

United States Congress, 2019. *H.R.2231 – 116th Congress (2019-2020): Algorithmic Accountability Act of 2019*, April 11. https://www.congress.gov/bill/116th-congress/house-bill/2231

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*, Vol. 11, No. 1, pp. 1–10. https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/abs/critical-take-on-the-policy-recommendations-of-the-eu-highlevel-expert-group-on-artificial-intelligence/FF6FF91A0C140E58B4B527C68E0C5321

Weiger, C., Smith, K.C., Cohen, J. E., Dredze, M. and Moran, M. B. 2020. How internet contracts impact research: Content analysis of terms of service on consumer product websites. *Public Health and Surveillance*, Vol. 6, No. 4, pp. 1–15. https://publichealth.jmir.org/2020/4/e23579/

Wieringa, M. 2020. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 1–18. https://dl.acm.org/doi/abs/10.1145/3351095.3372833

Xiang, A. and Raji, I. D. 2019. On the legal compatibility of fairness definitions. https://arxiv.org/abs/1912.00761

THE AI INDUSTRY THROUGH THE LENS OF ETHICS AND FAIRNESS

GOLNOOSH FARNADI

Assistant Professor of Machine Learning at HEC Montréal and Adjunct Professor at Université de Montréal. Core faculty member at MILA – Québec Institute of Artificial Intelligence, and holder of Canada's CIFAR AI chair.

AMANDA LEAL DE LIMA ALVES

Research assistant of the FATE – Fairness, Accountability, Transparency, and Ethics Lab led by Professor Golnoosh Farnadi. Member of the AI for Humanity team at Mila – Québec Institute of Artificial Intelligence.

REBECCA SALGANIK

M.Sc. student at Mila – Québec Institute of Artificial Intelligence and Université de Montréal. Member of the FATE – Fairness, Accountability, Transparency, and Ethics Lab led by Professor Golnoosh Farnadi.

SDG9 - Industry, Innovation and Infrastructure

SDG10 - Reduced Inequalities

SDG11 - Sustainable Cities and Communities

SDG12 - Responsible Consumption and Production

SDG16 - Peace, Justice and Strong Institutions

SDG17 - Partnerships for the Goals

THE AI INDUSTRY THROUGH THE LENS OF ETHICS AND FAIRNESS

ABSTRACT

Nowadays, every sector is part of the artificial intelligence (AI) industry. Agriculture, healthcare, finance, education, and art, among many others, not only use machine learning and AI models throughout their supply chain, but each one of us interacts with a whole ecosystem of their algorithms, perhaps without even realizing it and, what's more, without an understanding of the principles that guide those models' functioning for better or for worse. In this work, we attempt to shed light on how the industry has been dealing with AI ethics, pinpointing the cracks where unfairness has sprung, and exploring possible ways forward. We speak to seven ethical AI researchers from both academia and industry to unveil the current scenario of AI ethics in industry: the challenges, the opportunities, and the stakeholders. We explore three of the most prevalent issues raised by the experts, followed by an analysis of three possible avenues for ethical AI.

Because AI models play an increasingly prevalent role in our lives, changing the way they work will only get harder. Numerous reports of discriminatory behavior have arisen, showing just how biases can become deeply embedded in the technological pipeline. The culture of "moving fast and breaking things" has come with a whole host of negative societal consequences. We describe how a lack of unified fairness metrics, as well as lack of diversity and ethical standards, have formed a perfect storm of unaccountability and inaction. We then explore how broader participation can pull back the curtains of the technology sector and democratize the discussion surrounding fairness in AI. We also look at how raising awareness and broadening access to an ethics-based AI education can create substantive change in the way we approach the design and deployment of technology. Finally, we argue that the democratization of participation and a establishment of a level playing field for discussions must be followed by concrete policy, regulation, and organizational reforms.

INTRODUCTION

In recent years, the deployment of AI algorithms, and the machine learning models they use to reason about the world, has grown into a massive technological movement that permeates all facets of our lives. Agriculture, healthcare, finance, education and art, among many other domains, have embraced the use of algorithmic decision-making. Every day, we are interacting with a whole ecosystem of algorithms, not only as active users, but also passively, without even realizing it. As a byproduct of this, the complexities of AI have come to directly affect people's quality of life.

However, even as we have welcomed these algorithms into our world economies, there are a series of crucial questions which have been left unaddressed: Who oversees the potential impacts of AI development and deployment throughout our major industries? How do we guarantee that the negative effects of these systems won't outweigh the potential benefits? How is the gap between AI development, deployment and its outcomes being addressed in the tech industry?

The urgency of these questions grows exponentially with the speed at which nascent technology is being introduced into our lives. There are inherent limitations to the understanding that systems, created to reason mathematically about a world which mathematical rules can't fully translate, can achieve. Without careful oversight, those systems, when deployed in such a vast repertoire of applications, are bound to produce outcomes that weren't predicted during the design process. In situations where these models are given immense power and little oversight, their mistakes can have truly dire consequences. Unfortunately, this is an aspect of AI that is often ignored.

Recently, the concrete ethical costs of certain AI systems have come to light. In the last few years, experts have uncovered cases in which models were perpetuating, and even cementing, the forms of discrimination that underlie our society (Angwin et al., 2016, Spielkamp, 2017, Yong, 2018, Grind et al., 2019). Even more concerning was the finding that their algorithmic discrimination was specifically targeting historically disenfranchised demographic groups. The outcry that followed these discoveries motivated the development of a new field: algorithmic fairness in machine learning.

Intuitively, the basis of algorithmic fairness is to ensure that machine learning models do not discriminate against certain individuals and groups across society. In the common discourse, algorithmic fairness has often been treated as a specialized, technical, or academic pursuit. But fairness is a topic that relates to us all, not just coders. In fact, the values and premises of algorithmic fairness have their roots in concepts such as equality and non-discrimination. Of course, one could ask why we need such a movement when there are so many consolidated social and legal concepts that guide anti-discriminatory practices on many levels. However, there seems to be a paradox in the accountability around unfair practices when it comes to AI applications. In situations where a human being would be held responsible for committing discrimination, those behind the design of an AI model that is perpetrating discriminatory behavior don't face the same consequences. This inconsistency has created a situation in which AI systems could be perpetrating large-scale unfairness against a group of people within our society without anyone ever knowing. In this way, algorithmic fairness is not a technical, but a socio-cultural issue.

While academic efforts are important first steps towards addressing fairness concerns, this alone is not enough. How do we bridge the gap between a nascent and promising academic field of algorithmic fairness, the industry practices that shape the AI-powered products and services, and all of us: consumers, businesses, civil society and governments? Whose needs should AI be serving?

The field of AI is still in its fledgling stages. But as these models play a more prevalent role in our lives, changing the way they work will only get harder. The opacity of AI pipelines and their entanglement with deployment structures often makes it difficult to pin down where exactly things are going wrong. This

is because discrimination can happen anywhere in the AI process: in the data, in the model, and even in the outcome. In order to identify actionable change, we need to unveil the inner workings of the AI industry and understand the challenges in producing ethical products.

In a sense, "the only way out is through." To prepare this chapter, we reached out to seven ethical AI researchers from both academia and industry with diverse backgrounds who all share our commitment to advancing an ethics-awareness movement in the AI industry. Their perspectives reflect not only their expertise, but also the voices of underrepresented groups in AI, be it in terms of gender, race or geographic location. We prepared a set of questions relating to their perspectives of ethical AI in the tech industry, including issues of fairness and diversity, the benefits and drawbacks of AI, and potential avenues towards regulatory frameworks. Their answers were collected either via written or video interviews. In writing this chapter, we have selected the highlights of our discussions with these domain experts, focusing on three major gaps that jeopardize the development of ethical AI. In addition, we also present their valuable insights into three areas of engagement that can inspire a better AI ethics future in the industry. Ultimately, our goal is to open the black box of the AI industry and shine a light on necessary discussions surrounding current practices of AI ethics.

Participants

Margaret Mitchell: Margaret Mitchell is a researcher working on ethical AI, currently focused on the ins and outs of ethics-informed AI development in tech. She has published over 50 papers on natural-language generation, assistive technology, computer vision, and AI ethics, and holds multiple patents in the areas of conversation generation and sentiment classification. She previously worked at Google AI as a Staff Research Scientist, where she founded and co-led Google's Ethical AI group, focused on foundational AI ethics research and operationalizing AI ethics internally at Google. Before joining Google, she was a researcher at Microsoft Research, focused on computer vision-to-language generation, and was a postdoc at Johns Hopkins, focused on Bayesian modeling and information extraction. She holds a PhD in Computer Science from the University of Aberdeen and a Master's in Computational Linguistics from the University of Washington. While earning her degrees, she also worked from 2005 to 2012 on machine learning, neurological disorders, and assistive technology at Oregon Health and Science University. She has spearheaded a number of workshops and initiatives at the intersections of diversity, inclusion, computer science, and ethics. Her work has received awards from Secretary of Defense Ash Carter and the American Foundation for the Blind, and has been implemented by multiple technology companies.

Rumman Chowdhury: Dr. Rumman Chowdhury's passion lies at the intersection of AI and humanity. She is a pioneer in the field of applied algorithmic ethics, creating cutting-edge socio-technical solutions for ethical, explainable and transparent AI. She is currently Director of the META (ML Ethics, Transparency, and Accountability) team at Twitter, leading a group of applied researchers and engineers to identify and mitigate algorithmic harms on the platform. Previously, she was CEO and founder of Parity, an enterprise algorithmic audit platform company. She formerly served as Global Lead for Responsible AI at Accenture Applied Intelligence, leading the design of the Fairness Tool, a first-in-industry algorithmic tool to identify and mitigate bias in AI systems. Dr. Chowdhury co-authored a *Harvard Business Review* article on its influences and impact.

Francisco Marmolejo-Cossío: Francisco is a Postdoctoral Fellow at Harvard School of Engineering and Applied Sciences (SEAS) and a Research Fellow at Input Output Hong Kong (IOHK). Prior to this, he was a Career Development Fellow in Computer Science at Balliol College at the University of Oxford. He completed a D.Phil. in Theoretical Computer Science under the supervision of Paul Goldberg, and a B.A. in Mathematics at Harvard University with a minor in Neuroscience in 2012. He also co-organizes the Mechanism Design for Social Good (MD4SG) research initiative.

Arisa Ema: Arisa Ema is Associate Professor at the University of Tokyo and Visiting Researcher at RIKEN Center for Advanced Intelligence Project in Japan. She is a researcher in Science and Technology Studies (STS), and her primary interest is investigating the benefits and risks of AI by organizing an interdisciplinary research group. She is a member of the Ethics Committee of the Japanese Society for Artificial Intelligence, which released the AI Ethical Guidelines in 2017. She is also a board member of the Japan Deep Learning Association (JDLA) and chairs the AI governance study group. She was also a member of the Council for Social Principles of Human-centric AI of the Cabinet Office, which released "Social Principles of Human-Centric AI" in 2019.

Vidushi Marda: Vidushi Marda is an Indian lawyer and researcher who investigates the societal impact of AI systems. She currently works as a Senior Program Officer at ARTICLE 19, a global human rights organization, where she leads research and engagement on the human rights implications of machine learning. She is a member of the Expert Group on Governance of Data and AI at United Nations Global Pulse, and part of the steering committee at RealML. Ms. Marda's work engages with technical, policy, academic, and advocacy communities, and has been cited by the Supreme Court of India in a seminal ruling on the Right to Privacy, the United Kingdom House of Lords Select Committee on Artificial Intelligence, and the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, among others.

Joanna Shields: Tech industry veteran Baroness Joanna Shields has helped build some of the world's leading high-growth companies, including Google, Aol and Facebook, and led multiple startups to successful exits. Ms. Shields is currently CEO of BenevolentAI, a leading clinical-stage AI drug discovery company that uses machine learning and artificial intelligence to develop more effective medicines. She sits as Chair of the Multistakeholder Experts Group Plenary and Co-Chair of the Steering Committee on the Global Partnership on AI (GPAI), supported by the OECD, and previously served as the United Kingdom's first Minister for Internet Safety and Security and Under-Secretary of State, UK Ambassador for Digital Industries, Special Advisor to the Prime Minister on the Digital Economy, and Chair & CEO of TechCityUK and non-executive director of the London Stock Exchange Group. In 2014, Ms. Shields founded WePROTECT.org, a global alliance working to protect children from online abuse and exploitation. In 2014, she was appointed to the United Kingdom's Order of the British Empire for services to digital industries and voluntary service to young people, and made a Life Peer in the House of Lords.

Ulrich Aïvodji: Assistant Professor of Computer Science at ÉTS Montréal in the Software and Information Technology Engineering Department. His research areas of interest are computer security, data privacy, combinatorial optimization, and machine learning. His current research focuses on several aspects of trustworthy machine learning, such as fairness, privacy-preserving machine learning, and explainability. Before holding his current position, he was a postdoctoral researcher at Université de Québec à Montréal, working with Sébastien Gambs on machine learning ethics and privacy. He earned his Ph.D. in Computer Science at Université Toulouse III, under the supervision of Marie-José Huguet and Marc-Olivier Killijian. During his Ph.D., he was affiliated with the Laboratoire d'analyse et d'architecture des systèmes of the Centre national de la recherche scientifique (LAAS-CNRS) as a TSF and ROC research group member and worked on privacy-enhancing technologies for ridesharing.

WHY DO WE CARE ABOUT FAIRNESS IN THE AI INDUSTRY?

The common pitfalls of AI can be exemplified by the infamous COMPAS model (Angwin et al., 2016). COMPAS, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, was a tool developed by a private company to predict the risk of a criminal defendant committing another offense. This model was used throughout the court system in the United States as evidence during bail

hearings. In 2016, ProPublica, an investigative news organization, broke a story in which it alleged that the model was more likely to predict a high recidivism likelihood among Black defendants (Larson, 2016). In fact, when they dug into the numbers, ProPublica reporters found that Black defendants were almost twice as likely as Caucasians to be labeled high risk without actually going on to reoffend. Colloquially, this suggested that the model had begun to perpetuate profoundly harmful, untrue racial stereotypes about which demographics of citizens were more likely to commit crimes.

What did this mean and why did it happen? When this news broke, the company responsible for the COMPAS software, Northpointe, put out a statement informing the public that, by their mathematical notions of fairness, their tool was unbiased (Dietrich, 2016). And, in fact, after looking into their mathematical notions of bias, auditors confirmed that through these notions, the model was making "unbiased" predictions.

In this complex situation, we can see a perfect illustration of the shortcomings of ethics in the AI industry that we will be discussing in this chapter: 1) Lack of fairness definitions; 2) Lack of diversity; 3) Lack of ethical standards.

It is noteworthy that the primary reason this situation occurred was the lack of consistency in what defines fairness. First, we will explore how it is possible for an algorithm to be considered "fair" while it continues to perpetuate discriminatory behavior on a demographic group. We will also explain how a lack of diversity in the tech industry can exacerbate the lack of oversight in algorithmic consequences. Then, we will explore how lack of ethical standards, lack of regulation, and algorithmic opacity allow for the owners of AI systems to unanimously define what fairness means with respect to their products. We will analyze how a lack of broader participation in regulatory discussions and an overall lack of awareness of algorithmic fairness create a vacuum that allows discriminatory models to continue, unchecked. Finally, for the remainder of the chapter, we will go through these points, drawing from our in-depth conversations with ethical AI experts and culminating in a discussion of potential directions for improvement.

1. Lack of Fairness Definitions

At the crux of achieving fairness lies the question: what exactly is fairness and how can we define it in AI and machine learning models?

Intuitively, fairness involves ensuring that an algorithmic system's predictions do not unethically discriminate against a certain group or an individual. But, from an algorithmic perspective, fairness often needs to be defined in mathematical terms. It involves a notion of true positives, false positives, true negatives, and false negatives. For example, predicting that someone will commit a crime, when in reality they don't, is a false positive. Meanwhile, predicting that someone doesn't have cancer, when in fact they do, is a false negative. One of the difficulties of defining fairness is that it has different, often incompatible, definitions that depend deeply on the problem at hand. In fact, by 2018, in the field of algorithmic fairness, academics had identified 21 unique definitions of fairness, almost all of which are completely incompatible with one another (Verma, 2018).

> Ethical guidelines often recommend a list of trustworthiness properties such as fairness, security, privacy, explainability, transparency, that AI systems must exhibit. However, our understanding on how these properties interact with each other is still at an embryonic stage. Deploying these technologies without understanding these interactions is a pure fiction that will do more harm than good.
>
> - Ulrich Aïvodji

Even more striking is that none of these mathematical notions of fairness can truly capture the essence of a "fair" experience. As Rumman Chowdhury remarks in her discussion of fairness audits, this is because, from a user's perspective, fairness is not just a mathematical objective – it's an experience. It is not about just objectively assessing whether a set of algorithms is fair, but rather overseeing all aspects that compound a user's experience when interacting with the technology.

> You have multiple algorithms operating in tandem for a given user experience. [...] That's one thing that is very lacking [in the] regulatory conversation. We treat an algorithm and say "it's the algorithm and the algorithm needs to be audited." [But] there is no Twitter algorithm. Twitter has many algorithms that are working, but [you only have] one Twitter experience.
>
> - Rumman Chowdhury

As Ms. Chowdhury explains, even if we were to perform a formal audit of a single algorithm, user interactions with AI models are multi-faceted; they are interacting with an ecosystem of models, not just one. This problem only exacerbates the fluctuating definition of fairness: what if we allow false positives with one model and false negatives with another? What will this mean for the user?

Francisco Marmolejo-Cossío highlights that our fairness needs are ever-shifting, precisely because they are so context-based. This creates a situation in which research and implementation objectives are continuously being modified to resemble each unique fairness need. As we discover an ever-broadening range of contexts in which fairness notions must be defined, the list of fairness definitions expands to take into account even more complex notions of morality.

> [Fairness] is a very interdisciplinary practice. It's not just the technical from the STEM perspective, it's not just the algorithmic techniques, the optimization techniques we bring in, but it's also the societal context that goes into the specific features or input that we have.
>
> - Francisco Marmolejo-Cossío

A crucial point brought up by our experts is the profound difference between fairness metrics and the standard metrics currently being employed in the training of AI models. Margaret Mitchell points out this contrast and expands on its consequences:

> One thing is that we have to develop standard protocols for evaluation. The state of the art currently still is using metrics that other people have defined, the sort of normal, default thing in the literature like F1 score or whatever. And just focusing on making that number go up. What so many of us are trying to do in the safety and fairness space [is to say]: "No, you actually have to disaggregate model performance. You have to evaluate how the model does in a bunch of different contexts using a bunch of different metrics." And that's not a thing yet. That hasn't dawned on the machine learning community yet. Evaluation needs to be informed by societal context. So who's most likely to be harmed and how the system is most likely to be misused and [...] make problematic mistakes.
>
> - Margaret Mitchell

What Ms. Mitchell brings to light are the methods for evaluation currently being used to train AI systems. In the current industry setting, a model is often evaluated based on its ability to optimize some complex mathematical function that is meant to represent user satisfaction. For example, a video recommendation system would be taught to accurately predict a user's probability of clicking on a video. However, this evaluation is completely agnostic to the societal implications of someone watching this video. In fact, if the user then returned to a video, the model could be doubly rewarded, even if the video contained violent or disturbing content which had downstream consequences in the life of the user. Ms. Mitchell stresses the importance of creating methods for evaluation that are informed by societal context and argues for the consideration of new metrics that integrate risk assessments associated with various

outcomes. According to Ms. Mitchell, although fairness objectives are researched and presented in academia, they are rarely integrated into industry models. This creates serious issues because. if a model has never been tested for fairness, it's impossible to prove it's unfair.

Even more concerning is that oftentimes these models end up behaving in ways that their designers weren't intending. For example, in 2018, Amazon deployed a new AI hiring tool to fix the gender imbalances in their coding workforce (Dastin, 2018). However, it was quickly discovered that the model was disproportionately selecting male candidates over female ones. Even though the designers had expressly intended to avoid this problem, they had trained this model on a dataset fed with historically male-dominated C.V. data, which adopted a typical male C.V. as the gold hiring standard. In practice, this meant that any C.V. with different attributes, such as having attended a women-only university, were ranked lower (Dastin, 2018). Without changing their evaluation function to integrate and address this imbalance, the model continued to perpetuate the imbalance it had been exposed to.

> It's very hard for an individual to have a form of redress if they do think bias or discrimination has happened, because then what they ask is "Well, prove that algorithmic bias exists in the system." When you are not a data scientist, you don't know what's going on.
>
> - Rumman Chowdhury

As Ms. Chowdhury explains, the lack of fairness definitions and the industry's lack of participation in the fairness domain bar users from confronting the discriminatory practices of an AI model. As we move forward, this situation will only exacerbate the unfairness of structures that play an ever-present role in our day-to-day lives.

2. Lack of Diversity

At the root of fairness in any algorithm are the people who build it. As Arisa Ema argues, "*Who* is discussing it cannot be irrelevant to *what* is being discussed." We can tie the notions of team diversity with outcomes of algorithmic fairness. Intuitively, it is clear that, given the complexity of making a truly fair algorithm, without employing a pool of diverse thinkers, a company has almost no chance of achieving it. In a best-case scenario, we would hope that engineering teams are made up of a multifaceted, interdisciplinary pool of workers who are deeply in touch with the needs and concerns of their users. But, as we show in this section, that is not the case. Through the course of this discussion, we explain how lack of diversity creates an environment in which unethical AI practices can further take over.

Rumman Chowdhury explains how true fairness starts at the point of an algorithm's inception. She explains the reasons why, although fairness is often outsourced to "ethics experts," this does not necessarily increase the ethics standards.

> And the problem with [separating functions to oversee AI, such as with privacy and security] is you end up with active resistance, passive resistance and general inaction. Or, as we're finding, people think this actually increases your workload because people don't want to do it. [...] One of the reasons we have chosen to [adopt this behavior] is [because] you center the model owner as a person with the most expertise about the model. They're the one who built it. They're the one who knows the most. [There is] this idea of moral outsourcing like, "Aren't you the ethics people? Shouldn't you go do that then, like, go make my model ethical?" But I don't have the staffing to go into your model, make these changes, et cetera, and then come back to you for your approval. It's a failure state. [...]
>
> - Rumman Chowdhury

If ethics are only considered after a product has already been developed, there is a much smaller chance that foundational changes will be made. Separating the functions of product development and ethics creates an artificial division between product design and its fairness outcomes. Furthermore,

it undermines collaboration between creators and auditors, placing the burden of fairness on people who are often given less organizational power. Ms. Chowdhury argues that evaluating the fairness of the algorithm should be a concern throughout its design and development, and not just outsourced to another team after the work has already been done. In other words, it is necessary to take a step back and reassess the value of interdisciplinarity: both for society, as it would guarantee more oversight over the entire process of creating technologies, and for the industry, as it would likely cost less in the long run.

The need for a fairness-aware workforce makes hiring decisions a critical step in the ethical AI pipeline. Without having fairness-aware employees who can take ownership of the AI model from the ethical perspective, companies are forced to rely on outsourcing their ethics assessments to external players.

> Team diversity is critical, not just within fairness, but in the industry in general. We are building products and providing services that change the way people live and the intent of these products is usually to apply to the world at large. So the teams that build these products need to reflect the audience that they want to apply their tools to.
>
> - Rumman Chowdhury

As Ms. Chowdhury and Francisco Marmolejo-Cossío explain, the first step to achieving a fairness-centered team is diversity. In this way, they frame diversity hiring as a crucial element in ensuring algorithmic fairness.

> [The] entire back and forth process – even for those of us who are a little bit more theoretically-minded – of refining models with more on-the-ground reality is very important. Diversity of background in that context is super important. This, of course, comes from a multitude of different angles. There's socioeconomic diversity, geographical diversity, gender diversity is also a huge issue. As a male helping lead some of these groups, I find that some of the best ways that worked for us is trying to maintain a very diverse leadership.
>
> - Francisco Marmolejo-Cossío

Here the tie between diversity and its effects on fairness come to surface. Companies that create AI models need to understand the consequences of their work. So, as Ms. Chowdhury and Mr. Marmolejo-Cossío point out, the teams who build these products need to, at least somewhat, reflect the audience their tools are applied to. In this way, the gap between creators of a technology and those who assess the impact of how it's deployed becomes a critical shortcoming that is directly related to the lack of diversity in AI teams.

In recent years, the tech industry has started responding to society's calls for hiring workers from a broader range of perspectives. But, unfortunately, gathering a diverse team has many obstacles, especially when it comes to highly specialized positions. As Margaret Mitchell points out, even when companies want to hire workers from minority groups, the tech environment may not empower them to grow in their roles.

> Diversity in tech is generally horrible, in part because while companies can (somewhat) understand what diversity is, there doesn't seem to be an understanding of what inclusion is, or how it works. That means that while companies may be able to *hire* people with characteristics that are underrepresented in tech, they struggle to retain. For those with underrepresented characteristics, diversity without inclusion is career torture.
>
> - Margaret Mitchell

This is compounded by the fact that only the people at the top get to make critical design decisions. Just hiring a diverse pool of new engineers will not change the fact that more senior positions would still be dominated by a lack of diversity.

> To make matters worse, diversity efforts tend to focus on people that leaders in a company see as "below" them, such as new engineers. As people at the lower levels have very little say in defining the culture, while people at the higher levels who define culture tend to be those with the dominant characteristics in tech, the company as a whole can become structurally racist, sexist, etc., pushing out those that don't align with the expectations and norms of the culturally required behaviors defined top-down.
>
> - Margaret Mitchell

In her discussion of the challenges of promoting trustworthy AI within companies, Ms. Mitchell overly ties inclusion and diversity of a workplace with algorithmic fairness outcomes.

> One of the main barriers in getting ethical AI to work within companies is that there isn't the bottom-up information flow. The higher the level you are, the more you are just given the power of someone with expertise. That same point means that if you're lower-level, you can't tell anyone anything. No one else will listen to you or care because their company has declared them the expert. And often people who are at the higher levels seem to think that they're the experts. You know, if they've been at the company that long, it's kind of like a groupthink. So there needs to be a possibility of having bottom-up information flow from the experts at the company to the leadership. And that's not possible right now.
>
> - Margaret Mitchell

Francisco Marmolejo-Cossío also raises the point of how diversity and fairness are intertwined. He notes that a lack of diversity in AI teams is one of the possible reasons why they won't look for inputs from people on the ground, which is crucial to test for fairness requisites in AI applications.

> Perspectives from people who work on the ground are super crucial. They bring in relevance with respect to whether implementing something in a specific way would entirely miss some on-the-ground reality. In some of the other groups that I have worked in the past, this consideration is lacking either for by simply not being there, because of systemic issues, a lack of diversity, but also sometimes by convenience too. A perspective based in a kind of mathematical convenience sometimes come into play when working with the models.
>
> - Francisco Marmolejo-Cossío

Vidushi Marda calls for a greater reform with respect to the power structures within tech companies which would extend beyond low-level hiring practices. In doing so, she highlights how power dynamics can have broad trickle-down effects into discussions of both fairness and diversity.

> While diversity in teams is an important strength needed in order to holistically understand the impact of AI in societies, the ethical AI industry needs a much more fundamental reckoning, as it is currently people in power who decide what standards are, how they are met, and when they are satisfactory. This represents a misalignment of incentives and efforts to preclude concrete regulation and accountability at the minimum.
>
> - Vidushi Marda

From Ms. Marda's discussion, it becomes clear that although diversity and fairness are incredibly linked, one cannot be treated as a proxy for the other. After all, simply hiring female candidates cannot ensure that the algorithms they create do not discriminate against women. As Margaret Mitchell explained,

a crucial element of hiring a diverse workforce is making sure that this diversity is spread evenly throughout the company hierarchy. Ms. Marda's point builds on this idea to confirm that, without a broader reckoning with the power structures that permeate organizations, there can be no true reform of a company's unethical AI practices.

Sadly, the awakening of industry players to ideas of diversity has not led to a deeper look into the underlying systems which create it. On the contrary, diverse hiring practices have become somewhat of a double-edged sword in the fairness discussions. To reflect this, in recent years tech companies have pivoted to hiring more minority employees. However, drawing on what Margaret Mitchell says regarding the need for "bottom-up information flow," hiring new employees is not enough to change the existing culture of an organization. In this way, by enacting a surface-level change, companies often end up avoiding calls for broader, more radical, structural changes related to the inherent power disparities in their organizational structure.

> There is a current social consensus on the importance of addressing the issue of fairness in AI. Simultaneously, this may have made it harder to notice our unconscious biases. [...] the growing recognition of the importance of AI ethics is overshadowed by the conventions and unconscious biases of the community itself. *Who* is discussing it [AI ethics] cannot be irrelevant to *what* is being discussed. "Principles to Practices" is one of the central themes of the recent debate on AI. To be aware of such unconscious biases, it is essential to not only set forth principles such as fairness, but also to establish appropriate governance mechanisms to put these principles into practice.
>
> - Arisa Ema

This has an incredible effect on the fairness of algorithms being produced because people who might have more insight into the consequences of a model are being excluded from the conversation. And the people who are defining the rhetoric of fairness are the people that perpetuate such exclusion. As Joanna Shields puts it:

> Today's tech innovators and creators are not necessarily representative of the general population. That lack of diversity profoundly affects how AI and machine learning products are conceived, developed, and implemented. Over the years, we have seen AI replicate historic power imbalances. [For instance], with image recognition services making offensive classifications of minorities and even top-performing facial recognition systems misidentifying people with darker skin.
>
> - Joanna Shields

Ultimately, we can see that the lack of diversity in the design and strategic deployment of AI can hinder the fulfillment of ethical AI premises and its effects.

3. Lack of Ethical Standards

The issues we have discussed begin to compound themselves. The absence of a universal definition of fairness, coupled with the homogeneity of industry leaders, has created a system in which each tech organization gets to decide how AI is built and deployed without virtually any external input. As our experts will explain, the lack of broadly agreed-upon standards of ethics in AI has created a massive power vacuum in which industry players are both defining and enforcing their own ethical norms[10].

10. Shortly after this chapter was written, UNESCO launched its Recommendation on the Ethics of AI, on November 24, 2021, an important first step. For more information, see: https://unesdoc.unesco.org/ark:/48223/pf0000379920.page=14

As the interviewees point out, the norms and standards which do exist are created to serve each business's purposes. Lastly, we will also address how the current culture of the tech industry can exacerbate these issues with its "moving fast and breaking things" approach to AI deployment.

As we presented in our first section, without the establishment of a baseline series of metrics, we cannot evaluate the current fairness of AI models being deployed by the tech industry. Margaret Mitchell highlights this in her discussion of the role which values play in defining the actions of a company. Vidushi Marda also brings up the importance of having an organized form of ethical standards which govern a company's behavior.

> In order to resolve disagreements based on values, it is important to have already defined and agreed on basic values up-front. When a company, organization, team, etc., is created, those creators bring with them implicit values that affect their decisions. Make those values explicit, and update them as people from different backgrounds weigh in. It is these set of values – an organization's "principles" – that help to define what to do.
>
> - Margaret Mitchell

> In the best-case scenario, ethical standards are built along the minimum requirements set by international human rights law (the most universal set of principles we have, that have shared understanding and legal grounding across jurisdictions). Second, they are built and reckoned in tandem with a fundamental reckoning with the institutional, structural, and historical incentives that underlie organizations and companies at the moment.
>
> - Vidushi Marda

However, the establishment of company values is not enough if there is no structure to enforce their importance. As Vidushi Marda explains, although it is common for companies in the tech industry to establish certain guiding principles, without an enforcement structure, they are not held accountable for sticking to them.

> Ethical standards play an important role in situating an organization's public relations. Providing an understanding of what the company/organization would ideally like to achieve, enables stakeholders to identify opportunities that are acceptable or not. However, in its current form, it does not put in place accountability mechanisms, transparency obligations, or address crucial questions of power.
>
> - Vidushi Marda

Ms. Marda goes on to explain that these "guiding principles" as set by each individual company are unlikely to create a code of conduct that can be upheld throughout the broader industry.

> It's important to recognize at the outset that "ethics" and "ethical AI" mean different things to different people within and across organizations. What this translates to in practice is that vague terms are assigned the gravitas of a standard worth striving for, without any shared understanding of the contours of the term itself. Even within organizations and companies, this means that there is little to no coordination on how and when ethical standards are met or flouted.
>
> - Vidushi Marda

Arisa Ema raises the point about the complexity of the AI supply chain and how any governance or regulatory approach must consider the intricate dynamics of different levels of private actors. Her input is very relevant in pointing out that the industry is not homogeneous. Besides the global big tech companies and those that develop products directly to consumers ("B2C"), there are many smaller, local players involved. It is important to consider how governance and regulation impact start-ups, for

instance. Among the smaller players, there are also service providers that connect different companies throughout the supply chain until the technology reaches consumers ("B2B2C"). She cautions about the extensive size of supply chains and the challenges it poses for accountability.

> In general, AI governance refers to the development of principles for ensuring the safety and reliability regarding AI within a company or organization, and the implementation of controls in development and utilization. It is also linked to the concern of companies that their products and services will not be accepted by society, if they do not address the challenges posed by AI. For this reason, the debate on the ethics of AI is now recognized as a problem directly related to management strategy," for companies. For this reason, global companies (in particular) have established individual ethics committees. However, it is often technically and economically difficult for small and medium-sized companies, as well as start-up companies with limited resources, to implement similar governance. In recent years, national and local governments have begun to require data and AI governance as a condition for procuring AI services, which is expected to serve as an incentive for companies to strengthen their AI governance. But for start-ups, the high demands on AI ethics and AI governance are also becoming an economic industry barrier; this is an irony of consequence, given that the idea of AI ethics is underpinned by principles that share a vision of a society that recognizes inclusiveness in diversity. [...] [However] In a long supply chain, the principles of AI development and utilization are not always shared by downstream companies; in the event of an accident or incident at a downstream company, the extent to which the responsibility can be traced back to the upstream company becomes unclear. As such, it is difficult for a single company or organization to address all risks [...].
>
> - Arisa Ema

Reflecting about potential dangers for ethics in the AI industry, Joanna Shields highlights how the concentration of power within the tech industry creates a situation in which the people defining fairness for each organization have no interest in fairness reform. Furthermore, Ulrich Aïvodji cautions that ethical guidelines are often so broad that they have no actionable consequences to those responsible for applying them. This seemingly unintended loophole might actually be strategic for non-compliant entities.

> There is a very real danger that the power to influence the use of AI and its impact could be concentrated in the hands of a few. To a great extent, this is our reality today for ubiquitous products, applications, and services we use every day. [...] I had a front-row seat in the first wave of the digital revolution, an entrepreneurial free-for-all with tech giants living by the arrogant motto of "move fast and break things." Their unshakeable goal was growing by whatever means necessary to dominate these emerging sectors. There was no international framework or blueprint for managing the technology that was advancing in the private sector. As a result, we are now forced to tackle the unintended consequences that threaten our privacy. These biases unfairly classify people and limit their opportunities and the spread of disinformation and illegal content.
>
> - Joanna Shields

> Current "ethical guidelines" are presented as a list of recommendations that companies are "encouraged" to follow. Such an approach presents an evident risk of ethics washing – a communication exercise where dishonest entities will give the false impression that they comply with a particular recommendation, while it may not be the case.
>
> - Ulrich Aïvodji

Together, our experts highlight a conflict of interest in relying solely on self-regulation to advance trustworthy AI. They explain that genuinely enforcing fairness in algorithms is a profoundly complex process which requires both financial and genuine philosophical commitment. Most importantly, they point out that to enforce fairness at this moment in the industry's growth would require a serious departure from the status quo – a move which may not be financially viable.

> If ethical standards are developed with the understanding that they must facilitate faster and frictionless adoption of AI systems, they are clearly not meant to be the accountability mechanism we envision. [...] It is also crucial to understand that those pushing for ethical AI are also the same actors that wield power – i.e., Big Tech, actors that buy into ideas of tech-solutionism. Even well-intentioned ethical initiatives are constrained by organizational and structural realities that value speed over scrutiny, and deployment over deliberation.
>
> - Vidushi Marda

As our interviewees explain, if the only actors defining ethical standards are those who bear inherently profit-driven interests in developing AI, there is little guarantee that ethics, diversity, and inclusion will be seriously addressed. In this sense, there is an urgency in establishing barriers and enforcing accountability if we want to avoid perpetuating discriminatory practices in our AI ecosystem.

> "Ethical AI," if unaccompanied by a critical outlook that is driven by accountability needs, can legitimize and cement problematic and even dangerous uses of AI. This is because if ethical standards are defined, interpreted and certified by the same actors in the absence of legal standards or mechanisms for scrutiny and redressal, there are virtually no checks and balances in place.
>
> - Vidushi Marda

Joanna Shields' experience in the industry corroborates the idea of enlarging ethics discussions beyond each company's walls. She notes how the major players of the industry have historically chosen to ignore fairness reform.

> Over the past decade, I believe that the private sector has not demonstrated an ability to self-regulate in emerging technologies. There have been many inflection points, and the tech sector as a whole needs to step up and do more to proactively address the issues that come with the technology it creates.
>
> - Joanna Shields

Now that we have highlighted three serious holes in the ethical AI fabric, we will discuss the various threads holding it together. Our discussion will lead us through the various other stakeholders, outside of the industry members themselves, who have a stake in algorithmic fairness issues. We will present the interviewees' visions for a broader engagement in AI ethics discussions and ways we can raise awareness and educate society to participate in building a better AI future. Finally, we will briefly explore what this participatory process would look like and what it can achieve.

WHAT CAN BE DONE

The previous sections revealed that allowing the very organizations benefiting from rapid AI deployment to be the sole actors in defining fairness is, ultimately, an unsustainable practice. The shortcomings discussed so far overlap with a more prominent issue that must be addressed: a lack of broader participation in AI governance. By enabling companies to define the status quo, we have created a power

vacuum that allows for serious issues to go unnoticed and unaddressed. The goal of this section is to provide actionable guidance on how to address the issues brought up earlier in this chapter. In doing so, we explore the various stakeholders that are also implicated in the AI narrative and show how they can play a greater role in the discussion about fairness in the AI industry.

Avenue 1: Integrating various stakeholder perspectives

Our first direction for implementing change relates to the creation of a broad regulatory framework. Although in common discourse, regulation is often associated solely with governmental bodies, our experts push back against this idea. Many of them highlight the complexity of regulatory impulses and caution against defining regulation within the scope of a single stakeholder (be it government, external auditing, or otherwise). In fact, in our discussions, we find that the first direction for promoting fairness in the AI pipeline is a collaborative approach between all the players involved in the creation and consumption of AI models. As Joanna Shields states: "AI is not a superpower that will one day democratize benefits for all, and we need to work with governments and the tech industry to ensure the AI we are building benefits all, not the top one percent." By coming together, these players can empower each other to create an informed, robust regulatory framework.

Margaret Mitchell asserts that governmental regulation alone cannot be the fix-all. Rather, we need a collaborative approach that incentivizes members of the tech industry at all levels of the AI ecosystem to participate in holding it accountable. In addition, Rumman Chowdhury voices the importance of understanding what it means to audit technology, before jumping head-first into regulating it.

> [Tech companies] have demonstrated that they're not capable of doing the bare minimum needed in this space. [...] [But] I don't think regulation is a silver bullet. [However] with top-down, higher-level goals put forward by regulation (governments) [we can] start to incentivize company behaviors around [fairness].
>
> - Margaret Mitchell

> We're not even in a place where we can even agree what an audit looks like and that, to me, is very, very worrisome, given how the regulatory world seems to be very excited to be doing regulation without having done the basics.
>
> - Rumman Chowdhury

It is clear that no stakeholder (be it a government or otherwise) should be exclusively responsible for defining fairness, crafting regulations and enforcing it on all other actors entangled with AI pipelines. We will now present excerpts from our discussions with our experts, which explore the roles and collaborations between various stakeholders, each with a unique role in empowering change within the AI industry.

State actors

Vidushi Marda advocates for a broad, governance-focused approach in dealing with the challenges the industry faces regarding fairness. She defines a framework in which different stakeholders can create a network of interests. On a similar note, Joanna Shields presents the complexities of integrating governmental bodies of a fragmented landscape into a common regulatory framework. Nevertheless, she argues that an international agreement is one of the key elements of creating a culture of ethical AI.

> In an ideal situation, we have what I like to call the "room of AI governance." The floor represents international human rights standards as the minimum requirement below which AI systems cannot go. The ceiling represents ethical standards that indicate where AI systems should go, and provide

> an idea of what AI systems should strive for. The walls represent various regulatory and policy levers that decide the extent and contours within which AI systems should function, and it is technical standards and specifications that facilitate how we get to any of these parts of the room.
>
> - Vidushi Marda

> The significant challenge for ethical AI lies in the fragmented approach to AI worldwide and the uneven pace at which legislation is emerging. Major powers are already demonstrating divergence in their approaches to regulation, and a global AI race is underway. Ensuring the development of beneficial, trustworthy, and robust AI requires collaboration between like-minded, democratic nations and a set of international standards that holds each government to account. These standards must carefully balance the need for AI to be developed in accordance with human rights and fundamental values whilst not stifling innovation. [...] We need nations worldwide to come together and build a global framework and international standards for how this technology is used. There needs to be an inter-governmental body that ensures consequences for businesses and governments who misuse AI and deploy it to repress people's human privacy, dignity, freedom, and rights.
>
> - Joanna Shields

Arisa Ema also brings up the importance of uniformity in ethical standards. She, too, comments on the profound importance of creating a cohesive approach to fairness that transcends geographic localities.

> To implement an ethical approach to AI within an organization, it is essential to establish an appropriate governance structure. However, due to the differences in policies, values, and industrial structures in each country and region, the state of governance is not uniformly determined. While respect for diverse values is important, overly fragmented governance frameworks not only hinder innovation, but may also lead to regulatory arbitrage.
>
> - Arisa Ema

Ulrich Aïvodji brings up a crucial point about the power disparities in geopolitics, which affects discussions surrounding the deployment of AI systems. Drawing from his knowledge of the African continent's technological development, he highlights the importance of giving countries that have not yet made it to the "AI race" (see: Savage, 2020) a powerful voice in fostering a more diverse and context-appropriate AI landscape, to mitigate an increasing dependency on foreign technologies. This would allow for the emergence of a plurality of perspectives at the international level and foster an independent AI development in the Global South regions.

> If a government does not invest on local initiatives to foster the development of AI systems that are adapted to local needs, technologies developed with Western perspectives and interests will be imported and further increase the dependence of the country on Western technologies.
>
> - Ulrich Aïvodji

Rumman Chowdhury also notes that harmonizing perspectives should not lead to a hegemonic uptake of a particular vision for AI fairness. In her discussion of regulatory measures, she brings up one of the shortcomings of unifying regulations worldwide.

> I do understand that a frictionless regulatory environment makes things easier from an application logistic perspective. I think, though, that you run a risk of ignoring diversity of choice. So when we were talking about that ideal state, I don't know how one creates one set of laws to govern all of

> algorithmic ethics and algorithmic use. That would also enable fairness to the degree that an individual understands fairness and then also gives us the choice and agency. And that's never existed in the history of the world, essentially a unified government agreement on anything.
>
> - Rumman Chowdhury

External auditors

Another key player in the building of a regulatory framework are external auditors. Joanna Shields raises the importance of independent audits conducted on large tech corporations. She explains that leaving experts without any form of redress against companies can create a dangerous situation where their warnings are simply ignored.

> In the recent case of Dr. Timnit Gebru and her dismissal from Google, the ethics of conducting research with big technology companies has been called into question. It has made a case for independent, publicly funded research into AI and its potential harms, alongside robust legislation such as what the EU has demonstrated recently, ensuring this tech is developed responsibly.
>
> - Joanna Shields

Domain experts

Domain experts are a stakeholder group that is closely tied to auditors. They have a crucial role in situating AI models and shaping their application to specific areas. As mentioned earlier in this chapter, there is a common understanding that the context-based character of AI application requires a closer collaboration with professionals working on the ground. Margaret Mitchell corroborates the importance of domain experts in the discussion surrounding AI deployment.

> For all of these different technologies, those with the relevant expertise (medical doctors, climatologists, etc.) and those who are affected (people with different ability statuses, people who live in more isolated areas, people who might be displaced by the technology, etc.) should be part of shaping what the technology does, how it will be used, and whether it should exist at all.
>
> - Margaret Mitchell

Ms. Mitchell also suggests an opportunity for broadening multi-stakeholder participation by proposing the strengthening of relationships between ethics experts and regulators. She also presents space for another stakeholder in this regulatory collaboration. She presents the potential for having external auditors play the role of enforcing the government's directives.

> There's a role to play here, both for companies and organizations and for regulation [...] I think governments have a role to play in defining the high-level goals of what they want from systems – transparency, robustness, whatever – these kinds of high-level [goals]. And tech systems or people developing tech systems can then provide evidence of that based on what they understand about their systems. It's a top-down meeting bottom-up [approach]. The regulatory party says "We want to see this and this" and the company presents the various metrics they think are appropriate. [...] So it's a little bit of self-regulation meeting external regulation.
>
> Currently, you don't have experts within a company. You have people with power in a company. So if you have cutting-edge researchers doing work on fairness, then those are the right experts to talk to regulators about fairness. Actually matching expertise with what government officials

> need is very key. [...] Then this idea of independent auditing where you have researchers saying to the auditor "Here are the various issues here, [this is] what we think we're seeing." And then the auditor can handle it. And that also allows a certain amount of privacy.
>
> - Margaret Mitchell

Civil society

Lastly, the largest group of stakeholders is civil society. This category encompasses everyone, either as users of technology or as individuals that might come in contact with AI applications deployed by third parties in different situations, with or without their knowledge. Francisco Marmolejo-Cossío brings up the importance of integrating users into the regulatory process, but also of being mindful of how users' preferences regarding their interaction with technology are context-based and dependent on awareness-raising and empowerment.

> I think putting some of the onus on users to try and bring about some of this change in industry will fundamentally be different given these different preferences across the population. [...] Maybe this is something where external organizations can come in and put something on a level playing field. So rather than, for example, completely imposing some external set of metrics or external conversation, providing enough incentives for those segments of the population that aren't necessarily involved in the conversation on privacy and on fairness to be involved in the conversation. That might give some impetus to this customer-based approach.
>
> - Francisco Marmolejo-Cossío

He highlights the roles that civil society groups can have in empowering users to level the playing field when it comes to AI ethics awareness. Such roles can also extend to whole segments of the population subjected to the deployment of AI in different spheres, with other actors from the private or the public sector following the lead of a user empowerment approach.

Ultimately, it is noteworthy that throughout our conversations, the experts highlighted that regulation is not a matter of outsourcing the task of expecting governmental bodies to shape the AI landscape, but rather of building a set of common norms and standards to be adopted by the technology industry at large. This consensus would help enforce an ethical alignment in industry practices and inhibit potentially harmful uses of AI.

Avenue 2: Raising awareness through education

An underlying notion of implementing broader participation is the requirement of a level playing field where all parties can engage in meaningful and constructive conversations. Many interviewees mentioned the importance of raising awareness, providing education, and promoting people's autonomy to make informed choices about their interactions with AI technologies. Given the often shrouded and isolated nature of the technology sector, it is easy for non-technical thinkers to be excluded from conversations surrounding the creation and deployment of AI systems. In fact, it is only in recent years that discussions of ethical AI have penetrated the common discourse. Without an audience that can participate in discussions related to AI, there can be no reform. Elevating discussions around fairness requires a critical outlook from society, which can only be fostered through providing access to fairness education on a broad scale. This education is necessary at all levels of our society, both technical and non-technical. As Arisa Ema points out, understanding the role that biases play in shaping technological development is necessary for both those who consume technology and those who create it.

> Technology does not develop on its own. Technological development needs a purpose, and it is influenced by the needs of society and the visions that people have for it. [...] Concerns about bias in AI algorithms and data are now a globally shared problem. One of the reasons for these biases is that our society itself is biased to begin with. It is difficult for us to be aware of the bias in our society.
>
> - Arisa Ema

Without awareness of the pitfalls in technological development and the biases that shape it, there are no avenues for reassessing our shared visions, let alone for formulating a proper regulatory framework. Unfortunately, it is difficult to bridge the gap between those who are embedded in the technical field and those who aren't. As Francisco Marmolejo-Cossío points out, there is a profound need for the creation of a dialogue between those who are affected by AI and those who create it.

> [We need] an awareness around lack of fairness and more AI informational workshops or just even spreading the word. This would allow us to have an audience or an impact [with] some of these communities that are specifically affected by issues of fairness... [Those who] might not be necessarily aware of the fact that there are discrepancies in outcomes of opportunities, given the mechanisms that are in place behind the scenes. It's taking [an] awareness approach... [taking time to] look at segments of the population that suffer from unfair outcomes and have conversations with them.
>
> - Francisco Marmolejo-Cossío

Mr. Marmolejo-Cossío further explains the importance of raising awareness among society about the unfair outcomes of the AI models they interact with. He goes on to highlight the importance of educating non-technical audiences about the process of automation and how it can affect their lives.

> [We can] offer other forms of education of this in the educational sphere. As we go forward and we think about how we change primary, secondary and high school, this could definitely be something fundamental in the education sphere. Thinking about automation, thinking about the impacts of automation, the ethics behind all of this, this should be something that as it becomes ever-present [...]
>
> A great scenario going forward [would be] having just a platform at the public national education level for [a conversation with pedagogical policymakers, but also technical individuals such as you and other people that work in this space], because this is just going to become ever more present in the following years.
>
> - Francisco Marmolejo-Cossío

When asked about how such an educational program should be defined, he expresses the importance of an interdisciplinary approach to explaining algorithmic techniques.

> [Building such a curriculum] is a very interdisciplinary practice. It's not just the technical from the STEM perspective, it's not just the algorithmic techniques, the optimization techniques we bring in, but it's also the societal context that goes into the specific features or input that we have. Ultimately, there is a decision-making process that is facilitated in part by the techniques that are available in the system. And that part can also be a part of the conversation. Like what choices went into creating the data set and thinking about what this brings, what benefits, what discrepancies? And there, I think there's much that we can do [with] the technical [part] from this pedagogical perspective.
>
> - Francisco Marmolejo-Cossío

Education is a crucial element to empowering broad participation. Mr. Marmolejo-Cossío provides an insightful account of how education can create a conversation with broad community involvement and empower individuals to engage with some of the larger issues of transparency and distrust in the AI sphere.

> Ultimately, trust is always an issue with these things. Do you trust the individuals, the committee, the power that you're putting into the people that might be creating something that, if implemented, is going to be reaching millions of people? If we just focus on the primary and secondary school, for example, some conversation around this at a national policy level, then it's a big policy to be made with big implications for poor perceptions or for potential lack of perception.
>
> - Francisco Marmolejo-Cossío

Ulrich Aïvodji, too, comments on the importance of also raising awareness, focusing on another important pursuit for an AI-knowledgeable society: educating society about the negative underside of ethical AI discussions.

> To minimize the risks of ethics washing, the least that can be done is to raise the awareness of the different ways in which ethical recommendations can be easily evaded by well-motivated entities. [...] Africa and the African diaspora already have a lot of researchers who are actively working to raise awareness on the harms that automated decision-making systems might cause. It is important to pay attention to their work and not only listen to tech evangelists with over-optimistic views about AI.
>
> - Ulrich Aïvodji

Ultimately, without education about the importance and dangers of unregulated unethical AI practices, no one but the leaders in the AI industry can speak up against harmful practices. Furthermore, without education on the matter, it's easy to get caught up in shallow displays of fairness practices that are meant to conceal more sinister underlying structural problems.

Avenue 3: Elevating data-related discussions and activating reforms

The growing discussion around data gathering in the public sphere is a great example of stakeholders engaging with the consequences of AI usage and fighting for autonomy over the way AI is deployed in their lives. After scandals such as Cambridge Analytica, non-technical users became aware of the problematic ways in which their data was being collected (Confessore, 2018). This acknowledgement of AI's impact catalyzed a whole broader movement of action, empowering people who might not otherwise have voiced their opinions in the AI space (Garret, 2018). Once we have an informed public, every day members of our society can participate in crafting the critical elements that form the backbone of any AI system: data. Ownership of one's data means being able to make informed decisions about which of their data are used and how, as Rumman Chowdhury argues.

> There's a bit of a social contract when it comes to AI. [...] "I understand that I'm giving up certain pieces of data and information, and, in return, you utilize that to make some sort of a prediction, whether it's improving your models with respect to improving your product as it relates to me". [...] So the way meaningful user agency can help impact algorithmic fairness [is that it] allows both groups [users and companies] to have an understanding and appreciation of the social contract as it's written and not [allow companies] to exploit it. Because frankly, a lot of the misuse of algorithms comes from negative externalities, those that go above and beyond what the average user would probably agree to using their data for.

> I think the best-case scenario is a world in which people have ownership of their experience and they have the right to do things like opt in and opt out. They have the right to do things like share their data, not share their data. They have a right to benefit when they want to benefit and be left alone when they want to be left alone. I do think that sometimes the policy goes a little extreme, so either fully participate or completely opt out. And I think an ideal world is one in which we create a spectrum of engagement where individuals get to choose how engaged they want to be in an algorithmic society.
>
> - Rumman Chowdhury

Francisco Marmolejo-Cossío shares a similar vision about autonomy over one's data. He mentions sustainability as a central aspect of this approach. Empowering citizens holds the potential of driving change in the AI ecosystem and contributing to a better environment for users and for companies. Treating users with respect and advancing information, education and transparency seem vital for a data-driven ecosystem.

> Some best-case scenario would be awareness in this setting and having this actually be a substantial feature that the consumers end up actually putting weight on in their consumer decisions. And I don't know how to necessarily push for something like this through policy, through education, through building awareness. But I think that would go a long way and would potentially be more sustainable. And in a certain sense, this whole kind of response. So, some transparent method whereby users can be aware of potential pitfalls of the systems.
>
> - Francisco Marmolejo-Cossío

CONCLUSION

In writing this chapter, we specifically selected three dominant issues surrounding the AI industry that we and the experts we spoke to believe that, if left unaddressed, will lead to profound and dangerous perpetration of discriminatory practices in all spheres of our society. We provided not only descriptive accounts of those issues, but also gathered the interviewees' main suggestions of strategies that can drive change in an opposite, positive direction.

First, we touched upon the challenges of pinning down definitions of fairness. Frustrated attempts to capture complex values in a mathematical notation are only one aspect of this issue. The concept of fairness predates the emergence of AI and is interpreted differently depending on domains, cultures and other contextual factors. The discussions with our experts revealed how far the industry is from applying fairness methods in its evaluation of models and that a lack of an appropriate fairness definition can have devastating effects in the deployment of AI models.

Second, we touched upon cultural problems in the industry, highlighting its consistent lack of diversity in hiring practices and in the structural organization of roles and functions throughout a product's development. Be it diversity of gender, multidisciplinarity, or other, addressing this issue requires a cultural change in a higher organizational level. The interviewees presented insightful accounts of how diversity can have a direct impact on the fairness of AI models and delved into its causes. It is noteworthy how effective diversity can unlock great potentials for the AI industry in a way that societal benefits are maximized.

Third, as talks of regulation permeate the field of AI fairness, we instigated a discussion with the experts about the industry's practices in ensuring ethical AI and the prospects of broader regulations. In almost each of our interviews, we identified concerns over the industry's inability to self-regulate. Another common ground among the experts is a critical look about the potentials and limits of regulation. Although regulation is a necessary venue for a fairer AI ecosystem, it cannot be treated simply as a silver bullet.

Finally, in addressing ways forward, we highlighted that, without broader participation in the creation of a regulatory framework, the current issues of AI will continue to fester, unhindered. Building upon this topic, we presented the stakeholder groups that should be empowered to collaborate on advancing fairness in AI: governments, auditors, experts, civil society.

Delving into how to go about the much-needed structural changes in the field of AI, we brought up the experts' accounts on how awareness-raising and education can level the playing field for non-technical stakeholders to partake in a more democratic decision-making process. Lastly, we highlighted data ownership and data gathering as examples of important topics for elevating discussions among stakeholders and activating a reform in current AI practices. Drawing from the interviewees' inputs, we argue that reassessing current practices involving data can be a first step towards a multi-stakeholder-led change for a more ethical approach to AI design and deployment.

Ultimately, identifying and combatting discrimination will always be an incredibly difficult task, both within and outside of the AI space. That is why the responsibility of ethical AI should be shared between all those who oversee AI development and those who engage with it. As Arisa Ema puts it: "No matter how equity-conscious the vision of AI development is, if it is not accompanied by action, it will fall flat." On the same note, Joanna Shields expresses our shared sense of urgency in addressing this issue: "Now is the time to implement frameworks that ensure technology is developed in accordance with our fundamental rights and prevent the unintended consequences of AI from damaging lives and the fabric of our society."

REFERENCES

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. 2016. Machine bias. *ProPublica*. May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Cofessore, N. 2018. Cambridge Analytica and Facebook: The scandal and the fallout so far. *New York Times*. April 4. https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. October 10. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Dietrich, W., Mendoza, C. and Brennan, T. 2018. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Performance of the COMPAS Risk Scales in Broward County*. Northpointe Inc. Research Department. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Garrett, G. 2018. The politics of data privacy in a post-Cambridge Analytica world. *Wharton Magazine*. May 8. https://magazine.wharton.upenn.edu/digital/the-politics-of-data-privacy-in-a-post-cambridge-analytica-world/

Grind, K., Schechner, S., McMillan, R. and West, J. 2019. How Google interferes with its search algorithms and changes your results. *Wall Street Journal.* November 15. https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753

Larson, J., Mattu, S., Kirchener, L. and Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica*. May 23. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Savage, N. 2020. The race to the top among world's leaders in artificial intelligence. *Nature*. December 9. https://www.nature.com/articles/d41586-020-03409-8

Smith, B and Browne, C. A. 2019. *Tools and Weapons: The Promise and the Peril of the Digital Age*. New York: Penguin Press.

Spielkamp, M. 2017. Inspecting algorithms for bias. *MIT Technology Review.* June 12. https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/

Skeem, J. and Lowenkamp, C. 2016. Risk, race, & recidivism: Predictive bias and disparate impact. *SSRN*. http://dx.doi.org/10.2139/ssrn.2687339

Telford, T. 2019. Apple Card algorithm sparks gender bias allegations against Goldman Sachs. *Washington Post*. November 11. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/

Verma, S., and Rubin, J. 2018. Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7. https://fairware.cs.umass.edu/papers/Verma.pdf

Yong, E. 2018. A popular algorithm is no better at predicting crimes than random people. *The Atlantic.* January 17. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

THE ATTENTION SKEW IN AI DEVELOPMENT: THREATS AND CORRECTIVE MEASURES

ADJI BOUSSO DIENG

Senegalese computer scientist and statistician with a PhD from Columbia University. She is an Assistant Professor of Computer Science at Princeton University, Research Scientist at Google AI, and the founder and President of the nonprofit The Africa I Know. Her lab works on devising AI methods for science and healthcare applications. She is funded by NSF and the Schmidt DataX Project.

SDG5 - Gender Equality
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

THE ATTENTION SKEW IN AI DEVELOPMENT: THREATS AND CORRECTIVE MEASURES

ABSTRACT

The developments in artificial intelligence (AI), both in industry and in academia, that currently dominate the narrative about the field are mainly centered around the goal of building AI systems that can outperform humans on some tasks. A majority of researchers and engineers, empowered by the media and funding sources, are investing all their energy and focus in the pursuit of artificial general intelligence (AGI) and the study of the harms that stem from advances made towards that pursuit. The attention skew in the field of AI causes real societal and socio-economic threats and prevents us from leveraging AI to solve many pressing problems facing humanity—problems whose solutions don't require the development of agents with superhuman intelligence in the first place. This chapter argues that it is possible for AI to be a technology that pushes humanity forward in a positive direction, but only if we shift attention away from the shiny AGI goal and embrace the philosophy and practices of a smaller community within the field of AI—a community that gives agency to humans and is concerned with accounting for human knowledge, desiderata, uncertainty and controllability in the development of AI systems.

THE AI FIELD SUFFERS FROM AN ATTENTION SKEW

Let's take a bird's eye view of the AI field as it is today. The majority of the developments in the field that we hear about in the media are driven by two different communities with different sets of goals. One community is focused on achieving artificial general intelligence (AGI) with the goal of developing AI systems that are as intelligent as humans, if not more intelligent. We find the drivers of this AGI agenda mainly in industry but also in academia. The second community is focused on highlighting the harms caused by advances made by the first community and studying the implications of the realization of AGI. Ultimately, the work of both of these communities centers around AGI. One develops tools towards achieving AGI while the other acts as a guard against the harms and potential future repercussions of AGI. We call this narrow focus in the field the AI attention skew.

To understand the AI attention skew problem, we need to identify the different stakeholders enabling it and what their incentives are. The dominant enablers of the AGI vision of AI are found mainly in industry, with companies such as Meta (formerly Facebook), DeepMind, Google AI and OpenAI at the driving seat. AI provides a unique opportunity for growth for these companies by enabling them to attract the scarce AI talent in the field, improve their existing products or embark on new ventures. For example, in 2019 Microsoft invested US$1 billion in a multiyear partnership with OpenAI for it to "deliver on the promise of artificial general intelligence" (Microsoft, 2019). Since then, OpenAI has developed GPT-3, a multi-purpose AI that can process text from various sources for the purpose of answering human-generated prompts. GPT-3 is powering several Microsoft products, including Azure—Microsoft's lucrative cloud computing service—and GitHub, a major software development platform acquired by Microsoft in 2018 (Langston, 2021a, 2021b; Nadella, 2018; Newman, 2021). In 2017, Google AI researchers developed the Transformer, an artificial neural network architecture that is at the core of several AI technologies, including GPT-3 (Uszkoreit, 2017). Transformers are now at the core of Google's search and translation engines (Nayak, 2019; Raghavan, 2020; Caswell and Liang, 2020). Meta uses AI on its different social network platforms for content recommendation, automatic photo tagging, multilingual translation, content moderation and more. Recently, Meta announced a new project that'll leverage the videos of users in its platforms to train AI that can aid in the development of several new products (Zweig *et al*., 2021). DeepMind, on the other hand, has mostly focused on building AI that can beat human champions in games such as Go and Chess (Gibbs, 2017; Hutson, 2017), with one of the core missions of the company being to "solve intelligence" (DeepMind, n.d.). The company has since put more focus towards advancing science through AI with the development of AlphaFold (AlphaFold team, 2020; Jumper *et al*., 2021), an AI that has made strides in tackling the decades-long protein folding challenge in biology. Building from the success of AlphaFold, DeepMind's CEO recently launched a for-profit spin-off company to leverage the technology for drug discovery (Khan, 2021).

These business incentives are driving the development of AI, defining what problems AI research should focus on solving and what advances should be valued. Evidence of this is the dominance of these tech companies in terms of the number of papers published in the major AI publication venues, such as Neural Information Processing Systems (NeurIPS) and International Conference on Machine Learning (ICML) (Rakicevic, 2021; Nguyen, 2021).

Although major tech companies have a role to play in AI's attention skew problem, they share responsibility with other enablers, e.g., the media. Both the positive and negative consequences of advances made towards the pursuit of AGI get talked about profusely in the media, causing both excitement and fear towards the technology as well as confusion about its capabilities and goals (Jordan, 2018). Media highlights of AI beating human world champions at games such as chess and Go (Gibbs, 2017; Hutson, 2017) and other sensationalized reports on technologies' capabilities have caused some people to wonder whether AI will replace them in the workplace. Others have started worrying about a dystopian future where robots will take over the world and kill living beings on the planet. Besides media coverage, another manifestation of fear towards AI's impact on society is the emergence of a rapidly growing field of AI ethics. The increasing discussions about governments developing a legal framework for AI (European Commission, n.d.; Candelon *et al*., 2021), the funding of research on mitigating the potential harms caused by AI by nonprofits such as the Open Philanthropy (Beckstead and Muehlhauser, n.d.), and the proliferation of AI ethics centers and academic institutes in several universities highlight the field's recent expansion. These efforts are needed to mitigate potential harms caused by advances towards AGI. However, they constitute fuel for AI's attention skew as they center the narrative around AGI.

As mentioned in the beginning of the chapter, AI's attention skew is a problem. However, it also draws a lot of enthusiasm towards all the opportunities that AI offers. Countries such as Egypt, Brazil, Canada and the U.K. have been working on their national AI strategies, signaling a strong belief in AI's potential to spark significant growth in both the public and private sectors (Invest in Canada, n.d.; GOV.UK, n.d.). Government funding agencies such as the National Science Foundation in the United States are

investing significant amounts of money in academics pursuing research advancing AI in different domains (National Science Foundation, n.d.). AI-related majors and subjects are in increasing demand amongst students, both at the undergraduate and the graduate level (Artificial Intelligence Index Report, 2021). AI-related jobs are in high demand and pay relatively high salaries (Chung, 2017). Employers are increasingly seeking proficiency in AI-related topics (Columbus, 2019). Finally, AI-based startups are proliferating and garnering significant funding from venture capitalists and other investors (Weiss, 2021; Wilhelm and Heim, 2021).

WHY AI'S ATTENTION SKEW SHOULD BE CAUSE FOR CONCERN

The pursuit of AGI has fostered a research culture in which methodological and empirical innovations—and the development of the theory underpinning those innovations—are centered around solving human-like tasks. For instance, writing coherent text from scratch, summarizing documents, carrying on conversations, recognizing faces, answering questions, describing images, playing games and so on. AGI pursuit has led to several advancements in AI, mainly in computer vision, natural language processing, recommendation systems and the intersection of those domains.

Indeed, language technologies are seeing significant improvements in their ability to produce text and speech that are indistinguishable from human text and speech for the purpose of conversation, translation, summarization and more. One example is OpenAI's development of GPT-3—a piece of AI software that can write coherent text from scratch, among other things (Pilipiszyn, 2021). Another example is the development of AI-powered recommender systems, which are used across social media platforms and other internet services and dictate what types of content we consume online. Furthermore, voice assistants are interacting with us in our homes and through our mobile devices, answering our questions ranging from weather to culture or general knowledge at large.

Nevertheless, these advances also bear shortcomings: they often negatively impact marginalized communities. As mentioned, this has led to the emergence of a rapidly growing field of ethical AI that has garnered the attention of researchers, academics, activists, governments, nonprofit organizations and civil society, whose goal is to thwart the threats posed by AI advances and develop legal regulatory frameworks for the development of the technology (Beckstead and Muehlhauser, n.d.; European Commission, n.d.; Candelon *et al*., 2021).

Despite the benefits they may bring, all of the relatively recent AI advances in computer vision and natural language processing pose ethical threats and drain resources, both human and economic. Those resources could go towards developing AI that can help alleviate important pressing problems facing humanity, e.g., climate and healthcare crises. By taking a close look at how new AI is developed in the research community, we can better understand the attention skew—or the misplaced focus—in the field of AI and its several societal and socio-economic consequences.

The new AI breakthroughs we often hear about in the media, such as AI producing images of people and objects that are indistinguishable from reality or writing entire stories when given a prompt (Karras *et al*., 2019; GPT-3, 2020), are enabled by methods that follow the same pipeline, which we call *task modeling*. Task modeling has four steps. The first step is task specification, where we decide the task we want to teach the AI system to perform. The second step is data collection, where large amounts of data that are relevant to the task are collected—by scraping the internet, for example. The third step is system development, which involves using modeling tools such as artificial neural networks to process and represent the data and devising an algorithm to adapt those tools to the task at hand using the data. The models used in this step are often quite complex; they have several degrees of freedom (called *parameters*) and require significant computer resources. Finally, the fourth and last step in task

modeling is to evaluate the system resulting from the previous step on the task specified in the first step. This evaluation often consists in paying humans to assess the system as it pertains to solving the specified task—for instance, by using Amazon Mechanical Turk or submitting the system for performance assessment against a leaderboard of benchmark metrics. Each step in the task modeling pipeline poses potential threats, discussed in the following sections.

Issues with task specification

The current dominant paradigm of building AI systems that can accomplish certain tasks prevents us from thinking about the ethical considerations pertaining to specifying a task. Not all tasks should be performed; some are obviously harmful and can marginalize certain communities and negatively impact humanity in general. There are several examples of publications in AI, often in prestigious venues such as the journal *Nature* or the premier AI conference Neural Information Processing Systems (NeurIPS), whose propositions should be cause for concern. A paper published in *Nature Communications* in 2020 proposes a method to track historical changes in trustworthiness using facial cues (Safra *et al.*, 2020). Another paper published in NeurIPS in 2019 proposes to reconstruct a person's face based only on a recording of their voice (Wen *et al.*, 2019). There are many similar instances of papers in high-impact AI venues aiming to predict a person's identity or character based on biological features, such as voice or facial features. These types of methods power systems that are used in the real world: notably, image-based internet applications or policing tools, whose negative impacts on marginalized communities are extensively documented (Ryan-Mosley, 2021; General and Sarlin, 2021; Galston, 2020; Dunn *et al.*, 2020). On January 9, 2020, Robert Williams, a 42-year-old Black man living in Farmington Hills in Michigan, was arrested and detained for 30 hours after he was wrongfully identified by facial recognition software as the suspect of a crime he didn't commit (Ryan-Mosley, 2021). His case isn't an isolated one. Wrongful arrests were reported before Mr. Williams; for instance, in 2019, Nijeer Parks, a 31-year-old Black man living in New Jersey, was arrested after being falsely identified by facial recognition. He spent 11 days in jail before he was finally released (General and Sarlin, 2021). There are still multiple instances of such wrongful arrests caused by police forces' use of facial recognition. These incidents have sparked an outcry and led to the launch of several campaigns from activists and civil rights organizations calling for the ban of facial recognition use by the police (Allyn, 2020; Snow, 2018) as well as lawsuits from the victims (Harwell, 2021).

Other examples of tasks that should not be performed by AI systems are those whose potential harms outweigh any benefits they may have on a given industry. Deepfakes—fake images or videos of people that are indistinguishable from reality for the viewers—have a lot to offer to the video production industry, and the film industry at large, by making it easy to edit in and out certain aspects of a video, such as the speaker's voice, tone or accent. However, deepfakes pose a real threat to democracy everywhere as they can amplify misinformation online. They also enable abuse online and have been at the core of the discussion around online sex trafficking and gender-based violence (Galston, 2020; Dunn *et al.*, 2020).

Finally, one of the limitations of a task-focused approach to AI is that it prevents us from leveraging AI to solve important problems that cannot be boiled down to performing a task. Problems that require an understanding of certain mechanisms, such as the ones found in the sciences and in healthcare, and problems requiring an understanding of cause and effect aren't always amenable to a task-based framework. A task-focused approach to AI, therefore, constitutes a missed opportunity in leveraging AI for critical domains such as healthcare.

THE BURDEN OF RELIANCE ON BIG DATA

Building AI systems that can perform certain tasks often requires collecting several terabytes of data. There are several drawbacks to this reliance on large datasets. First, we must consider the many issues that arise from the data collection process itself.

Often AI researchers and engineers scrape the internet, gathering data from different sources to train AI systems. Sometimes this is done without regard to users' privacy and consent. The facial recognition company Clearview AI, founded in 2017, allegedly scrapes billions of photos of users of social media networks and the internet without their consent. The photos are used to train their facial recognition system, with law enforcement agencies among its customer base (Hill, 2020). There are many other such databases containing photos of internet users who aren't aware of the breach of their privacy (McQuaid, 2021). More recently, the Microsoft subsidiary GitHub and OpenAI partnered to develop and release GitHub Copilot, an AI-based code completion tool that helps coders write code more easily. The system has received great reviews on how accurate it is but has also raised eyebrows concerning copyright issues (Taft, 2021). Indeed, the training data of GitHub Copilot is based on publicly released code on GitHub, some of which have licenses that do not allow derivative works (Taft, 2021).

Another problem of scraping the internet in search of data is that we miss paying attention to the quality of the data collected. Often these data encode biases in the form of negative stereotypes towards certain communities, or harmful speech, such as towards women and Black people. For example, the state-of-the-art natural language AI system GPT-3 was originally trained using five different data sources, totaling almost 500 billion words. However, the data sources contain toxic language that is amplified by the model. For example, GPT-3 produces harmful speech towards Muslims and encodes gender and race biases (Samuel, 2021; Lucy and Bamman, 2021).

Furthermore, indiscriminate data collection raises the problem of the quality of these data. More specifically, data quality is compromised because the data are often not representative of everyone, and they often don't contain information about certain communities. For example, datasets used to train AI systems tend to be skewed towards white men and Western society. Entire cultures aren't currently represented on the internet—for example, many on the African continent. This is a problem that multiple data sources face.

In the domain of natural language processing (NLP) in particular, the lack of data representativity is very severe. Although AI has advanced language technologies, these advances are mainly in English and other languages that are significantly represented on the internet, e.g., Mandarin, German and French. Languages from the African continent aren't represented in the data used to train AI-based language systems, which excludes an entire continent from leveraging the benefits of advances in AI for language understanding.

Finally, building systems that require a large amount of data raises the problem of the huge cost associated with data collection. First, there is a monetary cost that comes with storage: one needs computer resources to store the data. Second, the data often need to be labeled, so researchers have to resort to tools such as Amazon Mechanical Turk for data labeling. Beyond the monetary cost, requiring human labor to label data constitutes a limitation for domains that require expertise, e.g., science and healthcare. Labeling images or text is easy and can be outsourced to a large pool of professionals. However, labeling molecules or X-rays, for example, requires domain expertise that few possess. Centering the development of AI on these data collection practices can therefore be limiting and exclusionary.

THE HARMS AND COSTS OF LARGE MODELS

Building AI systems that can perform certain tasks often requires models with high levels of complexity. One paradigm that enables the development of such complex models is *deep learning,* which leverages artificial neural networks—layers of computations that mimic neurons in the brain—to specify flexible models that can be trained on data to accomplish tasks. As the task at hand gets more and more complicated, more complexity needs to be added to the model. This leads to very large models whose decisions and behaviors cannot be understood and, more importantly, cannot be controlled. There are numerous examples of failures of these large models and the negative impact they can have on society (Bender *et al*., 2021).

For example, the computer vision systems powering technologies such as self-driving cars encode and amplify harmful human biases, especially as it pertains to race. Indeed, computer vision systems often fail to correctly recognize Black people. Recently, Facebook showed a video prompt that asked its users who just watched a video featuring Black men and published by the British tabloid *The Daily Mail* whether they would like to watch other videos of "primates" when there were no primates in the video (Mac, 2021). Another illustrative example of this is when, in 2015, the Google Photos app mislabelled several photos of two African-Americans as showing gorillas (Zhang, 2015). There are many other such examples where computer vision systems fail miserably when it comes to Black people. A self-driving car using a recognition system that fails to identify certain people may cause deadly accidents for those people it fails to identify. These are often people in marginalized communities who are not represented in the data the recognition system was trained on.

Computer vision isn't the only domain where large models fail. Natural language technologies based on AI have also shown limitations and continue to raise concerns. On March 23, 2016, Microsoft released a chatbot—a piece of software that can conduct an online conversation with a human via text or speech—called Tay, whose Twitter account started posting inflammatory and harmful tweets right after it was launched, which led to Microsoft taking it down only 16 hours after its release to the public (Hunt, 2016). Yet another example is when Facebook's AI-powered translation system wrongly translated the post of one Palestinian user as a call for violence when in fact the original post was a simple "good morning." This mistranslation led to the wrongful arrest of the user by Israeli forces who were alerted after the post was made (Ong, 2017).

Yet other unexpected outcomes of AI systems can be found in voice assistants. Although ubiquitous, voice assistants have a hard time recognizing certain accents. I have a personal account of this problem: Apple's phone assistant Siri often fails to recognize my Wolof accent. This is an illustration of AI advances benefitting only certain people and not everyone. Finally, it's been documented that GPT-3 has a negative bias against Muslims. When prompted to describe Muslims, GPT-3 spews harmful speech that amplifies the negative stereotypes equating Islam with violence (Samuel, 2021). GPT-3 also has race and gender-related negative biases. A paper in the premier NLP conference Association of Computing Linguistics (ACL) published in 2021 found that stories generated using GPT-3 encode and amplify social biases pertaining to race and gender (Lucy and Bamman, 2021). These harmful biases of GPT-3 should be cause for concern, especially given the numerous applications of the technology. A blog post from OpenAI reports that GPT-3 is powering more than 300 applications across several industries, including education, search, conversation, text completion and more (Pilipiszyn, 2021).

One more subtle way in which large neural network-based models negatively impact marginalized communities is that they often memorize the rare training samples in their training data (Feldman and Zhang, 2020). This can be exploited to reverse-engineer these systems to identify the training samples in question. The problem is these rare training samples often correspond to people from marginalized

communities who are underrepresented in the data. Large models, therefore, make people from marginalized communities more exposed to violation of privacy, surveillance and other potential harms related to identification.

Not only are these large models negatively biased against marginalized communities, they also have a large carbon footprint (Strubell *et al*., 2019). In 2019, a study showed that training a state-of-the-art neural-network-based AI model for NLP emits five times more CO2 than a car using fuel in its entire lifetime and 56 times more than a human in a year on average. This huge carbon footprint isn't singular to NLP systems; it is also a problem for state-of-the-art AI systems that are increasingly relying on a specific neural network architecture called Transformer Vaswani *et al*., 2017; Patterson *et al*., 2021), and researchers are looking into more energy-efficient alternatives (Patterson *et al*., 2021). Not only do these models have a high environmental impact, but they are also costly to train. Therefore, they put the future of AI as a technology in the hands of a few who have the means to develop it. In the same 2019 study of AI-based NLP models' carbon footprint, Strubell *et al*. also provided an estimate of the monetary cost of training them. For example, training one multi-purpose NLP model that received the best paper award at a prestigious AI conference (EMNLP) in 2018 required between US $103,000 and $350,000. These costs have increased since then as AI models get larger and larger. In October 2021, Microsoft and NVIDIA introduced Megatron-Turing NLG 530B, the world's most powerful language model (Alvi and Kharya, 2021). The transformer-based model has more than 530 billion parameters and requires $100 million to train (Alvi and Kharya, 2021; Simon, 2021). This increasing cost in training AI models heightens the power gap within the AI research community and for people who can benefit from the technology. Developing AI systems that require significant monetary cost excludes entire communities, such as the African continent, from the AI ecosystem.

RESHAPE AI TO MAKE IT A TECHNOLOGY THAT BENEFITS ALL

We are currently witnessing the emergence of an AI-driven technology revolution. Just like the revolutions preceding it, it will entirely transform the way we do business, interact with the world and live our day-to-day lives. However, to make AI a technology that benefits all, we would need to shift focus away from the pursuit of AGI and the task-driven culture that comes with it. What should be the defining characteristics of an AI field that benefits all?

AI should be more inclusive in terms of who it serves and who participates in its development. The huge monetary cost involved in the entire AI model development and deployment pipeline—from the cost of data collection to the cost of training and evaluation—puts the advancement and benefits of AI in the hands of a very few who have abundant resources. The innovations made in AI are driven by and centered around the interests of those elites. For AI to be more inclusive we need to make AI access a priority. There are concrete ways to improve AI access, such as by setting up and funding computer resources accessible by all for free.

In addition, AI access can be improved through funding for the creation and maintenance of databases for different domains, with carefully curated large datasets that are representative and respectful of privacy norms. Such funding would include making the datasets available to everyone at no cost, especially in areas with low resources. There is a growing awareness of the importance of high-quality data as calls towards data-centric AI gain traction. But we need to go beyond improving the quality of data for the interests of a powerful few and instead decentralize data collection and curation by encouraging and supporting the efforts of local communities, such as those in Africa.

Another AI access strategy involves centering innovation efforts around techniques that are more resource-efficient. Researchers and engineers should put efforts into developing methodologies that are less data-hungry and less computer-hungry. Funding agencies should reward research in this direction. Making AI more inclusive by prioritizing AI access will make it possible for AI to tackle a more diverse set of problems and push the field forward.

AI must be safe in order for it to benefit all. The current culture, driven by the goal of learning to perform human-like tasks, has led to advances that are beneficial to a few and harmful to many in marginalized communities. A safe AI would have frameworks in place for enforcing transparency in model development and performance; incorporating ways, within the evaluation process, to check the model for ethical attributes such as fairness and privacy; and having frameworks in place for explaining the cause of harm and easily intervening. The latter is possible if we enforce controllability in AI systems.

AI should empower humanity. This is only possible if we shift the focus away from task modeling and collaborate with domain experts to develop AI to solve humanity's most pressing problems, e.g., solve the climate crisis and healthcare, uplift the marginalized, unlock novel scientific discoveries and improve society. Developing AI that will allow us to seamlessly engineer information from data and humans into efficient solutions to humanity's problems is an opportunity the field should seize.

One thing that can potentially enable AI that benefits all is to foster broader adoption of the practices of a smaller community within the field of AI whose efforts we hear less about in mainstream media: the probabilistic machine learning community. This community is composed of researchers, the majority of whom are in academia, who are concerned with incorporating uncertainty and domain knowledge into decision-making systems. The methodologies developed by this community make it possible to learn even with very small amounts of data, making their approach to AI more inclusive towards resource-scarce communities. It is for this reason that we find techniques such as Gaussian processes and Bayesian neural networks developed by this community in critical domains such as healthcare and in the sciences. This community treats data as a first-class citizen and not as a mere tool for learning to perform a task. Their work often involves collaborating with domain experts to guide the development of methodologies targeted toward solving a problem. Empowering this community and adopting their data-centric and human-centric approach to AI will move us closer to a field that benefits all.

Reshaping AI is possible and will require the efforts of all stakeholders, including researchers, engineers, governments, the media, funding agencies and the private sector.

REFERENCES

Allyn, B. 2020. Amazon halts police use of its facial recognition technology. *NPR*, June 10. https://www.npr.org/2020/06/10/874418013/amazon-halts-police-use-of-its-facial-recognition-technology

AlphaFold team, 2020. AlphaFold: a solution to a 50-year-old grand challenge in biology. DeepMind blog, November 30. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Alvi, A. and Kharya, P. 2021. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, the world's largest and most powerful generative language model. Microsoft Research Blog, October 11. https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

Artificial Intelligence Index Report. 2021. Chapter 4: AI Education. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-_Chapter-4.pdf

Beckstead, N. and Muehlhauser, L. n.d. Potential risks from advanced artificial intelligence. Open Philanthropy. https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. March 2021, pp. 610–623. https://dl.acm.org/doi/10.1145/3442188.3445922

Candelon, F., di Carlo, R. C., De Bondt, M., and Evgeniou, T. 2021. AI regulation is coming. *Harvard Business Review*, September-October issue. https://hbr.org/2021/09/ai-regulation-is-coming

Caswell, I. and Liang, B. 2020. Recent Advances in Google Translate. Google AI Blog, June 8. https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html

Chung, C. 2017. Tech giants are paying huge salaries for scare A.I. talent. *New York Times*, October 22. https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html

Columbus, L. 2019. AI skills among the most in-demand for 2020. *Forbes*, November 27. https://www.forbes.com/sites/louiscolumbus/2019/11/27/ai-skills-among-the-most-in-demand-for-2020/?sh=69de02b36b44

DeepMind, n.d. Home page. https://deepmind.com/

Dunn, S., Carswell, N., Doagoo, B. C., Shoker, S. and Tatsis, S. 2020. Deepfakes and digital harms: Emerging technologies and gender-based violence. Online presentation and discussion (video), November 25. Centre for International Governance Innovation. https://www.cigionline.org/events/deepfakes-and-digital-harms-emerging-technologies-and-gender-based-violence/

European Commission, n.d. Regulatory framework proposal on artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Feldman, V. and Zhang, C. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *arXiv:2008.03703 [*cs.LG*]*. https://arxiv.org/abs/2008.03703

Galston, W. A. 2020. Is seeing still believing? The deepfake challenge to truth in politics. The Brookings Institution, January 8. https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/

General, J. and Sarlin, J. 2021. A false facial recognition match sent this innocent Black man to jail. *CNN Business*, April 29. https://www.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html

Gibbs, S. 2017. AlphaZero AI beats champion chess program after teaching itself in four hours. *The Guardian*, December 7. https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours

GOV.UK. n.d. Guidance: National AI Strategy. https://www.gov.uk/government/publications/national-ai-strategy

GPT-3. 2020. A robot wrote this entire article. Are you scared yet, human? *The Guardian*, September 8. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Harwell, D. 2021. Amazon extends ban on police use of its facial recognition technology indefinitely. *Washington Post*, May 18. https://www.washingtonpost.com/technology/2021/05/18/amazon-facial-recognition-ban/

Hill, K. 2020. The secretive company that might end privacy as we know it. *New York Times*, January 18. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hunt, E. 2016. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*, March 24. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=

Hutson, M. 2017. This computer program can beat humans at Go—with no human instruction. *Science*, October 18. https://www.science.org/content/article/computer-program-can-beat-humans-go-no-human-instruction

Invest in Canada. n.d. Pan-Canadian AI Strategy. https://www.investcanada.ca/programs-incentives/pan-canadian-ai-strategy

Jordan, M. 2018. Artificial intelligence—The revolution hasn't happened yet. Medium.com (personal blog), April 19. https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* Vol. 596, pp. 583–589. https://www.nature.com/articles/s41586-021-03819-2

Kahn, J. 2021. DeepMind spins out new Alphabet company focused on drug discovery. *Fortune*, November 4. https://fortune.com/2021/11/04/deepmind-spins-out-alphabet-company-isomorphic-drug-discovery-company/

Karras et al. and Nvidia. 2019. Imagined by a GAN (generative adversarial network) StyleGAN2. https://thispersondoesnotexist.com/

Langston, J. 2021a. From conversation to code: Microsoft introduces its first product features powered by GPT-3. Microsoft/The AI Blog, May 25. https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/

Langston, J. 2021b. New Azure OpenAI Service combines access to powerful GPT-3 language models with Azure's enterprise capabilities. Microsoft/The AI Blog, November 2. https://blogs.microsoft.com/ai/new-azure-openai-service/

Lucy, L. and Bamman, D. 2021. Gender and representation bias in GPT-3 generated stories. *Proceedings of the 3rd Workshop on Narrative Understanding, Association for Computational Linguistics*, June 11. pp. 48–55. https://aclanthology.org/2021.nuse-1.5.pdf

Mac, R. 2021. Facebook apologizes after A.I. puts "primates" label on video of Black men. *New York Times*, September 3. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

McQuaid, J. 2021. Limits to growth: Can AI's voracious appetite for data be tamed? *Undark*, October 18. https://undark.org/2021/10/18/computer-scientists-try-to-sidestep-ai-data-dilemma/

Microsoft. 2019. OpenAI forms exclusive computing partnership with Microsoft to build new Azure AI supercomputing technologies. Microsoft News Center, July 22. https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/

Nadella, S. 2018. Microsoft + GitHub = empowering developers. Microsoft/Official Microsoft Blog, June 4. https://blogs.microsoft.com/blog/2018/06/04/microsoft-github-empowering-developers/

National Science Foundation (NSF). n.d. Artificial intelligence at NSF. https://www.nsf.gov/cise/ai.jsp

Nayak, P. 2019. Understanding searches better than ever before. Google/The Keyword blog, October 25. https://blog.google/products/search/search-language-understanding-bert/

Newman, J. 2021. GitHub's new tool uses AI to craft code. Some developers are furious. *Fast Company*, July 9. https://www.fastcompany.com/90653878/github-copilot-microsoft-openai-coding-tool-backlash

Nguyen, T. 2021. An overview of ICML 2021's publications. VinAI Research/Achievements blog, July 18. https://www.vinai.io/an-overview-of-icml-2021s-publications

Ong, T. 2017. Facebook apologizes after wrong translation sees Palestinian man arrested for posting "good morning." *The Verge*, October 24. https://www.theverge.com/us-world/2017/10/24/16533496/facebook-apology-wrong-translation-palestinian-arrested-post-good-morning

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. and Dean, J. 2021. Carbon emissions and large neural network training. *arXiv:2104.10350 [*cs.LG*]*. https://arxiv.org/abs/2104.10350

Pilipiszyn, A. 2021. GPT-3 Powers the Next Generation of Apps. OpenAI blog, March 25. https://openai.com/blog/gpt-3-apps/

Raghavan, P. 2020. How AI is powering a more helpful Google. Google/The Keyword blog, October 15. https://blog.google/products/search/search-on/

Rakicevic, N. 2021. NeurIPS Conference: Historical data analysis. Medium.com/Towards Data Science, February 27. https://towardsdatascience.com/neurips-conference-historical-data-analysis-e45f7641d232

Ryan-Mosley, T. 2021. The new lawsuit that shows facial recognition is officially a civil rights issue. *MIT Technology Review*, April 14. https://www.technologyreview.com/2021/04/14/1022676/robert-williams-facial-recognition-lawsuit-aclu-detroit-police/

Safra, L., Chevallier, C., Grèzes, J. and Baumard, N. 2020. Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. *Nature Communications* Vol. 11, Art. No. 4728. https://www.nature.com/articles/s41467-020-18566-7

Samuel, S. 2021. AI's Islamophobia problem. *Vox*, September 18. https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim

Simon, J. 2021. Large language models: A new Moore's Law? Hugging Face blog, October 26. https://huggingface.co/blog/large-language-models

Snow, J. 2018. Amazon's face recognition falsely matched 28 Members of Congress with mugshots. ACLU blog, July 26. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Strubell, E., Ganesh, A. and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019. *arXiv:1906.02243 [*cs.CL*]* https://arxiv.org/abs/1906.02243

Taft, D. K. 2021. GitHub Copilot: A powerful, controversial autocomplete for developers. *The New Stack*, July 1. https://thenewstack.io/github-copilot-a-powerful-controversial-autocomplete-for-developers/

Uszkoreit, J. 2017. Transformer: A novel neural network architecture for language understanding. Google AI Blog, August 31. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. 2017. Attention is all you need. *arXiv:1706.03762 [*cs.CL*]*. https://arxiv.org/abs/1706.03762

Weiss, T. R. 2021. AI startups continue to rack up millions in VC funding in August 2021. *EnterpriseAI*, August 26. https://www.enterpriseai.news/2021/08/26/ai-startups-continue-to-rack-up-millions-in-vc-funding-in-august-2021/

Wen, Y., Raj, B., and Singh, R. 2019. Face Reconstruction from Voice using Generative Adversarial Networks. *NeurIPS Proceedings*. https://proceedings.neurips.cc/paper/2019/hash/eb9fc349601c69352c859c1faa287874-Abstract.html

Wilhelm, A. and Heim, A. 2021. Huge deals are pushing more AI startups into IPO territory. *TechCrunch*, November 6. https://techcrunch.com/2021/11/16/huge-deals-are-pushing-more-ai-startups-into-ipo-territory/

Zhang, M. 2015. Google Photos tags two African-Americans as gorillas through facial recognition software. *Forbes*, July 1. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=90e05f713d8b

Zweig, G., Auli, M. and Fagan, F. 2021. Learning from videos to understand the world. Meta AI, March 12. https://ai.facebook.com/blog/learning-from-videos-to-understand-the-world

BIG AI CAN CENTRALIZE DECISION-MAKING AND POWER, AND THAT'S A PROBLEM

ERIK BRYNJOLFSSON

Jerry Yang and Akiko Yamazaki Professor and Senior Fellow at the Stanford Institute for Human-Centered AI (HAI), Director of the Stanford Digital Economy Lab, and a Research Associate at the National Bureau of Economic Research (NBER).

ANDREW NG

Founder and CEO of DeepLearning.AI, Founder and CEO of Landing AI, and Adjunct Professor at Stanford University's Computer Science Department.

SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

BIG AI CAN CENTRALIZE DECISION-MAKING AND POWER, AND THAT'S A PROBLEM

ABSTRACT[11]

Over the past few years, artificial intelligence (AI) systems have grown dramatically larger and more powerful and now have the potential to substantially increase the centralization of decision-making. This can be more efficient, but it also can lead to concentration of wealth and power. While the current path of technology is opening the door for an increase in the centralization of decision-making and power, we don't think that outcome is inevitable. It may be possible to design decentralized, interoperable, or federated technologies that maintain the decentralization of decision-making and power. More fundamentally, we can strengthen democracy and other political institutions to serve as a check on machine-based decision-making.

In this chapter, we discuss some of the recent trends in AI and other technologies that may tip the balance between centralization and decentralization. We contrast non-human decision-making systems with human systems, review some of the empirical evidence on concentration to date, and lay out some technological, economic and policy options. Our purpose is not to predict a particular future, but to warn of a set of outcomes in which AI contributes to an unprecedented level of concentration. These outcomes are possible if we don't act responsibly.

11. Acknowledgements. The authors are grateful to Lynn He for a tremendous amount of help editing and revising this chapter.

INTRODUCTION

Over the past few years, AI systems have grown dramatically larger and more powerful. OpenAI's GPT-3 language model has over 175 billion parameters while Google has trained models with over one trillion parameters (Talagala, 2021). These models are achieving human-level or even superhuman performance in more and more domains, albeit narrow ones. For instance, AlphaZero, a computer program developed by DeepMind to master certain games, played 4.9 million games against itself and within a day had a chess rating superior to any human (Silver et al., 2017). These machine learning (ML) models are trained on historical data or simulated data to predict future outcomes and detect patterns. As the number of model parameters and data processing capacity increase, this helps the model to generalize better. AlphaZero is not only better than humans at chess, but also at Go and other games such as shogi. Furthermore, it can defeat special-purpose AI systems, such as Stockfish, that were developed over many years for the specific purpose of playing chess. Likewise, GPT-3 can generate text that is often indistinguishable from text created by humans, expanding on the purely analytical language understanding capabilities of most of its predecessors.

These ever larger and more powerful models have the potential to significantly increase productivity in areas as diverse as software development (Belton, 2021), medical diagnosis (Ronneberger et al., 2015), and predicting natural disasters (Devries et al., 2018); these can also in some cases improve human living standards. But they also can lead to a substantially greater centralization of decision-making. Because of any human brain's limitations in terms of the amount of information it can process and the number of decisions it can make per day, decision-making has historically been decentralized via markets and other distributed systems such as hierarchical organizations. However, the growing power of machines to analyze larger and larger amounts of information, as well as make thousands of decisions per second, suggest that machine-based decision-making might, in principle, be far more centralized.

While centralized decision-making can be more efficient, for instance because it can take into account dependencies across different units, it also can lead to a concentration of wealth and power. This is not a good outcome for those who have lost some or all of their decision-making authority.

While the current path of technology is opening the door for an increase in the centralization of decision-making and power, we don't think that outcome is inevitable. It may be possible to design decentralized, interoperable, or federated technologies that maintain the decentralization of decision-making and power. Indeed, a number of thinkers have argued that developments such as the internet, blockchain, and related technologies can drive increased decentralization (Malone, 2003; Pueyo, 2021; Srinivasan, 2019; Lera et al., 2020). It is also possible to support the decentralization of economic power by investing in the more widespread distribution of human, physical and financial capital, including tools such as basic income or vouchers for skill development or political contributions. More fundamentally, we can strengthen democracy and other political institutions to serve as a check on machine-based decision-making.

In this chapter, we discuss some of the recent trends in AI and other technologies that may tip the balance between centralization and decentralization, contrast non-human decision-making systems with human systems, review some of the empirical evidence on concentration to date, and lay out some technological, economic and policy options. Our purpose is not to predict a particular future, but to warn of a set of outcomes in which AI contributes to an unprecedented level of concentration. These outcomes are possible if we don't act responsibly.

AI SYSTEMS ARE BECOMING LARGER

AI systems underpin many core operations of the modern economy. From the humble and familiar line of best fit $y = mx + b$, which features just two parameters, to state-of-the-art neural networks with billions of parameters, machine-learning models and the datasets they consume have scaled to sizes oftentimes beyond human comprehension.

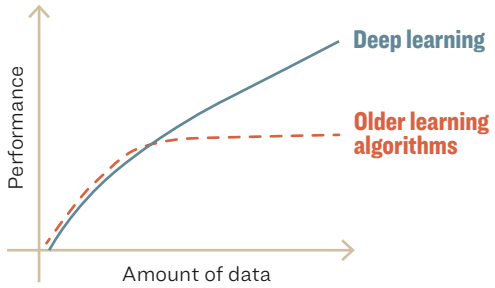
This explosion in scale can be explained by a simple logic: larger models can process larger quantities of data. If ML can be summarized as training models to make predictions based on historical data, it follows that the greater the quantity and diversity of the data (the more examples a model sees), the better an algorithm's performance. This line of reasoning has induced an academic and industry race to build bigger models which can generalize to more data. Numerous bottlenecks in innovation and scale were eliminated by harnessing more computational power. In 2009, Andrew Ng and his group at Stanford recognized the potential of computer graphics processing unit chips (GPUs)—invented to process video game graphics—to parallelize deep neural network computations, effectively reducing training time from weeks to days (Ng et al., 2009). This helped usher in a new era of ML classed as deep learning and allowed for innovations in model architecture and dataset size which produced remarkable results.[12]

Today, the use of GPUs is commonplace, even basic, and deep learning operates many of the ordinary functions of our lives, including content recommendation on social media platforms, organizing ride-hailing services, online shopping, and many other applications (Brynjolfsson and McAfee, 2017a). The sustained dominance of this particular data science technique can be explained in part by the fact that deep learning, more than its predecessors, scales with data (see Figure 1).

12. For example, deep learning achieved dramatically better results in image classification compared to earlier approaches. See Krizhevsky, Sutskever and Hinton (2012).

| FIGURE 1 |

Why deep learning? Deep learning algorithm performance continually improves with more data, whereas older learning algorithms plateau in performance (Ng, 2015).

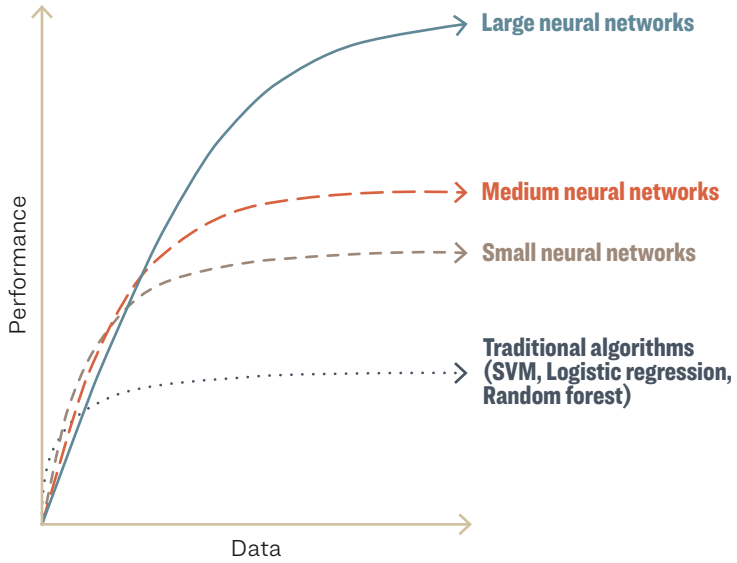


How do data science techniques scale with amount of data?

This follows the simple logic posited earlier: the more examples a model can train on, the better its performance. The larger the model, the more parameters it can employ to infer from increasing dataset sizes. This logic has become a conventional AI wisdom of sorts and spurred intense parallel contests in both computational capacity and model size. Just as the gap from traditional ML algorithms to deep learning techniques was bridged by advancing from central processing units (CPUs, with one1 million connections) to graphic processing units (GPUs, with 10 million connections), further expansions of model architectures can be expected to arise from exploiting cloud computing (many CPUs, with one billion connections) and high-performance computing (HPCs) (many GPUs, with 100 billion connections), and more (see Figure 2).

| FIGURE 2 |

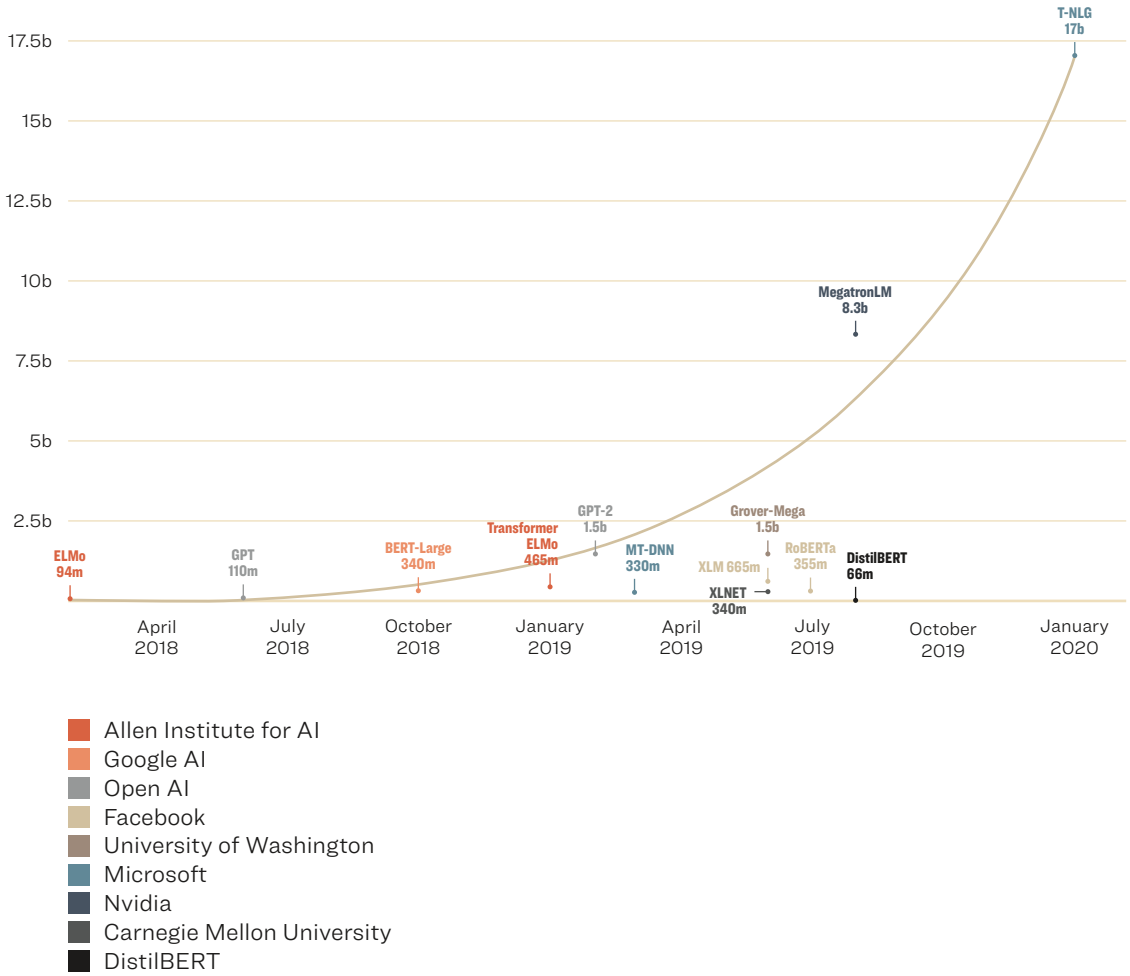
Model performance becomes saturated if there are not enough parameters to infer from the data (Ng, 2016).



Already, the rapidly increasing size of language models over the past few years is a reflection of these growing computational capabilities and of the aspirations of ML technologists to train on ever more data. Figure 3 shows how the size of Natural Language Processing (NLP) models has grown at an increasing rate over the past few years. Note that the figure does not include OpenAI's GPT-3 (175 billion parameters), which exceeds Microsoft's T-NLG by another order of ten.

| FIGURE 3 |

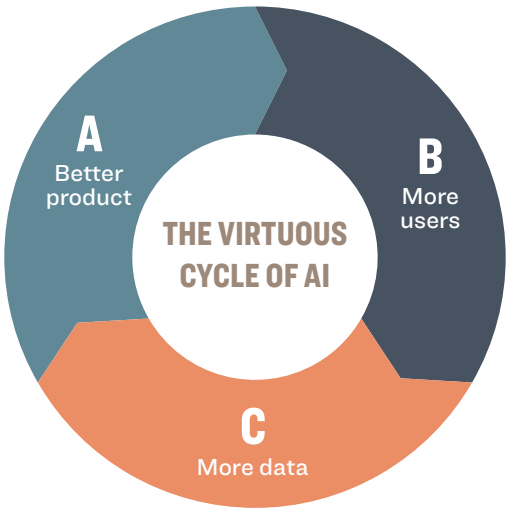
Language model sizes (Ng, 2020).



The near trivial cost of data accumulation only exacerbates this tendency towards scale. Deep learning models have the capability to continuously integrate new data for training without significant computing cost—a quality that is often capitalized on in the "virtuous cycle of AI": customer-facing AI applications organically generate a constant stream of customer data, which can then be incorporated into AI systems at a low cost to create more accurate predictions (see Figure 4). The superior model attracts more consumers, which in turn yields more user data. Without similar access to quality customer data, it is difficult for potential competitors to break into the positive feedback loop of data accumulation.

| FIGURE 4 |

The virtuous cycle of AI (Ng, 2020).



These dynamics are similar to those for web search: leading web search engines such as Google, Bing, and Baidu have vast amounts of data that show them what links a user clicks on after each search query. This data helps the companies build a more accurate search engine product (A), which in turn helps them acquire more users (B), which in turn results in their having even more of the most relevant and valuable user data on the market (C).

When the gains of model performance do not diminish, the inherent benefits of this cycle produce a natural winner-takes-all market dynamic, where even small gains can be decisive (Frank and Cook, 2013). These markets exhibit a superstar effect, where there is little to no incentive for the customer to choose a product that is even marginally worse than the best option. Web search serves again as a prime example of this phenomenon since the customer has low incentives to pick a search engine that is even slightly less accurate than Google. Even an artificial rewards system, such as the one created by Bing to encourage user searches and generate data, cannot replicate the velocity of natural data accumulation from the virtuous cycle.

To maximize the effects of positive feedback, companies must also employ the appropriate data strategy. In particular, centralizing data is crucial to taking advantage of the scaling properties of deep learning. If databases are siloed under the management of many different executives or divisions, it is nearly impossible for engineers to retrieve the data and draw meaningful connections. Unifying data warehouses supports the creation of one big AI system that allows for the maximum number of inferences, and outperforms many small systems through properties of scale (Figure 2).

While model performance gains do not diminish as a function of scale, innovations in computing power, data, and model architecture have tended toward centralization. CPUs were improved upon by GPUs, only to be supplanted by their centralized counterparts: cloud computing and high-performance computing. The virtuous cycle of AI promotes constant mining and aggregation of data while models

grow exponentially larger to accommodate and generalize on these expanded datasets. Access to deployment and profit from AI systems have followed the tendencies exhibited by their innovation and become increasingly concentrated in the hands of a few highly educated workers and stakeholders. Thus, the progression towards centralization in AI and its downstream effects give substantial cause for investigation.

HUMAN-BASED DECISION-MAKING IS DECENTRALIZED

The potential for the centralization of non-human decision-making contrasts with the relative decentralization of decision-making in human markets and organizations. At its core, this reflects the fact that human information processing capacity is bounded. As Simon (1955, p. 99) noted, any realistic model of decision-making must consider these limits:

> Broadly stated, the task is to replace the global rationality of economic man with the kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist.

For example, it has been estimated that the memory capacity of the human brain is 2.5 petabytes (Reber, 2010), and some estimates of the unconscious processing capacity are as high as 11 million bits per second (Wilson, 2004). However, the brain also has a number of bottlenecks (Marois and Ivanoff, 2005). As a result, conscious or "intelligent" information processing, such as reading, may be as low as 50 to 60 bits per second (Markowsky, 2021; Emerging Technology from the arXiv, 2009).

Regardless of the exact number, it is, of course, finite and small enough that no one human, no matter how intelligent, can make all the business and economic decisions for even a small organization, let alone an industry or an economy.

We can see the implications of this fact in the design of human organizations. Within firms, the locus of decision-making is distributed: some people may interact with customers while others design products. Customer teams and design teams themselves have limited responsibilities: each manages only a subset of customers or products. Other people focus on managing inventory, manufacturing, or supply chains, while still others make decisions about finance, marketing, hiring or the overall strategic direction of the firm. Some decisions are very localized, such as whether to empty a particular wastebasket, while others have broader effects on the company, such as whether to enter a new market.

As noted by Smith (1776), Hayek (1945), and many others, one of the benefits of markets and the price system is that they make it possible to further decentralize and distribute decision-making beyond the boundaries of the firm. A pencil manufacturer doesn't have to make the myriad decisions needed to create the wood, graphite, tin and rubber that pencils are made from, or the costs and benefits across the alternative sources of supply or demand for each at any given time. The pencil factory owner only needs to know that there is a market price for each material that serves as a sufficient statistic for these trade-offs. Likewise, an individual consumer will often simply focus on estimating the personal benefits from purchasing a pencil and then comparing it with the market price that reflects myriad trade-offs made by other decision-makers who design, produce, deliver and sell that good. The first fundamental theorem of welfare economics states that when there are no externalities, perfect information and no market power, the resulting equilibrium will be Pareto optimal (Hammond, 1997).[13]

13. In practice, all these conditions are not met, which limits their practical significance to some extent.

Literally millions of decision-makers can thus each focus their attention—their bounded rationality, to use Simon's language (1955)—on just a small aspect of the broader decision problem needed to run the economy, and largely ignore everything else (or more precisely, assume that these other factors are sufficiently summarized by prices).

In the 1930s and 1940s, there was a vigorous debate about whether all these decisions could, in principle, be fully centralized. One side of this "socialist calculation" debate, led by Lange (1936), Lerner (1938), and others, argued that all of the information needed could be transmitted to a central decision-maker, or team of decision-makers, who would then calculate the necessary costs and benefits, determine the optimal allocation, and transmit instructions to people in the rest of the economy about what to produce, transport and consume. The other side, led by von Mises (1951), Hayek (1945), and others, argued, in essence, that there was simply too much information for such a calculation to be feasible.

In particular, Hayek's classic article (1945, p. 522), "The Use of Knowledge in Society," pointed out that every person has various nuggets of unique information which might be beneficial in some circumstance: "To know of and put to use a machine not fully employed, or somebody's skill which could be better utilized, or to be aware of a surplus stock which can be drawn upon during an interruption of supplies...."

This dispersed knowledge defies statistical aggregation: there is value in knowing about a particular empty truck at a particular location. Knowing that on average, there are many trucks returning empty from journeys somewhere in America, or that there is a general formula for optimal truck routing, is no substitute for the specific knowledge of which truck is empty. So, Hayek argued that the important role of an economic system was to put the decision rights, and the accompanying incentives to use those decision rights, in the hands of the people who had the relevant knowledge. Since the knowledge was dispersed, decision rights must be dispersed as well. Jensen and Meckling (1992) generalized this insight to decision rights with a large firm, where the problem is complicated by the difficulty in assigning alienable ownership rights.

In contrast, an early attempt to centralize economic decision-making was Project Cybersyn, developed in the early 1970s to manage the Chilean economy (Morozov, 2014). It used a mainframe computer in the capital city of Santiago connected to factories throughout the country via a national network of telex machines. Information such as raw material supplies or labor productivity would be fed into an economic simulation software program, which would in turn generate directives to the factories and other organizations.

Cybersyn was abandoned in 1973[14] and with the collapse of the Soviet Union in 1989, it appeared that the socialist calculation debate was settled decisively in favor of those who argued for decentralized decision-making. The central planning approach, while working reasonably well in some areas, such as heavy industry, appeared to be outmatched in highly innovative or rapidly changing parts of the economy such as the computer hardware, software, and digital networks which the US and its more decentralized approach came to dominate.

However, the same digital technology industries that were enabled by this system also make possible far more centralization of decision-making than in the past. As noted by Brynjolfsson and Mendelson (1993), while one way to achieve the collocation of information and decision rights is by moving decision rights to the information, another option is to move the information itself. Large information systems and networks, such as enterprise resource planning systems and the internet, make this feasible

14. While it has ended operations, Cybersyn lives on in science fiction. For example, Jorge Baradit described it as "creating the first cybernetic state, a universal example, the true third way, a miracle" in his novel *Synco: The Game of Reverse* (2009).

for more and more types of information. At the same time, technologies—for instance, sensors such as RFID (Radio Frequency Identification) tags and the Internet of Things (IoT)—make it easier to record and digitize the types of dispersed information that Hayek emphasized in his classic article.

To give a concrete example, for most of the 20th century, a local small grocer would know better than anyone how popular different flavors of chewing gum or ice cream were in their neighborhood, and which local vendors could deliver those goods reliably and cheaply, exactly as Hayek described. But in the 1990s, Walmart developed and implemented sophisticated systems that tracked demand at the point of sale and shipments from individual vendors as they worked their way across the country. Today, it's likely that a database in Bentonville, Arkansas has more detailed knowledge of each neighborhood's purchases of peppermint ice cream yesterday than the typical local grocer, not to mention the predictive analytics needed to predict demand tomorrow as a function of seasonal patterns, interactive marketing campaigns, mobile phone traffic, weather forecasts and myriad other variables (Brynjolfsson et al., 2021). Not surprisingly, small grocers, and small retailers of all types, are losing market share to large chains (Decker et al., 2020).

A handful of large online retailers take this detailed knowledge even further. There's little chance a salesperson from a physical bookstore could make book recommendations with the precision and insight of Amazon's recommender tools, which draw not only on each customer's detailed purchase histories and preferences but also make inferences across customers using state-of-the-art ML systems trained on terabytes of data. Likewise, systems are being developed to predict engine failures before they happen (GE Research, n.d.), inventory levels in warehouses (Chang, 2020), traffic conditions (Lau, 2020), and myriad other local conditions that previously required on-the-spot data, expertise, and decision-making.

It's not just the information that can now be centralized, but more fundamentally, as discussed in the previous section, the decisions themselves can often be done by machines. While the processing capacity of the human brain hasn't changed in millennia, computer processing power has been doubling roughly every two years since Gordon Moore made his eponymous forecast in 1965.

The Grossman-Hart-Moore framework provides a way to map from decision rights to property ownership, bargaining power, and firm boundaries (Grossman and Hart, 1986; Hart and Moore, 1990). In particular, the authors show that it is optimal to assign ownership of assets, and with them the residual rights of control they entail, to the most important decision-makers in any economic decision. This gives them the bargaining power to claim a share of the resulting value they create from their decisions and thus helps align their incentives so that more value will be created.

Brynjolfsson (1994) extended this framework to show that when knowledge moves from human brains to non-human assets (such as a database or AI system), then it becomes not only feasible but often more economically efficient to centralize not only the knowledge but also the accompanying ownership of other assets. The Grossman-Hart-Moore framework associates separate ownership of assets with market transactions that occur between distinct firms. In contrast, unified asset ownership is the distinguishing characteristic of transactions that occur within a single firm. Thus, by centralizing knowledge and decision-making, AI systems can also make it optimal to centralize asset ownership. According to this framework, that means moving transactions inside the boundary of the firm and reducing reliance on market transactions.

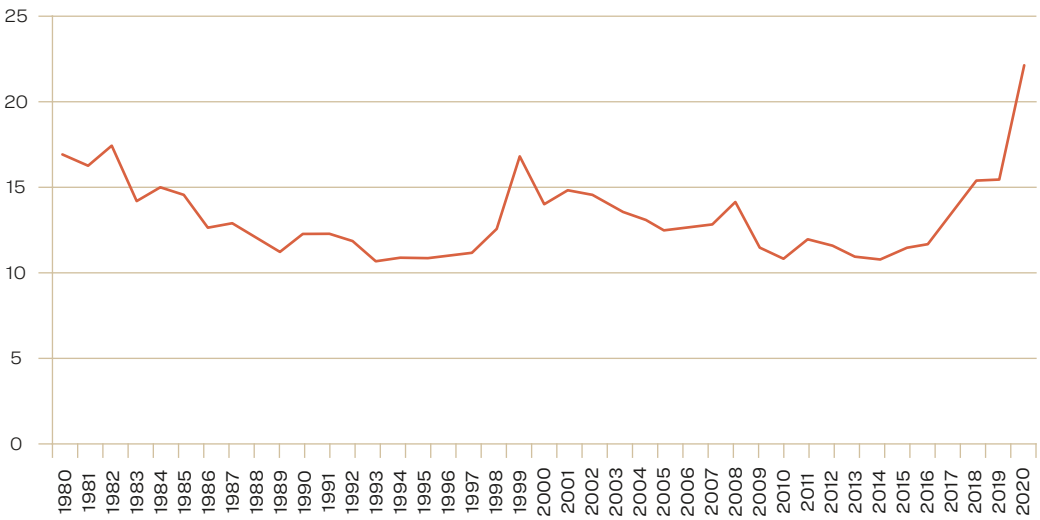
CENTRALIZATION IS ALREADY HAPPENING IN MANY AREAS

We have argued that AI can lead to a centralization of decision-making. This is a function of two things: the shift of knowledge and information from human brains to machines and, separately, the shift of the processing power needed to make decisions from human brains to machines.

We see examples of this phenomenon not only in the rise of multi-store chains and online retailers replacing local stores but also more broadly. For instance, for the first time in American history, there are five companies with over a trillion dollars in market cap (Apple, Amazon, Alphabet, Microsoft, and Facebook). The top five companies in the S&P 500 now account for over 22% of the total index capitalization, a modern-day record (Scheid, 2020), as seen in Figure 5.

| FIGURE 5 |

Yearly percentage of the average share of the S&P 500's market cap represented by that year's five largest companies (Scheid, 2020).



Data compiled July 24, 2020.
Yearly percentage reflects the average share of the S&P 500's market cap represented by that year's five largest companies.
Source: S&P Dow Jones Indices

Concentration is growing not only in digital industries but also in the American economy overall. For instance, 80 percent of the beef market is controlled by four packers (MacDonald et al., 1999) and many Americans are limited to one choice for their broadband provider (Rogers, 2017). These figures are supported by studies that demonstrate that the profits of larger companies have grown over the last 40 years, limiting opportunities for smaller firms to create disruptive competition (Boushey and Knudsen, 2021). In fact, in 2020, market concentration among the top 3000 firms by market cap hit its highest point since 1986 according to Bank of America (*The Economist*, 2021).

That said, most of the increased centralization in the US in recent years is not driven by AI per se, but rather by technological forces other than AI and by policy choices, such as deregulation and changes in antitrust enforcement (Stuck and Ezrachi, 2017). Specifically, on the technology side, there are increasingly important supply-side and demand-side economies of scale that favor larger firms.

On the supply side, the production of computer hardware and especially software tends to entail high fixed costs and low, or even zero, marginal costs. The first copy of a new microprocessor or a new kind of database system might require millions or even billions of dollars in research, development, and other initial costs to produce.[15] But additional copies can be made at little or no additional cost. This means that bigger companies that sell more copies will have lower average costs, enabling them to charge lower prices while making greater profits. This kind of cost structure is common in digital industries.

As "software eats the world," in the words of Marc Andreessen, it is also coming to other industries which are increasingly digitizing (Brynjolfsson et al., 2008). For instance, in automobiles, the cost contribution of electronics increased from 18 to 40 percent over the past 20 years (Deloitte, 2019). Retailers such as CVS pharmacy have digitized their business processes, making it possible for them to replicate and scale much faster (McAfee and Brynjolfsson, 2008).

On the demand side, more and more companies, from Facebook and Apple to Uber and Airbnb, benefit from scale. Because of network effects, the value of their products and services increases as more users use their products. These effects can be direct, as when friends and relatives all join the same social network and share posts, or indirect (two-sided) as when Uber riders benefit from the increased number of drivers or when advertisers benefit from the increased number of users. The scale effects are often loosely modeled as following Metcalfe's Law, which predicts that total value grows with the square of the number of users, while costs only increase linearly. In many cases, a preferential attachment model (i.e., new customers are more likely to join the networks that already have more existing customers) will lead to a winner-takes-most outcome that can be well described by a power-law distribution. While AI is not at the core of most network effects, in many cases, such as with the routing engine that facilitates ride-hailing, network effects are amplified by AI.

As with supply-side economies of scale, network effects (a.k.a. demand-side economies of scale) are especially common among industries that use digital technologies. While supply-side economies and network effects favor a single large production facility or a single large network, they do not necessarily entail centralized decision-making. For instance, multiple companies can use a shared facility while separate networks can be made interoperable, reaping the benefits of network effects even while ownership is independent.[16] Thus, regulatory approaches to creating and maintaining decentralized decision-making power, such as requiring mobile phone number portability, can work technically and economically (Federal Communications Commission, 1996).

In contrast, big AI is different. It's not simply about centralizing the operations, but rather about centralizing decision-making itself. While one can often find ways to distribute control even if operations are centralized, it's much harder to distribute control if the decision-making is centralized. If a large AI-based system, drawing on large amounts of data, consistently makes better decisions than any of the entities overseeing it, then how can we create credible checks and balances? And how can we ensure that the values and goals of the entity making these decisions are aligned with the values and goals of the broader population?[17]

15. For instance, Intel's new chip manufacturing campus will cost between $60 billion and $120 billion (Shilov, 2021).

16. Consider for instance, how the numerous independent network operators coordinate data traffic on the internet.

17. Russell (2019), Bostrom (2016) and Yudkowsky (2016) are among those who have emphasized this challenge.

HIGHLY CENTRALIZED DECISION-MAKING CAN BE BAD FOR SOCIETY

There are many benefits to centralizing decision making. Most notably, centralization makes it easier to consider interdependencies and interactions across units and optimize globally, not just locally.

However, centralization also creates risks. With the centralization of decision-making comes the centralization of power. If the decision-maker given this power is benevolent and seeks to create value for all those affected, this is not necessarily a bad thing. But, as Lord Acton (1887) argued, "power tends to corrupt and absolute power corrupts absolutely." Thus, it can be risky to rely on the good intentions of centralized decision-makers. Even when they initially seek to preserve freedom and equal rights, if there is no natural check on their power, those intentions may not last.

In particular, concentration of economic decision making can undermine democracy. As Louis Brandeis put it (Dilliard, 1941, p. 42–45): "We can have democracy in this country, or we can have great wealth concentrated in the hands of a few, but we can't have both." Furthermore, the tools of AI are often used directly to control information flows in ways that can be harmful to democracy (Acemoglu, 2021).

Historically, the concentration of power in a country or system has also been bad for human rights. While it's not a universal rule, people who don't have a role in decision-making are often not treated well (Acemoglu and Robinson, 2012). Thus, some political scientists have argued that it is worth having a less efficient system, with lots of checks and balances, rather than a more streamlined system, to better preserve the rights of ordinary people (see, e.g., Reich et al., 2021). While these inefficiencies can be frustrating, it's not clear that there is a better alternative. As Winston Churchill quipped, "democracy is the worst form of government—except for all the others that have been tried."

RECOMMENDATIONS

As noted in the introduction, it is not inevitable that decision-making will become more centralized. There are powerful forces that go in the opposite direction and the ultimate outcome will depend greatly on the choices we make today. In particular, three broad sets of approaches can be applied to reduce the centralization of decision-making or mitigate its risks: technological, economic and political.

Technological approaches

We can design technology to encourage decentralized decision-making by making it easier for many different people to invent, innovate and improve new goods and services. Platforms can be designed to allow ideas and entrepreneurship to flourish and to empower the innovators with decision rights and economic rewards (Phelps, 2015).

Interoperability and consistent standards can preserve the benefits of network effects while allowing decentralized ownership. Paradoxically, a very clear and rigid set of standards often makes it easier for people to be creative in designing components that work with those standards. TCP/IP, the standard at the heart of the internet, is a compelling example of this effect. Innovation was spurred as millions of distributed entrepreneurs, researchers, and developers created applications, from email and the world wide web to voice-over-IP and the Internet of Things, that rapidly gained scale and impact as long as they adhered to a few core standards.

Just as interoperability has preserved the benefits of network effects while distributing ownership and power, data replication and sharing can enable multiple ML systems to benefit from very large data sets. Indeed, one of the great benefits of digital data is that it can be replicated at very low cost. As a result, making multiple copies of image data, language data or other types of data so that different teams and organizations can each train competing models need not be expensive.

In tandem, it would be beneficial to widely distribute AI tools, such as large pre-trained models, as well as big data resources in order to allow smaller players to be more competitive with dominant firms, counteracting the concentrating effects of preferential attachment. For instance, Lera et al. (2020) propose a federated system where there is a network of participants that share global data descriptions (not the underlying data) and each entity locally combines these global descriptions with their own local, private data.

Other technologies have been invented specifically with the intent to decentralize power, most notably the blockchain. Instead of centralized gatekeepers approving transactions, as is required by traditional database systems and ledgers, the blockchain seeks to decentralize this process. While the promise of the blockchain is intriguing, and some have even argued that it could lead to the re-emergence of a more decentralized financial infrastructure (Pentland, 2015) or even the demise of the nation-state (Pueyo, 2021), in practice it often has ended up even more centralized than conventional infrastructure. For instance, it is estimated that at one point just four bitcoin miners in China controlled more than 50 percent of mining (Sharma, 2019).

We can also seek to design AI systems to augment human decision-making rather than replace it, thereby potentially supporting more decentralization of power.[18] Indeed, while more and more decisions can be made by machines, humans are still superior at most tasks, and still more are best completed by a combination of human and machine. This means that there are still many types of decisions that will remain decentralized. To the extent that technologists focus on systems that complement rather than replace humans, this may be a sustainable path for some time. In particular, humans are often better than ML systems at managing unstructured tasks and problem definition, the kinds of creative work that are particularly important for innovation and entrepreneurship (Brynjolfsson et al., 2017).

While these technological approaches can be helpful, it's also likely that they will also be opposed by entities that benefit from preventing them. For instance, while telephone number portability from one mobile carrier to another has obvious benefits for consumers, it was not implemented by mobile phone operators and was opposed for many years (Douglass, 2002). Only after the United States Congress required number portability in 2003 did the carriers comply (Federal Communications Commission, 2009). The types of data sharing and data and knowledge interoperability required to enable distributed ML systems are likely to be substantially more difficult to implement than simple mobile phone number portability. This can create implementation lags and unintended consequences. For instance, some have argued that GDPR, the General Data Protection Regulation, which was intended to protect the privacy rights of consumers, may have further entrenched the power of large digital platforms (Nouwens et al., 2020).

Ultimately, the core weakness of purely technological approaches is that, while they may be helpful in mitigating some types of centralizing forces, such as network effects (via interoperability), they have a more difficult time addressing the centralization of decision-making itself.[19]

18. For more on this approach, see, for example, Brynjolfsson and McAfee (2011) and Brynjolfsson (2022).

19. For instance, thus far at least, attempts to use federated learning have been less successful in distributing learning than hoped for.

Economic approaches

Another set of approaches to combating the centralization of power draws on economics. For instance, by investing in broad-based education, a society can create a more widespread capacity for effective decision-making. The more people who have human capital—knowledge and skills—needed to make thoughtful decisions, the more likely it is that decision rights will be more widely distributed. Similarly, more widespread ownership of physical and financial capital could boost entrepreneurship and, with it, further distribute decision-making. Done right, these sorts of policies would not only broaden bargaining power, but also boost innovation, productivity, and growth—a double win. Historically, even when there have been significant economies of scale, the equilibrium in many markets has been to have two to three dominant firms rather than a single monopolist. And as Schumpeter (1942) noted, even monopolists can be toppled periodically through the process of creative destruction when new platforms and paradigms emerge. Strong support of new entrants and competition could put a check on the power of potential monopolies.

A complementary approach would seek to rein in organizations that are centralizing power, much as regulators have long sought to rein in natural monopolies in areas such as power generation or telephone services. This could be done with direct regulation of decisions and profits, as is often done with electrical utilities, or a more flexible approach such as the tax on digital ad revenues proposed by Romer (2019), with progressively higher rates on organizations with larger revenues in certain domains. Because the profits of a monopolist are higher when there is little competition, businesses often have an incentive to hinder interoperability or prevent data-sharing. This suggests a role for regulators to encourage or mandate data-sharing, just as anti-trust authorities intervene to maintain or increase competition.

One weakness of subsidizing decentralized decision-making or penalizing centralized decision-making is that these approaches may inefficiently undermine the benefits of scale. If centralized decision-making really does become increasingly superior at allocating resources, then societies that prevent this outcome may fall behind economically.

In that case, it might make sense to instead distribute economic power directly via a universal basic income (UBI) (see, e.g., Lowrey, 2018). If dollars are like votes directing economic decision-makers toward certain areas for the kinds of products and services that are in demand and the kinds of innovations that will be profitable, then basic income more broadly distributes these votes, and with them, the accompanying decision rights. In this way, even if the unfettered market would give less and less power to people who didn't own or control large ML systems, UBI could restore some of that power, at least in the economic sphere.

A basic income system could be combined with the goal of creating more widely distributed human capital if a portion of the payment were earmarked or conditional on skill development, turning the UBI into more of a conditional basic income. One could also amplify the ability of basic income to distribute political power by earmarking some of the funds specifically for political contributions. Widespread vouchers for campaign contributions would dilute the power of large corporations and wealthy individuals to amplify certain political messages or candidates.

While all these economic approaches have promise, they may also be vulnerable if those with economic power don't need the economic or decision-making contributions of others. That could lead them to re-write the rules to bolster their economic power. There's a real risk that economic concentration

of power would in turn lead to political concentration of power. This is a recurring problem with regulatory capture, as the regulators come under the sway of those they are supposed to regulate. *Quis custodiet ipsos custodes?*[20]

Political approaches

Ultimately, preserving and strengthening democracy may be the best counterbalance to the centralization of power in other spheres. While ML systems can concentrate decision-making in many realms, it's a political, not a technological or economic decision, to vest ultimate power in the people via democratic principles and institutions. The forces of markets and capitalism may find it more efficient to centralize more and more decision-making via large ML systems, and with it, create increasing disparities of wealth and power. But as Khosla (2017) notes, "capitalism is by permission of democracy and democracy should have the tools to correct for disparity."

The essence of democracy is one person, one vote—not one dollar, one vote. While the market rewards people and grants them decision-making power based on some calculus of their economic bargaining power, democracy treats people as ends, not means. Each individual has a voice regardless of whether they know valuable information, possess useful economic capabilities, control important assets, or even contribute at all.

However, simply establishing the institutions of a democracy, or its counterpart, a republic, is not sufficient to maintain or perpetuate a distributed governance.[21] There are myriad ways to undermine voting rights and access or to shift power away from some people or groups and toward others. An ML itself can be used to turbocharge anti-democratic techniques, for instance by using fine-grained data to predict voting preferences and then gerrymandering districts, tailoring marketing strategies, or curating candidates to induce the preferred outcome of those who control the relevant data and technology. Democracy also depends on a host of companion institutions and norms, such as a free press, freedom of association, and freedom to dissent. These may also be threatened by the use and abuse of increasingly powerful AI systems.

What's more, to make the democracy project even more challenging, the growing power of ML is increasingly not just a national matter, but one with international ramifications. Modern data networks and social networks are spanning international boundaries, making it possible for both state and non-state actors to extend their influence globally. The concentration of decision-making can affect organizations across political and geographic boundaries, so a successful system may need to include a global framework for creating fair rules.

If ML systems lead to increased centralization of decision-making in markets and firms, it will be even more important to work for democracy in the political sphere. It is the ultimate counterbalance to centralized power and is a safeguard for preserving individual liberty.

20. "Who watches the watchers?"

21. According to a possibly apocryphal account of the American constitutional convention, a lady asked Dr. [Ben] Franklin, "Well, Doctor, what have we got? a republic or a monarchy?" "A republic," replied the Doctor, "if you can keep it."

CONCLUSION

As long as there have been human organizations, there has been a tension between centralizing and decentralizing decision-making. While centralized decision-making can take into account interdependencies across units and thereby increase efficiency, it has historically had two big disadvantages: first, there are limits to computational power that prevent any one entity, human or machine, from making more than a finite subset of the possible decisions that need to be made, and second, no one entity has all the relevant knowledge or expertise.

With ever more powerful computers, ML systems, and data-gathering systems, these two constraints are less and less binding. In particular, while decision-making by humans is inherently decentralized by the computational, information gathering, and information sharing limits on the human mind, modern AI systems continue to improve on all these dimensions. For an increasingly large set of problems, big AI systems can take into account more information and make better decisions—not only better than those made by individual humans, but also those made by groups of humans working together.

The implications of this for the economy, our governance, and even the international order could be profound. For instance, in the 1950s through 1980s, there was an ongoing struggle between the economic and political systems of the West, which largely relied on the paired systems of distributed governance (democracy) and distributed ownership of the means of production (capitalism), vs. the Soviet system, which centralized both types of decision-making to a much greater extent. By 1989, it was clear that the former combination was the winner of this contest. While the prevailing storyline often emphasized the virtues of freedom and democracy, it was arguably the superior innovation capacity of free enterprise, especially when it came to high-tech innovation and overall wealth creation, that proved decisive.

Knowing the risks that centralized power can veer toward totalitarianism, even when it starts with noble goals, most of us are happy that the decentralized approach prevailed. But would the same outcome occur if the contest were re-run in 2030 or 2040? The decision-making capabilities of technology are already very different today than they were 40 years ago. What's more, they are on an express lane to ever larger and more powerful AI systems that can make decisions in ways that no centralized machine ever could have before.

We don't believe any outcome is inevitable. But that is no cause for complacency. The private incentives for centralizing decision-making will often exceed the social benefits of greater decentralization. Thus, we cannot necessarily count on the unfettered market to prevent increased centralization of decision-making, power, and wealth. As technology evolves, it is our responsibility to consider both the opportunities and the risks of increasingly powerful and centralized machine decision-making, and consciously work toward outcomes that support freedom and human flourishing.

REFERENCES

Acemoglu, D. 2021. *Redesigning AI*. Cambridge: MIT Press.

Acemoglu, D. and Robinson, J. A. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Currency.

Acton, J. 1887. Acton-Creighton Correspondence. Unknown.

Baradi, J. 2009. *Synco: El juego del revés* [Synco: *The Game of Reverse].* El Mercurio Revista de Libros. (In Spanish.)

Begenau, J. 2018. Big data in finance and the growth of large firms. *Journal of Monetary Economics*, Vol. 97, pp. 71–87.

Belton, P. 2021. Why coders love the AI that could put them out of a job. *BBC News.* September 7. https://www.bbc.com/news/business-57914432

Bostrom, N. 2016. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Boushey, H. and Knudsen, H. 2021. The importance of competition for the American economy. The White House. Blog. July 9. https://www.whitehouse.gov/cea/blog/2021/07/09/the-importance-of-competition-for-the-american-economy/

Brynjolfsson, E. 1994. Information assets, technology and organization. *Management Science*, Vol. 40, No. 12, pp. 1645–1662.

Brynjolfsson, E. 2022. The Turing trap: The promise & peril of creating human-like artificial intelligence. *Daedalus*. Spring issue. https://www.amacad.org/publication/turing-trap-promise-peril-human-artificial-intelligence

Brynjolfsson, E., Jin, W. and McElheran, K. S. 2021. The power of prediction: Predictive analytics, workplace complements, and business performance. *SSRN.* http://dx.doi.org/10.2139/ssrn.3849716

Brynjolfsson, E. and McAfee, A. 2011. *Race Against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Digital Frontier Press.

––––. 2017a. Artificial intelligence, for real. *Harvard Business Review*, pp. 1-31. https://store.hbr.org/product/artificial-intelligence-for-real/BG1704

––––. 2017b. What's driving the machine learning explosion? *Harvard Business Review.* The Big Idea Series. July 18. https://hbr.org/2017/07/whats-driving-the-machine-learning-explosion

Brynjolfsson, E., McAfee, A., Sorell, M. and Zhu, F. 2008. Scale without mass: Business process replication and industry dynamics. Harvard Business School Technology & Operations Mgt. Unit Research Paper, No. 07-016.

Brynjolfsson, E. and Mendelson, H. 1993. Information systems and the organization of modern enterprise. *Journal of Organizational Computing*, Vol. 3, No. 3, pp. 245–255.

Brynjolfsson, E. and Mitchell, T. 2017. What can machine learning do? Workforce implications. *Science*, Vol. 358, No. 6370, pp. 1530–1534.

Chang, C. 2020. 3 ways AI can help solve inventory management challenges. IBM Supply Chain Blog. March 4. https://www.ibm.com/blogs/supply-chain/3-ways-ai-solves-inventory-management-challenges/

Decker, R. A., Haltiwanger, J., Jarmin, R.S. and Miranda, J. 2020. Changing business dynamism and productivity: Shocks versus responsiveness. *American Economic Review*, Vol. 110, No. 12, pp. 3952–3990.

Deloitte. 2019. *Semiconductors – The Next Wave: Opportunities and Winning Strategies for Semiconductor Companies*. https://www2.deloitte.com/content/dam/Deloitte/tw/Documents/technology-media-telecommunications/tw-semiconductor-report-EN.pdf

Devries, P. M., Viégas, F., Wattenberg, M. and Meade, B. J. 2018. Deep learning of aftershock patterns following large earthquakes. *Nature,* Vol. 560, No. 7720, pp. 632–634.

DiSalvo, D. 2013. Your brain sees even when you don't. *Forbes*. June 22. https://www.forbes.com/sites/daviddisalvo/2013/06/22/your-brain-sees-even-when-you-dont/?sh=6c44bbe6116a

Dilliard, I. 1941. *Mr. Justice Brandeis: Great American.* The Modern View Press.

Douglass, E. 2002. Carriers aim to kill number portability. *Los Angeles Times*. January 16. https://www.latimes.com/archives/la-xpm-2002-jan-16-fi-cell16-story.html

Economist (The). 2021. Is America Inc getting less dynamic, less global and more monopolistic? September 18. https://www.economist.com/business/is-america-inc-getting-less-dynamic-less-global-and-more-monopolistic/21804757

Emerging Technology from the arXiv. 2009. New measure of human brain processing speed. *MIT Technology Review*. August 25. https://www.technologyreview.com/2009/08/25/210267/new-measure-of-human-brain-processing-speed/

Federal Communications Commission (USA). 1996. Telephone number portability. https://www.fcc.gov/document/telephone-number-portability-19

Federal Communications Commission (USA). 2009. Wireless local number portability. https://docs.fcc.gov/public/attachments/FCC-96-286A1.pdf

Frank, R. and Cook, P. 2013. Winner-take-all markets. *Studies in Microeconomics*, Vol. 1, No. 2, pp. 131–154.

GE Research. n.d. Predictive maintenance. https://www.ge.com/research/project/predictive-maintenance

Grossman, S. J. and Hart, O. D. 1986. The costs and benefits of ownership: A theory of lateral and vertical integration. *Journal of Political Economy*, Vol. 94, No. 4, pp. 691–719.

Hammond, P. 1997. The efficiency theorems and market failure. Preprint chapter. Kirman, A (ed.). *Elements of General Equilibrium Analysis*. Toronto: Wiley. 1998. http://web.stanford.edu/~hammond/effMktFail.pdf

Hart, O. 1989. An economist's perspective on the theory of the firm. *Columbia Law Review*, Vol. 89, No. 7, pp. 1757–1774.

Hart, O. and Moore J. 1990. Property rights and the nature of the firm. *Journal of Political Economy*, Vol. 98, No. 6, pp. 1119–1158.

Hayek, F. A. 1945. The Use of Knowledge in Society. *The American Economic Review*. Vol. 35, No. 4, pp. 519–530.

Jensen, M. and Meckling, W. 1992. Specific and general knowledge and organizational structure. L. Werin and H. Wijkander (ed.), *Contract Economics*. Oxford: Blackwell Publishers.

Khosla, V. 2017. AI: Scary for the right reasons. Khosla Ventures blog, September 12. https://www.khoslaventures.com/ai-scary-for-the-right-reasons

Krizhevsky A., Sutskever I. and Hinton G. E. 2012. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*., Vol. 60, No. 6, pp. 84–90.

Lange, O. 1936. On the economic theory of socialism. *The Review of Economic Studies*, Vol. 4, No. 1.

Lau, J. 2020. Google Maps 101: How AI helps predict traffic and determine routes. Google Blogs, September 3. https://blog.google/products/maps/google-maps-101-how-ai-helps-predict-traffic-and-determine-routes/

Lera, S. C., Pentland, A. and Sornette, D. 2020. Prediction and prevention of disproportionally dominant agents in complex networks. *Proceedings of the National Academy of Sciences*, Vol. 117, No. 44, pp. 27090–27095.

Lerner, A. 1938. Theory and practice in socialist economics. *Review of Economic Studies*, Vol. 6, No. 1.

Lowrey, A. 2018. *Give People Money: How a Universal Basic Income Would End Poverty, Revolutionize Work, and Remake the World*. New York: Broadway Books.

MacDonald, J. M., Ollinger, M.E., Nelson, K.E. and Handy, C. R. 1999. *Consolidation in U.S. Meatpacking*. Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture, Agricultural Economic Report No. 785.

Malone, T. 2003. The decentralization imperative. *MIT Technology Review*. October 24. https://www.technologyreview.com/2003/10/24/274985/the-decentralization-imperative/

Markowsky G. 2021. Physiology. Britannica. https://www.britannica.com/science/information-theory/Physiology

Marois, R. and Ivanoff, J. 2005. Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, Vol. 9, No. 9, pp. 415.

McAfee, A. and Brynjolfsson E. 2008. Investing in the IT that makes a competitive difference. *Harvard Business Review*. July. https://hbr.org/2008/07/investing-in-the-it-that-makes-a-competitive-difference

Morozov, E. 2014. The planning machine: Project Cybersyn and the origins of the Big Data nation. *New Yorker*. October 13. https://www.newyorker.com/magazine/2014/10/13/planning-machine

Ng, A., Madhavan, A. and Raina, R. 2009. Large-scale deep unsupervised learning using graphics processors. Proceedings of the 26th International Conference on Machine Learning.

––––. 2015. What data scientists should know about deep learning. Speech presented at Extract Data Conference, November 24. https://www.slideshare.net/ExtractConf/andrew-ng-chief-scientist-at-baidu

––––. 2016. Presentation given at Bay Area Deep Learning School.

––––. 2020a. *Landing AI Transformation Playbook 5*. https://landing.ai/wp-content/uploads/2020/05/LandingAI_Transformation_Playbook_11-19.pdf

––––. 2020b. State of AI. Presentation at AI Fund Update.

Nouwens, M., Liccardi, I., Veale, M., Karger, D. and Kagal, L. 2020. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. pp. 1–13.

Pentland, A., Nathan, O. and Zyskind, G. 2015. Enigma: Decentralized computation platform with guaranteed privacy. *arXiv:1506.03471*

Phelps, E. 2015. *Mass Flourishing: How Grassroots Innovation Created Jobs, Challenge and Change*. New Jersey: Princeton University Press.

Pueyo, T. 2021. Internet and blockchain will kill nation-states. Uncharted Territories blog, August 29. https://unchartedterritories.tomaspueyo.com/p/internet-blockchain-kill-nation-states

Reber, P. 2010. What is the memory capacity of the human brain? *Scientific American.* May 1. https://www.scientificamerican.com/article/what-is-the-memory-capacity/

Reich, R., Sahami, M. and Weinstein, J. M. 2021. *System Error: Where Big Tech Went wrong and How We Can Reboot*. New York: Harper Collins.

Rogers, K. 2017. More than 100 million Americans can only get internet service from companies that have violated net neutrality. *Vice*. December 11. https://www.vice.com/en/article/bjdjd4/100-million-americans-only-have-one-isp-option-internet-broadband-net-neutrality

Romer, P. 2019. A tax that could fix big tech. *The New York Times.* May 6. https://www.nytimes.com/2019/05/06/opinion/tax-facebook-google.html

Ronneberger, O., Fischer, P. and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention – MICCAI* 2015. pp. 234–241.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin Books.

Schumpeter, J. 1942. *Capitalism, Socialism and Democracy*. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship. New York: Harper & Brothers.

Sharma, R. 2019. Bitcoin won't win worldwide adoption because China controls it: Ripple CEO. *Investopedia*. June 25. https://www.investopedia.com/news/bitcoin-wont-win-worldwide-adoption-because-china-controls-it-ripple-ceo/

Shilov, A. 2021. Intel: Upcoming US fab will be a small city, to cost $60 to $120 billion. *Tom's Hardware*. August 6. https://www.tomshardware.com/news/intel-to-spend-up-to-120-billion-on-new-us-manufacturing-hub

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., (...) Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature.* Vol. 550, No. 7676, pp. 354–359.

Simon, H. A. 1955. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, Vol. 69, No.1, pp. 99–118.

Smith, A. 1776 (2008). *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Oxford University Press.

Srinivasan, B. 2019. Balaji Srinivasan on the argument for decentralization – Part 1. The Pomp Podcast, episode 295.

Stucke, M. E. and Ezrachi, A. 2017. The rise, fall, and rebirth of the U.S. antitrust movement. *Harvard Business Review*. December 15. https://hbr.org/2017/12/the-rise-fall-and-rebirth-of-the-u-s-antitrust-movement

Scheid, B. 2020. Top 5 tech stocks' S&P dominance raises fear of bursting bubble. S&P Global Market Intelligence blog. July 27. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/top-5-tech-stocks-s-p-500-dominance-raises-fears-of-bursting-bubble-59591523

Talagala, N. 2021. Google built a trillion parameter AI model. 7 things you should know. *Forbes*. July 8. https://www.forbes.com/sites/nishatalagala/2021/07/08/google-built-a-trillion-parameter-ai-model-7-things-you-should-know/?sh=4909de7b7974

Tambe, P., Hitt, L., Rock, D. and Brynjolfsson, E. 2020. Digital capital and superstar firms. National Bureau of Economic Research. NBER Working Paper Series. https://www.nber.org/papers/w28285

von Mises, L., 1951. *Socialism: An Economic and Sociological Analysis*. New Haven: Yale University Press. Ludwig von Mises Institute.

Werin, L. and Wijkander, H. (eds). 1992. *Journal of Applied Corporate Finance*, Vol. 8, No. 2.

Wilson, T. D. 2004. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge: Harvard University Press.

Yudkowsky, E. 2016. The AI alignment problem: Why it's hard, and where to start. Talk presented at Stanford University, May 5. Video. https://intelligence.org/stanford-talk/

RESOLVING DILEMMAS IN RESPONSIBLE ARTIFICIAL INTELLIGENCE DEVELOPMENT: A MISSING LINK DURING THE PANDEMIC

NATHALIE VOARINO

Postdoctoral fellow, Health Hub: Policy, Organizations and Law (H-POD), Université de Montréal Faculty of Law, International Observatory on the Societal Impacts of AI and Digital Technology (OBVIA – Fonds de recherche du Québec).

CATHERINE RÉGIS

Full professor, Université de Montréal, Canada Research Chair in Collaborative Culture in Health Law and Policy, researcher at Mila (Quebec Artificial Intelligence Institute) and at the International Observatory on the Societal Impacts of AI and Digital Technology, co-director of the Health Hub: Policy, Organizations and Law (H-POD).

SDG3 - Good Health and Well-being
SDG8 - Decent Work and Economic Growth
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG13 - Climate Action
SDG14 - Life Below Water
SDG15 - Life on Land
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

RESOLVING DILEMMAS IN RESPONSIBLE ARTIFICIAL INTELLIGENCE DEVELOPMENT: A MISSING LINK DURING THE PANDEMIC

ABSTRACT

Focusing on the dilemmas of AI ethics and how we can reconcile the potential tensions that emerge from principles implementation helps to identify some of the AI governance missing links. The COVID-19 pandemic constitutes a paradigmatic case for the study of these missing links, with several countries having chosen to rely on AI technologies to support ongoing public health efforts. Guided by the ten ethical principles of the Montréal Declaration for a Responsible Development of AI, in this text, we present two examples of key dilemmas emerging from the use of AI as part of the pandemic response, namely: 1) the dilemma between the protection of the privacy and intimacy principle and the solidarity principle; 2) the dilemma between the equity principle and the sustainable development principle. We then identify possible solutions, based on the capability approach, in order to address them. Resolving these dilemmas is essential if we are to harness the full potential of AI systems to fight pandemics—whether this one or the next ones—and ensure that AI is beneficial to everyone's health.

INTRODUCTION

The field of artificial intelligence (AI) has experienced major advances in the last decade, due, in part, to the sophistication of computer-based tools and the growing amount of available data (Cardon et al., 2018). Whether we are talking about health, human resources, the environment or education, the potential benefits of this new springtime for AI will not bypass any sector of society. However, the systemic implementation of increasingly autonomous AI systems (AIS) to automate repetitive tasks hitherto entrusted to humans raises a number of issues that have been widely identified in recent years. These include the risk to privacy (Stahl and Wright, 2018); the risk to social justice in relation to the potential biases perpetuated by algorithms (Kim, 2016; Risse, 2019); the risk of the dehumanization of activities

(Coeckelbergh, 2015)—especially considering the decrease in human supervision; or the risk of the erosion of accountability of AI users (Noorman, 2016) due to the lack of transparency of algorithmic decisions (fostered by the famous algorithm "black box") (Annanny and Crawford, 2018).

As a response to these risks and issues, several initiatives around the world have defined guiding ethical principles for the responsible development of AI: 84 documents were listed in 2019 (Jobin et al., 2019) and 167 were listed in 2020 (AlgorithmWatch, 2020), including the Montréal Declaration for a Responsible Development of AI (2018). Whether these documents have a national or international scope, their ethical principles are intended to guide AI governance, i.e. to guide the development of different mechanisms such as public policies, laws and regulations or technical standards (AIHLEG, 2019). It is noteworthy that there are different ways to implement the ethical principles of AI (AIHLEG, 2019). Namely, through technical methods (e.g. procedures included in algorithm architectures) or non-technical methods (e.g. legal mechanisms). In this chapter, we are mainly dealing with the second category.

While work done in recent years regarding the identification of issues and the definition of ethical principles has been considerable, a gap between these principles and their practical implementation remains difficult to overcome, as highlighted by several experts (Mittelstadt, 2019; Morley et al., 2020; Hagendorff, 2020; Shneiderman, 2020; Siau and Wang, 2020; Langlois and Régis, 2021). This gap is related namely to the fact that the principles are too abstract or too vague, which makes it difficult to interpret them in order to guide the development of the aforementioned mechanisms (Mittelstadt, 2019; Morley et al., 2020). The gap is also compounded by the difficulty of prioritizing one principle over another when they conflict (Whittlestone et al., 2019; Yeung et al., 2020), i.e. when facing ethical dilemmas.[22]

While this level of abstraction is inherent to ethical principles (Massé, 2003),[23] the emergence of dilemmas when it comes to implementing them is a major issue. How will these principles guide the development of policies or standards when they find themselves in contradiction? This issue could prompt the rejection or a lack of interest in the principles, leading to the risk of them no longer being considered relevant to guide action, thus considerably reducing their potential contribution. This risk of losing interest in ethics—namely due to the lack of a clear definition of the allocation of responsibility—is sometimes called 'ethics shirking', i.e. the risk of no longer using ethical practices because they are deemed to be ineffective in a given context (Floridi, 2019).

According to Whittlestone et al. (2019), identifying these dilemmas is indeed one of the essential next steps in AI ethics for an effective governance. In addition to the fact that it would contribute to bridging the gap between principles and practice, identifying dilemmas would make it possible to highlight situations in need of new solutions—regardless of their nature—where ethical principles alone are not enough to guide action (Whittlestone et al., 2019). Focusing on the dilemmas of AI ethics and on the way to reconcile the potential tensions that arise when principles are applied would help identify some of the missing links in AI governance.

While these dilemmas are observed during operationalization of principles in real-world situations, the COVID-19 pandemic constitutes a paradigmatic case in which to study them. Indeed, because crises require urgent action and often occur in situations of uncertainty, they limit the time and evidence available to assess the risks associated with new uses of AI (Tzachor et al., 2020; Cave et al., 2021).

22. An ethical dilemma emerges when the application of one principle (or upholding a value) hinders the implementation of another principle (or another value)—and neither of the conflicting principles stands above the rest because there are "good" arguments in favour of both alternatives (Durand, 2007).

23. Indeed, "the principles are designed to be sufficiently abstract to enable their sustainability and the flexibility of their interpretation with a view to broad appropriation (or even universal appropriation)" (Voarino, 2020 p. 182).

Moreover, deploying solutions (technological or other) at scale (local or international) increases the impact of unexpected harmful consequences (Tzachor et al., 2020; Cave et al., 2021) as well as of envisaged benefits.

Therefore, the pandemic can act as a powerful indicator of ethical dilemmas, especially those arising from the use of AI. Several AIS have indeed been identified and deployed to support ongoing public health efforts. Whether they intervene at the molecular (e.g. optimization of vaccine development), clinical (e.g. diagnosis support) or societal (e.g. epidemiological modelling) level (Bullock et al., 2020), they offer several promising perspectives in the fight against the spread of the virus. However, the use of AIS has raised many concerns relating namely to the protection of personal data, respect for citizen's consent and autonomy or to the infringement of various individual freedoms and fundamental human rights (Gasser et al., 2020; Naudé, 2020; Cave et al., 2021; von Struensee, 2021). In order to inform the responses to these concerns, it was particularly fitting to refer to the ethical principles that garnered so much attention before the start of the pandemic. However, several significant dilemmas emerge when consulting these principles to guide the responsible development of AI to limit the spread of COVID-19.

Guided by the ten ethical principles of the Montréal Declaration (2018), in this text we present two examples of key dilemmas that emerge from the use of AI to respond to the current pandemic. The Montréal Declaration, based on a co-construction process involving more than 500 citizens, has received scientific and international attention (Else, 2018; Fjeld et al., 2020) and has been identified as an important tool for the responsible development of AI (The Future Society, 2020). This exercise enables us to highlight some of AI ethics' missing links, resulting namely from the tendency of certain principles to eclipse others. We then identify possible solutions to address these missing links. We believe that resolving these dilemmas is essential to harnessing the full potential of AIS to fight pandemics—whether this one or the next ones, and ensure that AI benefits everyone.

TWO EXAMPLES OF KEY DILEMMAS RELATED TO THE USE OF AI TO RESPOND TO THE COVID-19 PANDEMIC

Solidarity in the Shadow of Privacy Protection

The first dilemma that emerged when using AI as part of the COVID-19 pandemic response is the one between protection of privacy and solidarity, which was especially controversial in the context of data sharing to implement public health surveillance mechanisms—e.g. via tracing apps.[24]

Irrespective of the current pandemic, the risk of privacy infringement is one of the major concerns associated with the advent of AI in healthcare (Christen et al., 2016; Azencott, 2018; Iyengar et al., 2018; Hager et al., 2019) and one of the most discussed issues in the literature of related fields such as big data (for example, see Mittelstadt and Floridi, 2016; or Stahl and Wright, 2018). Echoing a fundamental human right—present, for instance, in Article 12 of the Universal Declaration of Human Rights and the European Union's General Data Protection Regulation (GDPR)—Principle 3 of the Montréal Declaration, *Protection of privacy and intimacy,* advocates for data protection beyond the simple guarantee

24. These apps may relate to contact tracing or location tracking, which make it possible to identify users who represent a risk of contagion. Such risk is measured through the establishment of a contact history or through tracking the location of people who have tested positive, respectively (Mondin and Marcellis-Warin, 2020). This type of app is not always supported by AI systems and other AI systems are also likely to support public health surveillance mechanisms.

of personal data confidentiality and anonymity. It calls for the protection of "personal spaces in which people are not subjected to surveillance;" and stipulates that "every person must be able to exercise extensive control over their personal data, especially when it comes to its collection, use, and dissemination."

The reason that this principle has been undermined during the current pandemic is because the use of AI and, ultimately, its performance, are heavily dependent on access to individuals' data (Bullock et al., 2020).[25] Often, the data that is the subject of privacy-related discussions and concerns is the data collected outside of the healthcare system, such as from the Internet, social media, or smart phones (Mittelstadt and Floridi, 2016; Ienca and Vayena, 2020; Scassa et al., 2020; Kassab and Graciano Neto, 2021). Their collection does indeed allow surveillance into *personal spaces*, an issue that had already been raised by several experts before the pandemic when system portability exited the traditional spaces of care to introduce horizontal and ubiquitous, and potentially intrusive, health data collection even when it is anonymous or low sensitivity data (Mittelstadt and Floridi, 2016; IEEE, 2017; Villani, 2018). In addition, this type of collection *limits citizens' potential control* over their data, specifically with regard to what is done with it when it is reused, which then becomes almost infinite (Christen et al., 2016; Rial-Sebbag, 2017). On this point, infringement has been more or less significant depending on the country, namely depending on whether (or not) the use of these AIS or data collection is compulsory, but also based on the authorities' or digital tools' level of transparency regarding the purpose of using them (Mondin and de Marcellis-Warin, 2020). Thus, surveillance into personal spaces also risks infringing Montréal Declaration's Principle 2, Respect for autonomy—which stipulates namely that AIS must not be "developed or used to impose a particular lifestyle on individuals, whether directly or indirectly, by implementing oppressive surveillance and evaluation or incentive mechanisms;" and that "public institutions must not use AIS to promote or discredit a particular conception of the good life."

According to Mello and Wang (2020), although using health data for disease surveillance is not new, "several countries have taken digital epidemiology to the next level in responding to COVID-19" (p. 951) with large-scale data collection from millions of users (Ienca and Vayena, 2020). This phenomenon has raised privacy concerns in several countries around the world such as Canada (CEST, 2020); China (Ienca and Vayena, 2020; Mello and Wang, 2020; Shachar et al., 2020); the United States (Shachar et al., 2020) and Zimbabwe (Mbunge et al., 2021).

Concerns have been raised regarding the risk of falling into an excess of tracing and surveillance (for example, see Scassa et al., 2020; Mbunge et al., 2021; Tran and Nguyen, 2021; CEST, 2020) while Principle 3 advocates limiting the potential intrusion of AIS in people's lives when these systems are capable of "causing harm" as part of uses that "impose moral judgments on people or on their lifestyle choices" (Montréal Declaration, 2018). This aspect of the principle seems incompatible with the use of AIS to monitor adherence to public health measures, as was the case, according to Mello and Wang (2020), in China, Poland and Russia.

However, such AIS could help limit the spread of the virus by identifying the emergence of future clusters (Vaishya et al., 2020) or by helping to better understand patterns of viral spread (Alimadadi et al., 2020), thus making the implementation of public health measures more effective. This could accelerate the end of liberty-infringing measures such as confinement (Shachar et al., 2020) or the end of restricted access to education and to economic and cultural activities. Not using these AIS for the sake of privacy protection would then risk undermining Principle 4 of the Montréal Declaration, *Solidarity*. According to this principle, the development of AI must "be compatible with maintaining the bonds of solidarity among people and generations" and "improve risk management and foster conditions

25. In their literature review, Bullock et al. (2020) identified several useful datasets for the analysis of AI systems used to limit the spread of COVID-19, including data related to the number of cases or their location.

for a society with a more equitable and mutual distribution of individual and collective risks." In the case of the current pandemic, every effective measure aimed at limiting the spread of the virus does indeed foster *solidarity among generations* (e.g. with the elderly who were particularly impacted by the pandemic—Jackman, 2020; Lagacé et al., 2020) or among different groups (e.g. with essential workers who could not be assigned to telework). As for the *mutual distribution of risks*, it appears to encourage the sharing of data (whether personal or not) with a view to the aforementioned collective benefits (i.e. to improve everyone's health). Naudé (2020) has identified privacy concerns as one of the barriers to the effectiveness of AIS used as part of the pandemic response.

This tension between the protection of privacy and solidarity was raised in the literature whether authors were discussing management of the COVID-19 pandemic in general (e.g. in Colombia, see de la Espriella, Llanos and Hernandez, 2021) or the specific use of tracing apps (see Kudina, 2021). Nevertheless, the dilemma was already apparent before the pandemic, especially when it came to AI in healthcare. Indeed, some have argued that protection of privacy is outdated at a time when the sharing of (personal) data on social networks is ubiquitous (Spiekermann et al., 2018) and others contend that such breaches of privacy are justified during crises (O'Doherty et al., 2016; Fiore and Goodman, 2016). Thus, for some, in a public health context and considering the benefits for the common good, sharing data is a moral duty that justifies privacy infringement (Fiore and Goodman, 2016; Hand, 2018; Mello and Wang, 2020). Recently, Terry and Coughlin (2021) even proposed a "recalibration" of privacy protection based on solidarity considerations, which were observed in the context of the COVID-19 pandemic.

Thus, compliance with Principle 3, *Protection of privacy and intimacy*, the ethical importance of which no longer needs to be demonstrated, would hinder compliance with Principle 4, *Solidarity*, which advocates for sharing the (personal) data of the greatest number of people in order to enhance the health-related benefits of AIS for everyone.

Sustainable Development in the Shadow of Equity

The use of AI in a global health context also fosters tension between the moral duty to ensure everyone's access to technologies that support AIS (and to the ensuing health benefits) according to Principle 6, *Equity*, while limiting the environmental impact of these AIS in accordance with Principle 10, *Sustainable Development*.

According to the Montréal Declaration, Principle 6, *Equity*, requires that "the development and use of AIS must contribute to the creation of a just and equitable society." This implies namely that AIS must produce "social and economic benefits for all by reducing social inequalities and vulnerabilities;" that "access to fundamental resources, knowledge and digital tools" must be "guaranteed for all" and that it should support "the development of common algorithms—and of open data needed to train them."

In a global health context, AI has been identified (in the vein of digital health) as a particularly promising tool to achieve universal health coverage (Global Observatory for eHealth, 2015, 2016; WHO, 2018), echoing the aforementioned equity imperatives. Therefore, it would be a matter of providing (excluded or marginalized) populations and groups who have little or no access to technologies and infrastructure with tools that are capable of supporting AIS to ensure they have better access to healthcare and services. According to the World Health Organization (WHO), this means ensuring access to human and technical resources as well as to the required infrastructure, including electrification, Internet connectivity, wireless and mobile networks and devices (WHO, 2021a). This objective is part of a broader project on the international scene which aims to overcome the "digital divide," defined by WHO as "the uneven distribution of access to, use of or effect of information and communication technologies among any number of distinct groups" (WHO, 2021a, p. 34). Indeed, as recommended by the United Nations Secretary-General's High-level Panel on Digital Cooperation:

> By 2030, every adult should have affordable access to digital networks, as well as digitally enabled financial and health services, as a means to make a substantial contribution to achieving the Sustainable Development Goals (United Nations, 2019, in WHO, 2021.a, p. 34).

This digital divide is apparent between different countries around the world (Makri, 2019) as well as between different groups within the same society. Although the issue of a digital divide has existed for nearly a quarter of a century[26], its effects were exacerbated during the COVID-19 pandemic, when the use of digital technology became more widespread in healthcare as well as in other sectors (Davis, 2020; Ramsetty and Adams, 2020). For example, teleconsultations were preferred over face-to-face consultations to limit the spread of the virus. In this context, many people who did not have access to digital technologies and infrastructure (let alone to technologies and infrastructure likely to support AIS) were excluded from the available healthcare solutions, namely the elderly (Martins Van Jaarsveld, 2020), rural residents (Lai and Widmar, 2021) or people with limited income (News, 2020).

Responding to this digital divide would require providing tools to a non-negligible part of the world's population (if not the whole world in an ethical ideal) and would inexorably be accompanied by a greater number of technologies and infrastructure that are essential to AIS deployment and algorithm training. Because the latter are dependent on the amount of data available, this "global shift toward new digital technologies in health" (Davis, 2020) is also likely to be accompanied by an increase in the data generated, collected, stored and analyzed. For example, this would be the case with the creation of very large pandemic-specific datasets, such as the WHO Hub for Pandemic and Epidemic Intelligence project (WHO, 2021b). This initiative should take the form of a global platform for the collection and analysis of data that can be useful for the prevention and management of future pandemics, with the objective namely of overcoming state restrictions relating to confidentiality and protection of privacy to ensure relevant and effective data sharing for the common good.

However, digital technologies are not environmentally neutral. In addition to the significant level of electronic waste associated with digital innovation (Dwivedi et al., 2022), the operation of data centres, as well as the production of computers and smartphones, consume a significant amount of energy and could contribute significantly to global warming (Gmach et al., 2010; The Shift Project, 2020). The training of AI models is also increasingly associated with greenhouse gas (GHG) emissions (Ligozat et al., 2021).

The energy transition is also associated with other harmful consequences to the environment. While the digitalization of operations is sometimes considered a solution to reduce these GHG emissions (Patsavellas and Salonitis, 2019; Ghobakhloo, 2020; IEA, 2021), it is well known that this energy transition requires a lot of critical minerals and rare earth elements (European Commission, 2020; Hund et al., 2020; IEA, 2021). Smartphones, like other computing devices supporting AIS, require, among others, the mining of lithium, which is largely used in the development of batteries but has disastrous consequences on ecosystems (Crawford, 2021; IEA, 2021). As stated in the International Energy Agency 2020 Report, such mining activities: 1) can impact biodiversity and result in the loss of animal habitats (especially endangered species); 2) require large volumes of water (which is unsustainable in a context of water scarcity); 3) can lead to acidic wastewater contamination, and; 4) generate hazardous waste that can increase with declining ore quality (IEA, 2021).

Complying with Principle 3 of the Montréal Declaration, *Equity*, could in turn hinder compliance with Principle 10, *Sustainable Development*, which namely requires that the development and use of AIS be carried out "so as to ensure strong environmental sustainability of the planet." Among other things, this means that it must "mitigate greenhouse gas emissions," "aim to generate the least amount of electric and electronic waste" and "minimize our impact on ecosystems and biodiversity." Although

26. The term "*digital divide*" was first used in the United States in 1995 (Dijk, 2020).

several possible solutions are emerging to limit the environmental consequences of digital technologies (The Shift Project, 2020; IEA, 2021), many experts question whether the digital and ecological transitions are compatible[27] and whether these solutions are sufficient and effective in the short term, considering the urgency for climate action (IPCC, 2021). Recently, during the United Nations Climate Change Conference (COP26) in 2021, several experts questioned the extent to which digital technologies can contribute to the climate change response or whether they are an integral part of the problem (Dwivedi et al., 2022).

The dilemma around sustainable development and equity is all the more important in the context of global health, when compliance with principles of sustainable development is directly linked to the population's health (Patz et al., 2014; Solomon and LaRocque, 2019). Degradation of the environment and biodiversity as well as global warming could foster the emergence of new pandemics (Mackenzie and Jeggo, 2019; Solomon and LaRocque, 2019; Charlier et al., 2020; Hébert, 2021). These concerns are at the heart of the Manhattan Principles, developed in 2004 during a symposium that gathered international experts to consider, among other things, the prevention of the emergence of infectious diseases such as zoonotic diseases (Manhattan Principles, 2004). These principles advocate for a comprehensive approach linking environmental and health concerns, focusing on the notion of "one world, one health" (Manhattan Principles, 2004). The importance of these issues led several international experts to write an open letter to the WHO taking stock of the health-related consequences of global warming (including the risk of a pandemic) and urging (international) organizations to focus on this problem (see Charlier et al., 2020).

In light of this, equity and sustainability seem difficult to reconcile, particularly in a global health context. While a reconciliation is partly the objective of the UN's Sustainable Development Goals, the latter perpetuate this dilemma through potentially contradictory indicators when it comes to digital technology. For example, how do we reconcile Indicator 5.b.1.—of increasing the proportion of individuals who own a mobile telephone, by sex—with Target 12.2.—of achieving the sustainable management and efficient use of natural resources) —(United Nations, 2021)? In the context of the Montréal Declaration, compliance with Principle 10, *Sustainable Development*, thus comes up against the (highly commendable) objective of overcoming the digital divide in accordance with Principle 6, *Equity*.

OVERCOMING ETHICAL DILEMMAS THROUGH THE PRISM OF THE CAPABILITY APPROACH

Addressing the dilemmas that have arisen from the use of AIS as part of the COVID-19 pandemic response is especially relevant to informing ethical governance in global health. This topic is being neglected by the AI ethics research community (Murphy et al., 2021). These dilemmas are in the crosshairs of a classic dilemma in this field of action, namely: "How to balance the needs of 'the many' against the rights of 'the individual'" (Stapleton et al., 2014, p. 4) or, in other words, how to reconcile the health of individuals with that of the community. The boundary between the two is not always impermeable, as the protection of individual rights can obviously contribute to the achievement of collective objectives. This being said, in the examples of dilemmas presented here, we do find principles with individual dimensions that echo fundamental rights (i.e. Principle 3, *Protection of privacy and intimacy*, or Principle 6, *Equity*), that conflict with principles guided by objectives that fall under more collective considerations (i.e. Principle 4, *Solidarity*, and Principle 10, *Sustainable development*).

27. For example, see publications on the subject by the "Chemins de transition" project: https://cheminsdetransition.org/numerique/

In public health, this is a recurring dilemma that sets "individualistic ethics" nurtured by traditions of autonomy and individual rights against a more collective ethic, based on the common good and solidarity (Kenny et al., 2010).[28] This tension between individual and collective interests is at the root of the ethical issues that arise from the use of AI in healthcare (Voarino, 2020). It has been exacerbated during the COVID-19 pandemic (Anand et al., 2020) as mentioned by Biggeri (2020, p. 277):

> We have been willing to renounce (individual) freedom of movement and association to preserve the health and longevity of the most vulnerable. We realise that public health systems and governance need to pay far greater attention to collective and individual well-being.

Resolving such dilemmas requires, in part, that we focus on balancing the individual and collective aspects of concerns related to the use of AI as part of the pandemic response[29]. To reflect on the achievement of this balance, we believe that the capability approach may be an interesting path to explore.

The capability approach stems from the work of Amartya Sen who challenged traditional economic indicators to assess human development (Sen, 1983). According to this approach, development is not measured in terms of the possession of resources or income, but rather in terms of what individuals are actually able to do and be, i.e. in terms of their capabilities (Oosterlaken, 2015, summarizing several studies by Sen and Nussbaum). This approach has frequently been used in the context of development, namely by international organizations such as the 2020 United Nations Development Program (UNDP, 2020). The capability approach is particularly relevant to technology assessment. The Appropriate Technology Movement (ATM) supports this belief (Oosterlaken, 2015). Based on the capability approach, the ATM is driven by the following fundamental question with respect to technology assessment: "Do such initiatives truly empower people—in all their human diversity—to lead the lives they have reason to value?" (Oosterlaken, 2015, p. 41). In other words, according to the ATM, appropriate technology development should ensure the expansion of human capabilities (Oosterlaken, 2015).

First, the capability approach is relevant because it enables us to focus on issues relating to the pandemic's management which introduced a significant loss of capabilities across many aspects of life (Anand et al., 2020; Biggeri, 2020). According to Anand et al. (2020), basic capabilities such as health, education, nutrition and social ties have been compromised during the COVID-19 pandemic. Whether through individual choices or government decisions, "many populations have had to give up certain freedoms temporarily to protect other freedoms that they have reason to value" (Anand et al., 2020, p. 294).

Second, the capabilities approach makes it possible to embrace and go beyond the binary opposition between the individual and collective dimensions of the dilemmas presented here. Although the capabilities approach has also sometimes been criticized for its individualistic emphasis, many argue that it enables us to consider social well-being as an organized production of collective well-being (Doucin, 2009) or as a collective responsibility towards individual freedoms (Fusulier and Sirna, 2010). A hindrance to equity or privacy, in a context of global health, could also hinder populations' collective well-being, thus opposing collective dimensions to each other. We believe that the capability approach makes it possible to go beyond a distributive approach to conflict resolution—which aims to resolve the opposition of ideas by choosing a solution proportional to the balance of power or merit. It also allows

28. For example, this ethical dilemma has been widely discussed in the context of mandatory vaccination, independently of the current pandemic (e.g. see Krantz, Sachs and Nilstun, 2004; Dawson, 2015; Boas, Rosenthal and Davidovitch, 2016; Sim, 2017).

29. This requires going beyond the simple prioritization of principles (i.e., favouring one principle over another) though it is possible in certain situations, as mentioned in the Montréal Declaration (2018): There is no hierarchy among the principles; however, it is possible, depending on the circumstances, to lend more weight to one principle than another as long as 'the interpretation [is] coherent' (Montréal Declaration, 2018).

us to reflect on the dilemmas through an integrative approach—which seeks to define a common unifying standard of arbitration that creates additional value for the two ideas initially in tension. The capability approach identifies the resulting common standard as the one that will increase human capabilities.

According to the capability approach, we must consider at least two dimensions to ensure that resources (in this case, AIS) are converted by individuals into effective "functionings" (i.e. what is actually done or achieved by individuals): On the one hand, these resources must introduce actual additional opportunities and, on the other hand, individuals must be free to access them and choose to do so (Bonvin and Farvaque, 2007; Fusulier and Sirna, 2010).

Regarding the first dimension, i.e. the actual additional opportunities introduced by AIS, we must mention that several experts have pointed out that few of the AIS developed to limit the spread of COVID-19 have actually been effective (Naudé, 2020; Wynants et al., 2020; Douglas Heaven, 2021). Their potential was limited (depending on the type of AIS involved) by various factors such as insufficient data, poor quality data (not timely or insufficiently robust), models with high risk of bias, inability to be used by laypeople or in resource-limited settings, and ethical and legal limitations (Chen and See, 2020; Naudé, 2020; Wynants et al., 2020). The majority of AIS used as part of the pandemic response were in the early stages of development, not advanced enough for use in real-world settings—especially, in clinical settings—thus limiting their scope (Gunasekeran et al., 2021; Hashiguchi et al., 2022; Bullock et al., 2020). According to the WHO, the actual impact of AIS used as part of the COVID-19 pandemic response has, for the moment, been "modest" (WHO, 2021a).

Therefore, in the context of the dilemmas presented here, guaranteeing that AIS will actually be resources that introduce opportunities requires focusing above all on the means of overcoming the barriers to their effectiveness. Otherwise, they would contribute little or nothing to the achievement of effective gains in solidarity or equity—thus allowing concerns regarding hindrances to privacy and sustainable development to justify a possible restriction on the uses of AIS. As the ATM presupposes that not all technologies represent progress in themselves (Oosterlaken, 2015), it is also important to take into account the hype surrounding the development of AIS (Gibert, 2019). Such hype could lead to overestimating the benefits AIS can bring. In addition, more appropriate alternatives could end up overlooked if AIS' use proves to be premature.

As for the second dimension of individuals' freedom to access AIS and choose them or not, assessing the translation of real capabilities into effective functionings requires an identification of the choices that individuals have actually made—as well as of the values and preferences that motivated those choices. According to the capability approach (used in the context of the ATM): Resources (technologies) are translated into real capabilities or freedoms through "conversion factors," which are essential preconditions for expanding capabilities, whether these conditions are environmental, social or cultural (Bonvin and Farvaque, 2007; Oosterlaken, 2015). These real capabilities or freedoms are translated into effective functionings when individuals choose to use them (namely according to their preferences, once the opportunity exists) (Bonvin and Farvaque, 2007; Oosterlaken, 2015). Among other things, this requires identifying citizens' expectations and fears regarding the use of the various AIS developed to fight the pandemic, but also to assess the actual use of these AIS once the opportunity of using them is introduced, and the reasons for low user adoption. This is especially relevant in the context of using AI in healthcare, as several factors that may affect healthcare professionals' trust in these devices have been identified (directly impacting their appropriation and use in clinical settings) (Asan et al., 2020). Furthermore, European surveys have shown that not all citizens are ready to use a contact tracing app due to privacy and security concerns and skepticism about their effectiveness (Craglia et al., 2020). Several of the countries that used this type of app on a voluntary basis also observed a low adoption rate (e.g. 16% of the population of Singapore and 4% of the Australian population in April 2021) (Akinbi and al., 2021). The example of tracing apps is particularly relevant, even if they are not all AI-based. This is because their effectiveness is highly dependent on citizens' propensity to use them: it is estimated

that 50% to 70% of the population must use it for the application to be effective (Akinbi et al., 2021). Furthermore, while the number of people who install the app is an indicator, it is not enough. For example, with 1.9 million downloads, the French application had only sent 14 notifications by August 2020 (Akinbi et al., 2021).

However, this second dimension of individuals' freedom and choice comprised in the capability approach invites us to consider certain avenues to resolve dilemmas. Namely, regarding the aspects that can lead individuals to choose whether or not to use AIS. As the ATM (in accordance with the capability approach) gives special importance to individual diversity, it requires the participation of affected populations in the development of technological solutions (Oosterlaken, 2015) in order to embrace human diversity and, therefore, the diversity of preferences. As Doucin (2009) acknowledges:

> Developing capabilities is not only providing training [...] it is initiating a dialogue [...] with identified populations, by addressing groups, while ensuring that individuals are not subdued, to then build policy-related tools with them (p. 447).

This is particularly relevant with regard to potential privacy infringement in the name of solidarity-related considerations, as the type of data collected, the purposes for which they are collected and the actors who have access to them have changed as part of using AIS in the pandemic response: This implies a form of renegotiation of the social contract concerning health data. Prior to the current pandemic, the French *Comité consultatif national d'éthique* had highlighted various disruptions between the management of traditional health data and the advent of big data in health, namely: a change of scale, conservation time, rapid dissemination beyond medical teams and borders (CNNE, 2019). This disruption was accentuated by the pandemic. For example, with the collection of data pertaining to citizens' geolocation and movements for health purposes, or with the use of data generated on social media for public decision-making (such as the analysis of sentiment towards vaccination—see Wilson and Wiysonge, 2020). More traditional health data (e.g. a diagnosis) were no longer collected solely to treat the particular patient but also for other purposes, such as placing the patient in confinement.

The ATM and the capability approach also require that we pay particular attention to social inequalities that might influence the conversion of a resource into an effective functioning for everyone, beyond the simple creation of resources or means (Fusulier and Sirna, 2010; Oosterlaken, 2015). This can lead to questioning the contextual relevance of providing every person with digital technology in the name of equity, including those who do not have access to the basic resources necessary for survival. Is digital access a priority or relevant in all contexts? Addressing inequalities also requires that infringing on intimacy and privacy be justified only if it results in a real solidarity-related gain, and this applies to all those concerned by this infringement.

However, beyond the simple digital divide, inequalities persist in terms of sharing the benefits of data analysis. There remains a significant asymmetry between the people who collect, store and use big data and those who generate the data or are targeted by the data collection. This phenomenon is called the "big data divide" (Andrejevic, 2014; McCarthy, 2016). It has also been pointed out that, by exacerbating pre-existing inequalities, the COVID-19 pandemic has had far more harmful consequences on precarious population groups—especially on their capabilities (Biggeri, 2020). The populations of southern countries are also the most immediately and significantly affected by the consequences of global warming (Goodman, 2009) or of the environmental degradation resulting from the extraction of critical minerals. For example, the main suppliers of the elements required to develop digital technologies are China (41%) and African countries (30%) (European Commission, 2020). Europe is also largely dependent on South-East Asia for high-tech components and assembly (European Commission, 2020). The big data divide may lead to questioning the actual equity-related gains of AIS for population groups excluded from digital technology, if they are the ones who suffer the most from the environmental consequences of AIS development.

Finally, Sen (2013) and Dubois (2006) suggest rethinking the impact of increased capabilities on sustainability. The capability approach enables us to consider the imperatives of equity not only between countries or groups that have more or less access to digital technology, but also between human generations. Thus, sustainable development is understood in terms of intergenerational equity, aiming to ensure that future generations have access to at least the same capabilities as current generations (Dubois, 2006). According to Sen (2013), maintenance of capabilities between generations should not be limited to the maintenance of "our ability to fulfil our felt needs," but should rather aim at "sustaining human freedoms." From this perspective, Principle 3, *Equity*, is no longer in conflict with Principle 10, *Sustainable Development*, but is an integral part of it, broadening the scope of the United Nations' "leave no one behind" principle to include future generations (United Nations, n.d.).

CONCLUSION

The implementation of ethical principles to guide responsible development of AI as part of the COVID-19 pandemic response reveals the existence of several dilemmas. Drawing from the principles of the Montréal Declaration, two key dilemmas were highlighted. First, compliance with Principle 3, *Protection of privacy and intimacy*, could hinder compliance with Principle 4, *Solidarity*, which calls for sharing the personal data of the greatest number of people. Second, compliance with Principle 10, *Sustainable Development*, comes up against the objective of overcoming the digital divide in accordance with Principle 6, *Equity*. The benefit of resolving these dilemmas is twofold. With respect to AI governance, resolving the dilemmas could help prevent potential ethical rejection or disinterest by ensuring greater coherence of existing guidelines. With regard to global health, the resolution of dilemmas is necessary in order to ensure responsible development of AI in healthcare and, therefore, best contribute to the management of future pandemics.

The capability approach is presented as a promising way to overcome the binary dilemmas explored in this chapter. According to this approach, we must consider at least two dimensions to ensure the development of AIS that enable individuals, in all their diversity, to lead the lives they value (Oosterlaken, 2015). On the one hand, we must assess the extent to which the AIS introduce actual additional opportunities, which requires going beyond the current limitations with regard to their effectiveness and eventually considering other alternatives—without which it is not possible to ensure real solidarity—(Principle 4) or equity-related gains from AIS (Principle 6). On the other hand, we must focus on the conditions that enable individuals to choose whether or not to use AIS. This requires involving citizens, identifying their preferences, and taking into account the context in which AIS are implemented—namely, pre-existing inequalities between different groups and between current and future generations—in order to collectively define expectations related to privacy (Principle 3) and sustainability (Principle 10).

While we recognize that this is only a first level of analysis, it nevertheless invites us to state or reiterate the importance of a few possible solutions, whether they aim to resolve the dilemmas presented here or, generally, to shed light on potential missing links in AI ethics:

Encourage funding and research efforts prior to the large-scale deployment of AIS (whether in response to the current pandemic or to future pandemics). Research should focus primarily on: (i) the limitations encountered by AIS used as part of the COVID-19 pandemic response and; (ii) the impact of the environmental consequences of digital technology on people's health, at the different stages of their life cycle. Acquired knowledge regarding these limitations and impact should make it possible to increase the number of effective opportunities introduced by AIS, thereby fostering the capabilities of current and future populations.

Systematize the implementation of co-construction, in conjunction with citizens, of digital solutions and public policies relating to the development of AIS in a global health context. More than simple consultation, co-construction involves active participation by populations and is essential to appropriate technology development. This would align the development of AIS with citizens' values and preferences, which are key dimensions of the capability approach. If a proportionate infringement on individual rights and freedoms is justified in the name of the common good, it is essential to collectively assess the form that this common good should take. As part of co-construction efforts, special attention must be paid to local populations as well as to marginalized and excluded groups within the societies concerned, and a two-way exchange between northern and southern countries must be ensured.

Choose AIS relevance by-design. Consider options other than digital technology when it does not represent a means of increasing real capabilities, in order to achieve a sustainable balance. This requires questioning the achievement of equity to overcome the digital divide solely by increasing access to digital technology for populations who have little or no access to them, and to consider, for example, measures aimed at limiting the overconsumption of digital technology, especially in northern countries.

Promote global approaches to respond to AI-related issues, especially in a global health context. This implies limiting silo-, project-, program-, or discipline-based approaches, which are conducive to missing links (e.g. addressing the digital divide on one side and sustainable development on the other). A global approach also requires not limiting oneself to a local conception of the described issues. One must consider the globalization of exchanges and digitization, the diffuseness of AI regarding governance[30] or the cross-border and cross-sectoral nature of pandemics.

It is essential to implement these different mechanisms for the medium and long term, even amid the urgency associated with crises such as pandemics. We believe they would help create viable solutions from a "one world, one health" perspective and ensure that AI benefits everyone.

30. Or the "diffuseness problem," described by Danaher (2015) as "the problem that arises when AI systems are developed using teams of researchers that are organisationally, geographically, and perhaps more importantly, jurisdictionally separate" enabling them to evade a country's regulations by taking advantage of this jurisdictional diffusion.

REFERENCES

AIHLEG (High-Level Expert Group on Artificial Intelligence). 2019. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

Akinbi, A. Forshaw, M. and Blinkhorn, V. 2021. Contact Tracing Apps for the COVID-19 Pandemic: A Systematic Literature Review of Challenges and Future Directions for Neo-Liberal Societies. *Health Information Science and Systems*, Vol. 9, No. 1, 18. https://doi.org/10.1007/s13755-021-00147-7.

AlgorithmWatch. 2020. AI Ethics Guidelines Global Inventory, https://inventory.algorithmwatch.org/

Alimadadi, A. *et al.* 2020. Artificial intelligence and machine learning to fight COVID-19. *Physiological Genomics*, Vol. 52, No. 4, pp. 200–202. doi:10.1152/physiolgenomics.00029.2020.

Anand, P., Ferrer, B. Gao, Q., Nogales, R. and Unterhalter, E. 2020. COVID-19 as a Capability Crisis: Using the Capability Framework to Understand Policy Challenges. *Journal of Human Development and Capabilities*, Vol. 21, No. 3, pp. 293–299. doi:10.1080/19452829.2020.1789079.

Ananny, M. and Crawford, K. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, Vol. 20, No. 3, pp. 973–989. doi:10.1177/1461444816676645.

Andrejevic, M. 2014. Big Data, Big Questions| The Big Data Divide. *International Journal of Communication*, Vol. 8, No. 0, p. 17.

Asan, O., Bayrak A. E., and Choudhury, A. 2020. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, Vol. 22, No. 6, e15154. https://doi.org/10.2196/15154

Azencott C.-A. 2018. Machine learning and genomics: precision medicine versus patient privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 376, No. 2128, p. 20170350. doi:10.1098/rsta.2017.0350.

Biggeri, M. 2020. Introduction: Capabilities and Covid-19. *Journal of Human Development and Capabilities*, Vol. 21, No. 3, pp. 277–279. doi:10.1080/19452829.2020.1790732.

Boas, H., Rosenthal, A. and Davidovitch, N. 2016. Between individualism and social solidarity in vaccination policy: the case of the 2013 OPV campaign in Israel. *Israel Journal of Health Policy Research*, Vol. 5, No. 1, p. 64. doi:10.1186/s13584-016-0119-y.

Bonvin, J.-M. and Farvaque, N. 2007. L'accès à l'emploi au prisme des capabilités, enjeux théoriques et méthodologiques. *Formation emploi. Revue française de sciences sociales*, Vol. 98, pp. 9–22. doi:10.4000/formationemploi.1550.

Bullock, J. *et al.* 2020. Mapping the landscape of Artificial Intelligence applications against COVID-19. *Journal of Artificial Intelligence Research*, Vol. 69, pp. 807–845. doi:10.1613/jair.1.12162.

Cardon, D., Cointet, J.-P. and Mazières, A. 2018. La revanche des neurones. *Reseaux*, Vol. 211, No. 5, pp. 173–220.

Cave, S. *et al.* 2021. Using AI ethically to tackle covid-19. *BMJ (Clinical research ed.)*, No. 372, p. 364. doi:10.1136/bmj.n364.

CCNE. 2019. Données massives (big data) et santé: une nouvelle approche des enjeux éthiques. Avis 130. Comité Consultatif National d'éthique français. https://www.ouvrirlascience.fr/wp-content/uploads/2019/06/CCNE_Donnees-massives-et-sant%C3%A9_avis130_29mai2019.pdf.

Charlier, P. *et al.* 2020. Global warming and planetary health: An open letter to the WHO from scientific and indigenous people urging for paleo-microbiology studies. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, Vol. 82, p. 104284. doi:10.1016/j.meegid.2020.104284.

Chen, J. and See, K.C. 2020. Artificial Intelligence for COVID-19: Rapid Review. *Journal of Medical Internet Research*, Vol. 22, No. 10, p. e21476. doi:10.2196/21476.

Christen, M. *et al.* 2016. On the Compatibility of Big Data Driven Research and Informed Consent: The Example of the Human Brain Project. in Mittelstadt, B.D. and Floridi, L. (eds). *The Ethics of Biomedical Big Data*. Cham: Springer International Publishing (Law, Governance and Technology Series), pp. 199–218. doi:10.1007/978-3-319-33525-4_9.

Coeckelbergh, M. 2015. Artificial agents, good care, and modernity. *Theoretical Medicine and Bioethics*, Vol. 36, No. 4, pp. 265–277. doi:10.1007/s11017-015-9331-y.

European Commission. 2020. *Critical raw materials for strategic technologies and sectors in the EU – A foresight study*. Luxembourg: Publications Office of the European Union. https://ec.europa.eu/docsroom/documents/42881.

Craglia M., de Nigris S., Gomez-Gonzalez E., Gomez E., Martens B., Iglesias Portela M., Vespe M., Schade S., Micheli M., and Kotzev A. 2020. Artificial Intelligence and Digital Transformation: early lessons from the COVID-19 crisis. *JRC Science for policy report. Publications Office of the European Union*.

Crawford, K. 2021. *Atlas of AI*. New Haven: Yale University Press.

Danaher, J. 2015. Philosophical Disquisitions: Is effective regulation of AI possible? Eight potential regulatory problems. *Philosophical Disquisitions*, http://philosophicaldisquisitions.blogspot.com/2015/07/is-effective-regulation-of-ai-possible.html

Davis, S.L.M. 2020. The Trojan Horse. *Health and Human Rights*, Vol. 22, No. 2, pp. 41–47.

Dawson, A.J. 2015. Ebola: what it tells us about medical ethics. *Journal of Medical Ethics*, Vol. 41, No. 1, pp. 107–110. doi:10.1136/medethics-2014-102304.

Montréal Declaration. 2018. *Montréal Declaration for a Responsible Development of Artificial Intelligence*. University of Montréal. https://www.declarationmontreal-iaresponsable.com/la-declaration.

Dijk, J. van. 2020. The Digital Divide. *John Wiley & Sons*. ISBN: 978-1-5095-3446-3.

Doucin, M. 2009. Review of Repenser l'action collective: une approche par les capabilités, (Réseau Impact, coll. « Éthique économique »). *Revue Tiers Monde*, Vol. 50, No. 198, pp. 444–448.

Douglas Heaven, W. 2021. Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

Dubois, J.-L. 2006. Approche par les capabilités et développement durable: La transmission intergénérationnelle des capabilités. *Amartya Sen: un économiste du développement*, pp. 201–213.

Dwivedi, Y. K., Hughes, L. Kar, A.K., Baabdullah, A.M., Grover, P. Abbas, R. Andreini, D. et al. 2022. Climate Change and COP26: Are Digital Technologies and Information Management Part of the Problem or the Solution? An Editorial Reflection and Call to Action. *International Journal of Information Management*, Vol. 63, 102456. https://doi.org/10.1016/j.ijinfomgt.2021.102456

El-Sayed, A. and Kamel, M. 2020. Future threat from the past. *Environmental Science and Pollution Research International*, pp. 1–5. doi:10.1007/s11356-020-11234-9.

Else, H. 2018. Europe's AI researchers launch professional body over fears of falling behind. *Nature*. doi:10.1038/d41586-018-07730-1.

de la Espriella, F.R.M., Llanos, A.Z.B. and Hernandez, J.C. 2021. Privacy as a human right and solidarity as a constitutional value in the era of Covid-19. *Juridicas Cuc*, pp. 17–17.

Fiore, R.N. and Goodman, K.W. 2016. Precision medicine ethics: selected issues and developments in next-generation sequencing, clinical oncology, and ethics. *Current Opinion in Oncology*, Vol. 28, No. 1, pp. 83–87. doi:10.1097/CCO.0000000000000247.

Fjeld, J. *et al.* 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* [Preprint], (2020–1).

Floridi, L. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology*, Vol. 32, No. 2, pp. 185–193. doi:10.1007/s13347-019-00354-x.

Fusulier, B. and Sirna, F. 2010. Contrer les inégalités du "pouvoir d'agir", augmenter les capabilités. *Les Politiques Sociales*, Vol. 3-4, No. 2, pp. 33–38.

Gasser, U. *et al.* 2020. Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. *The Lancet Digital Health*, Vol. 2, No. 8, pp. e425–e434. doi:10.1016/S2589-7500(20)30137-0.

Ghobakhloo, M. 2020. Industry 4.0, digitization, and opportunities for sustainability. *Journal of Cleaner Production*, Vol. 252, p. 119869. doi:10.1016/j.jclepro.2019.119869.

Gibert, M. 2019. Faut-il avoir peur de la peur de l'IA? *La Quatrième Blessure*. https://medium.com/@martin.gibert/faut-il-avoir-peur-de-la-peur-de-lia-1687abc35342.

IPCC (Intergovernmental Panel on Climate Change). 2021. Sixth Assessment Report, Climate Change 2021: The Physical Science Basis. https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf.

Global Observatory for eHealth. 2015. Atlas of eHealth country profiles: the use of eHealth in support of universal health coverage. *World Health Organization.* https://www.who.int/publications/i/item/9789241565219

Global observatory for eHealth. 2016. Global diffusion of eHealth: Making universal health coverage achievable. *World Health Organization*. http://library.health.go.ug/download/file/fid/2620

Gmach, D. *et al.* 2010. Profiling Sustainability of Data Centers in *Proceedings of the 2010 IEEE International Symposium on Sustainable Systems and Technology*. pp. 1–6. doi:10.1109/ISSST.2010.5507750.

Goodman, J. 2009. From Global Justice to Climate Justice? Justice Ecologism in an Era of Global Warming. *New Political Science*, Vol. 31, No. 4, pp. 499–514. doi:10.1080/07393140903322570.

Gunasekeran, D. V., Wei Wen Tseng R. M., Tham Y-C., and Wong T. Y. 2021. Applications of Digital Health for Public Health Responses to COVID-19: A Systematic Scoping Review of Artificial Intelligence, Telehealth and Related Technologies. *Npj Digital Medicine*, Vol. 4, No. 1, pp. 1-6. https://doi.org/10.1038/s41746-021-00412-9.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, Vol. 30, No. 1, pp. 99–120. doi:10.1007/s11023-020-09517-8.

Hager, G.D., Drobnis A., Fang F., Ghani R. *et al.* 2019. Artificial Intelligence for Social Good. *arXiv:1901.05406 [cs]* [Preprint]. http://arxiv.org/abs/1901.05406.

Hand, D.J. 2018. Aspects of Data Ethics in a Changing World: Where Are We Now?. *Big Data*, Vol. 6, No. 3, pp. 176–190. doi:10.1089/big.2018.0083.

Hashiguchi, T. Cravo, O., Oderkirk J., and Slawomirski, L. 2022. Fulfilling the Promise of Artificial Intelligence in the Health Sector: Let's Get Real. *Value in Health*, No. 25, Vol. 3, pp. 368-73. https://doi.org/10.1016/j.jval.2021.11.1369.

Hébert, C. 2021. Un pour tous, tous pour Une seule santé. *Hinnovic*. https://www.hinnovic.org/post/un-pour-tous-tous-pour-une-seule-santé.

Hund, K. Laporta, D. Fabregas T.P. Laing, T. and Drexhage J. 2020. Minerals for Climate Action: The Mineral Intensity of the Clean Energy Transition. *The World Bank*. https://pubdocs.worldbank.org/en/961711588875536384/Minerals-for-Climate-Action-The-Mineral-Intensity-of-the-Clean-Energy-Transition.pdf.

IEA. 2021. The Role of Critical Minerals in Clean Energy Transitions – Analysis. *International Energy Agency*. https://www.iea.org/reports/the-role-of-critical-minerals-in-clean-energy-transitions/executive-summary.

IEEE. 2017. Ethically aligned design – Version 2 – For Public Discussion. *I. of E. and E.E.* https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

Ienca, M. and Vayena, E. 2020. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, Vol. 26, No. 4, pp. 463–464. doi:10.1038/s41591-020-0832-5.

Iyengar, A., Kundu, A. and Pallis, G. 2018. Healthcare Informatics and Privacy. *IEEE Internet Computing*, Vol. 22, No. 2, pp. 29–31. doi:10.1109/MIC.2018.022021660.

Jackman, M. 2020. Fault Lines: COVID-19, the Charter, and Long-term Care in *Vulnerable: The Law, Policy and Ethics of COVID-19 de Colleen M. Flood et al.* University of Ottawa Press, pp. 339–354. https://muse.jhu.edu/book/76885.

Jobin, A., Ienca, M. and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* Vol. 1, No. 9, pp. 389-99. https://doi.org/10.1038/s42256-019-0088-2.

Kassab, M. and Graciano Neto, V.V. 2021. Digital Surveillance Technologies to Combat COVID-19: A Contemporary View. *Procedia Computer Science*, Vol. 185, pp. 37–44. doi:10.1016/j.procs.2021.05.005.

Kenny, N. P., Sherwin, S.B. and Baylis, F.E. 2010. Re-visioning Public Health Ethics: A Relational Perspective. *Canadian Journal of Public Health*, Vol. 101, No. 1, pp. 9–11. doi:10.1007/BF03405552.

Kim, P.T. 2016. Data-Driven Discrimination at Work. *William & Mary Law Review*, Vol. 58, pp. 857–936.

Krantz, I., Sachs, L. and Nilstun, T. 2004. Ethics and vaccination. *Scandinavian Journal of Public Health*, Vol. 32, No. 3, pp. 172–178. doi:10.1080/14034940310018192.

Kudina, O. 2021. Bridging Privacy and Solidarity in COVID-19 Contact-tracing Apps through the Sociotechnical Systems Perspective. *Glimpse*, Vol. 22, No. 2, pp. 43–54. doi:10.5840/glimpse202122224.

Lagacé, M., Garcia, L. and Bélanger-Hardy, L. 2020. COVID-19 et âgisme: crise annoncée dans les centres de soins de longue durée et réponse improvisée? in *Vulnerable: The Law, Policy and Ethics of COVID-19 de Colleen M. Flood et al.* University of Ottawa Press, pp. 329–338. https://muse.jhu.edu/book/76885.

Lai, J. and Widmar, N.O. 2021. Revisiting the Digital Divide in the COVID-19 Era. *Applied Economic Perspectives and Policy*, Vol. 43, No. 1, pp. 458–464. doi:10.1002/aepp.13104.

Langlois, L. and Régis, C. 2021. Analyzing the Contribution of Ethical Charters to Building the Future of Artificial Intelligence Governance in Braunschweig, B. and Ghallab, M. (eds) *Reflections on Artificial Intelligence for Humanity*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 150–170. doi:10.1007/978-3-030-69128-8_10.

CEST. 2020. Les enjeux éthiques de l'utilisation d'une application mobile de traçage des contacts dans le cadre de la pandémie de COVID-19 au Québec. *Commission de l'éthique en science et en technologie*. https://www.ethique.gouv.qc.ca/fr/publications/l-utilisation-d-une-application-mobile-de-tracage-des-contacts-dans-le-cadre-d-une-pandemie/.

Ligozat, A.-L., Lefèvre, J. Bugeau, A. and Combaz, J. 2021. Unraveling the hidden environmental impacts of AI solutions for environment. *arXiv:2110.11822* [cs], (octobre). http://arxiv.org/abs/2110.11822.

Mackenzie, J.S. and Jeggo, M. 2019. The One Health Approach—Why Is It So Important? *Tropical Medicine and Infectious Disease*, Vol. 4, No. 2, p. 88. doi:10.3390/tropicalmed4020088.

Makri, A. 2019. Bridging the digital divide in health care. *The Lancet Digital Health*, Vol. 1, No. 5, pp. e204–e205. doi:10.1016/S2589-7500(19)30111-6.

Martins Van Jaarsveld, G. 2020. The Effects of COVID-19 Among the Elderly Population: A Case for Closing the Digital Divide. *Frontiers in Psychiatry*, Vol. 11, p. 577427. doi:10.3389/fpsyt.2020.577427.

Massé, R. 2003. Valeurs universelles et relativisme culturel en recherche internationale: les contributions d'un principisme sensible aux contextes socioculturels. *Autrepart*, Vol. 28, No. 4, pp. 21–35.

Mbunge, E. *et al.* 2021. Ethics for integrating emerging technologies to contain COVID-19 in Zimbabwe. *Human Behavior and Emerging Technologies*. doi:10.1002/hbe2.277.

McCarthy, M.T. 2016. The big data divide and its consequences. *Sociology Compass*, Vol. 10, No. 12, pp. 1131–1140. doi:10.1111/soc4.12436.

Mello, M.M. and Wang, C.J. 2020. Ethics and governance for digital disease surveillance. *Science (New York, N.Y.)*, Vol. 368, No. 6494, pp. 951–954. doi:10.1126/science.abb9045.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, Vol. 1, No. 11, pp. 501–507. doi:10.1038/s42256-019-0114-4.

Mittelstadt, B.D. and Floridi, L. 2016. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, Vol. 22, No. 2, pp. 303–341. doi:10.1007/s11948-015-9652-2.

Mondin, C. and de Marcellis-Warin, N. 2020. Recension des solutions technologiques développées dans le monde afin de limiter la propagation de la COVID-19 et typologie des applications de traçage. *Observatoire international sur les impacts sociétaux de l'IA et du numérique (OBVIA)*. https://www.docdroid.com/VLokunh/recension-des-solutions-technologiques-developpees-dans-le-monde-afin-de-limiter-la-propagation-de-la-covid-19-et-typologie-des-applications-de-tracage-pdf.

Morley, J., Floridi L., Kinsey L. and Elhalal A. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, Vol. 26, No. 4, pp. 2141-68. https://doi.org/10.1007/s11948-019-00165-5.

Murphy, K. *et al.* 2021. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Medical Ethics*, Vol. 22, No. 1, p. 14. doi:10.1186/s12910-021-00577-8.

United Nations. 2021. Cadre mondial d'indicateurs relatifs aux objectifs et aux cibles du Programme de développement durable à l'horizon 2030. https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202021%20refinement_Fre.pdf.

Naudé, W. 2020. Artificial intelligence vs COVID-19: limitations, constraints and pitfalls. *AI & SOCIETY*, Vol. 35, No. 3, pp. 761–765. doi:10.1007/s00146-020-00978-0.

News, L. 2020. *Covid-19 is magnifying the digital divide*, *LaptrinhX/News*. https://laptrinhx.com/news/covid-19-is-magnifying-the-digital-divide-pGZbpeA/.

Noorman, M. 2016. Computing and Moral Responsibility in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*. Winter 2016. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/.

O'Doherty, K.C., Christofides, E., Yen J., Beate Bentzen H. *et al.* 2016. If you build it, they will come: unintended future uses of organised health data collections. *Bmc Medical Ethics*, Vol. 17, p. 54. doi:10.1186/s12910-016-0137-x.

WHO. 2018. Big data and artificial intelligence for achieving universal health coverage: an international consultation on ethics. *World Health Organization*. http://apps.who.int/iris/bitstream/handle/10665/275417/WHO-HMM-IER-REK-2018.2-eng.pdf?ua=1.

––––. 2021a. Ethics and governance of artificial intelligence for health: WHO guidance. *World Health Organization*. https://www.who.int/publications/i/item/9789240029200.

––––. 2021b. WHO HUB FOR PANDEMIC AND EPIDEMIC INTELLIGENCE. Better Data. Better Analytics. Better Decisions. *World Health Organization, Health Emergencies Programme*. https://cdn.who.int/media/docs/default-source/2021-dha-docs/who_hub.pdf?sfvrsn=8dc28ab6_5.

Oosterlaken, I. 2015. *Technology and human development*. Routledge, Taylor & Francis. USA.

Patsavellas, J. and Salonitis, K. 2019. The Carbon Footprint of Manufacturing Digitalization: critical literature review and future research agenda. *Procedia CIRP*, Vol. 81, pp. 1354–1359. doi:10.1016/j.procir.2019.04.026.

Patz, J.A. *et al.* 2014. Climate Change: Challenges and Opportunities for Global Health. *JAMA*, Vol. 312, No. 15, pp. 1565–1580. doi:10.1001/jama.2014.13186.

de Pooter, H. 2015. *Le droit international face aux pandémies: vers un système de sécurité sanitaire collective?* Editions A. Pedone.

The Manhattan Principles. 2004. https://oneworldonehealth.wcs.org/About-Us/Mission/The-Manhattan-Principles.aspx.

Ramsetty, A. and Adams, C. 2020. Impact of the digital divide in the age of COVID-19. *Journal of the American Medical Informatics Association*, Vol. 27, No. 7, pp. 1147–1148. doi:10.1093/jamia/ocaa078.

Rial-Sebbag, E. 2017. Chapitre 4. La gouvernance des Big data utilisées en santé, un enjeu national et international. *Journal international de bioethique et d'ethique des sciences*, Vol. 28, No. 3, pp. 39–50.

Risse, M. 2019. Human Rights and Artificial Intelligence: An Urgently Needed Agenda. *Human Rights Quarterly*, Vol. 41, No. 1, pp. 1–16. doi:10.1353/hrq.2019.0000.

Scassa, T., Millar, J. and Bronson, K. 2020. Privacy, Ethics, and Contact-tracing Apps. *SSRN Scholarly Paper* ID 3651457. Rochester, NY: Social Science Research Network. doi:10.2139/ssrn.3651457.

Sen, A. 1983. Poor, Relatively Speaking. *Oxford Economic Papers*, Vol. 35, No. 2, pp. 153–169.

Sen, A. 2013. The Ends and Means of Sustainability. *Journal of Human Development and Capabilities*, Vol. 14, No. 1, pp. 6–20. doi:10.1080/19452829.2012.747492.

Shachar, C., Gerke, S. and Adashi, E.Y. 2020. AI Surveillance during Pandemics: Ethical Implementation Imperatives. *The Hastings Center Report*, Vol. 50, No. 3, pp. 18–21. doi:10.1002/hast.1125.

Shneiderman, B. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, Vol. 10, No. 4, p. 26:1-26:31. doi:10.1145/3419764.

Siau, K. and Wang, W. 2020. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management (JDM)*, Vol. 31, No. 2, pp. 74–87. doi:10.4018/JDM.2020040105.

Sim, F.M. 2017. Individualism and social solidarity in vaccination policy: some further considerations. *Israel Journal of Health Policy Research*, Vol. 6, No. 1, p. 21. doi:10.1186/s13584-017-0147-2.

Solomon, C.G. and LaRocque, R.C. 2019. Climate Change — A Health Emergency. *New England Journal of Medicine*, Vol. 380, No. 3, pp. 209–211. doi:10.1056/NEJMp1817067.

Solomon, D.H. *et al.* 2020. The "Infodemic" of COVID-19. *Arthritis & Rheumatology*, Vol. 72, No. 11, pp. 1806–1808. doi:10.1002/art.41468.

Spiekermann, S., Korunovska, J. and Langheinrich, M. 2018. Inside the Organization: Why Privacy and Security Engineering Is a Challenge for Engineers. *Proceedings of the IEEE*, pp. 1–16. doi:10.1109/JPROC.2018.2866769.

Stahl, B.C. and Wright, D. 2018. Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security Privacy*, Vol. 16, No. 3, pp. 26–33. doi:10.1109/MSP.2018.2701164.

Stapleton G., Schröder-Bäck P., Laaser U., Meershoek A. and Popa D. 2014. Global health ethics: an introduction to prominent theories and relevant topics. *Global Health Action*, Vol. 7(s2), p. 23569. doi:10.3402/gha.v7.23569.

von Struensee, S. 2021. Mapping Artificial Intelligence Applications Deployed Against COVID-19 Alongside Ethics and Human Rights Considerations. *SSRN Scholarly Paper ID 3889441*. Rochester, NY: Social Science Research Network. doi:10.2139/ssrn.3889441.

Tagliabue, F., Galassi, L. and Mariani, P. 2020. The "Pandemic" of Disinformation in COVID-19. *Sn Comprehensive Clinical Medicine*, pp. 1–3. doi:10.1007/s42399-020-00439-1.

Terry, N. and Coughlin, C.N. 2021. A Virtuous Circle: How Health Solidarity Could Prompt Recalibration of Privacy and Improve Data and Research. *SSRN Scholarly Paper ID 3774366. Rochester*, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=3774366.

The Future Society. 2020. Areas for future action in the responsible AI ecosystem. https://thefuturesociety.org/wp-content/uploads/2021/02/Areas-for-Future-Action-in-the-Responsible-AI-Ecosystem.pdf.

The Shift Project. 2020. Déployer la sobriété numérique. https://theshiftproject.org/article/deployer-la-sobriete-numerique-rapport-shift.

The Shift Project. 2021. Impact environnemental du numérique et gouvernance de la 5G. https://theshiftproject.org/article/impact-environnemental-du-numerique-5g-nouvelle-etude-du-shift/.

Tran, C.D. and Nguyen, T.T. 2021. Health vs. privacy? The risk-risk tradeoff in using COVID-19 contact-tracing apps. *Technology in Society*, Vol. 67, p. 101755. doi:10.1016/j.techsoc.2021.101755.

Tzachor, A. *et al.* 2020. Artificial intelligence in a crisis needs ethics with urgency. *Nature Machine Intelligence*, Vol. 2, No. 7, pp. 365–366. doi:10.1038/s42256-020-0195-0.

UNDP. 2020. *Human Development Report 2020. The next frontier Human development and the Anthropocene*. https://reliefweb.int/report/world/human-development-report-2020-next-frontier-human-development-and-anthropocene.

United Nations. n.d. Universal Values, Principle 2: Leave No One Behind. *United Nations Sustainable Development Group.* https://unsdg.un.org/2030-agenda/universal-values/leave-no-one-behind

Vaishya, R. *et al.* 2020. Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, Vol. 14, No. 4, pp. 337–339. doi:10.1016/j.dsx.2020.04.012.

Villani, C. 2018. Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

Voarino, N. 2020. Systèmes d'intelligence artificielle et santé: les enjeux d'une innovation responsable. Thèse. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/23526.

Whittlestone, J. *et al.* 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery (AIES '19), pp. 195–200. doi:10.1145/3306618.3314289.

Wilson, S. L., et Wiysonge, C. 2020. Social Media and Vaccine Hesitancy. *BMJ Global Health*, Vol. 5, No. 10: e004206. https://doi.org/10.1136/bmjgh-2020-004206.

Wynants, L. *et al.* 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*, Vol. 369, p. m1328. doi:10.1136/bmj.m1328.

K. Yeung; A. Howes and G. Pogrebna. AI governance by Human Rights-Centered Design, Deliberation, and Oversight in *Dubber Pasquale and Das (eds)*, The Oxford Handbook of Ethics of AI, Oxford University Press, 2020, p. 77-106.

Yu, H. *et al.* 2018. Building Ethics into Artificial Intelligence. *arXiv:1812.02953 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1812.02953.

DATA: FROM THE ATLAS OF AI[31]

KATE CRAWFORD

Principal researcher at Microsoft Research, the co-founder and director of research at the AI Now Institute at NYU. Author of “Atlas of AI” (2021). New Haven, Yale University Press.

SDG3 - Good Health and Well-being
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG12 - Responsible Consumption and Production
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

31. This chapter is a reproduction of “Atlas of AI,” chapter 3. Copyrights granted to Mila – Québec Artificial Intelligence Institute by Yale University Press on 23 November 2021 for reproduction in this publication.

DATA: FROM THE ATLAS OF AI

ABSTRACT

This chapter explores how data is the essential foundation of AI systems. It exposes the underlying politics, ethical issues, epistemological limitations, and range of harms that arise from the logics of data extraction and accumulation in the AI industry. In the race to algorithmic performance, more data is better. Hence, everything is presumed to be there for the taking. Extremely large datasets are seen as neutral infrastructures: interpreted as "things" devoid of context and meaning, despite the deeply personal and sometimes horrifying images they contain. By excavating the data layers, we discover the stories: individual and collective accounts of historical injustice, discrimination, and structural inequities. The widely accepted understanding of data as a resource to be consumed, a flow to be controlled, and an investment to be harnessed has produced a kind of hubris – a statistical ideology where only scale matters.

INTRODUCTION

A young woman gazes upward, eyes focused on something outside the frame, as though she is refusing to acknowledge the camera. In the next photograph, her eyes are locked on the middle distance. Another image shows her with disheveled hair and a downcast expression. Over the sequence of photos we see her aging over time, and the lines around her mouth turn down and deepen. In the final frame she appears injured and dispirited. These are mug shots of a woman across multiple arrests over many years of her life. Her images are contained in a collection known as NIST Special Database 32–Multiple Encounter Dataset, which is shared on the internet for researchers who would like to test their facial recognition software (NIST, 2010).

This dataset is one of several maintained by the National Institute of Standards and Technology (NIST), one of the oldest and most respected physical science laboratories in the United States and now part of the Department of Commerce. NIST was created in 1901 to bolster the nation's measurement infrastructure and to create standards that could compete with economic rivals in the industrialized world, such as Germany and the United Kingdom. Everything from electronic health records to earthquake-resistant skyscrapers to atomic clocks is under the purview of NIST. It became the agency of measurement: of time, of communications protocols, of inorganic crystal structures, of nanotechnology (Russell, 2014). NIST's purpose is to make systems interoperable through defining and supporting standards, and this now includes developing standards for artificial intelligence. One of the testing infrastructures it maintains is for biometric data.

| FIGURE 1 |

Images from NIST Special Database 32—Multiple Encounter Dataset (MEDS). National Institute of Standards and Technology, U.S. Department of Commerce.

I first discovered the mug shot databases in 2017 when I was researching NIST's data archives. Their biometric collections are extensive. For more than fifty years, NIST has collaborated with the Federal Bureau of Investigation on automated fingerprint recognition and has developed methods to assess the quality of fingerprint scanners and imaging systems (Garris and Wilson, 2005, p. 1). After the terrorist attacks of September 11, 2001, NIST became part of the national response to create biometric standards to verify and track people entering the United States (Garris and Wilson, 2005, p. 1). This was a turning point for research on facial recognition; it widened out from a focus on law enforcement to controlling people crossing national borders (Garris and Wilson, 2005, p. 12).

The mug shot images themselves are devastating. Some people have visible wounds, bruising, and black eyes; some are distressed and crying. Others stare blankly back at the camera. Special Dataset 32 contains thousands of photographs of deceased people with multiple arrests, as they endured repeated encounters with the criminal justice system. The people in the mug shot datasets are presented as data points; there are no stories, contexts, or names. Because mug shots are taken at the time of arrest, it's not clear if these people were charged, acquitted, or imprisoned. They are all presented alike.

The inclusion of these images in the NIST database has shifted their meaning from being used to identify individuals in systems of law enforcement to becoming the technical baseline to test commercial and academic AI systems for detecting faces. In his account of police photography, Allan Sekula has argued that mug shots are part of a tradition of technical realism that aimed to "provide a standard physiognomic gauge of the criminal" (Sekula, 1986, p. 17). There are two distinct approaches in the history of the police photograph, Sekula observes. Criminologists like Alphonse Bertillon, who invented the mug shot, saw it as a kind of biographical machine of identification, necessary to spot repeat offenders. On the other hand, Francis Galton, the statistician and founding figure of eugenics, used composite portraiture of prisoners as a way to detect a biologically determined "criminal type" (Sekula, pp. 18-19). Galton was working within a physiognomist paradigm in which the goal was to find a generalized look that could be used to identify deep character traits from external appearances. When mug shots are used as training data, they function no longer as tools of identification but rather to fine-tune an automated form of vision. We might think of this as Galtonian formalism. They are used to detect the basic mathematical components of faces, to "reduce nature to its geometrical essence" (Sekula, 1986, p. 17).

Mug shots form part of the archive that is used to test facial-recognition algorithms. The faces in the Multiple Encounter Dataset have become standardized images, a technical substrate for comparing algorithmic accuracy. NIST, in collaboration with the Intelligence Advanced Research Projects Activity (IARPA), has run competitions with these mug shots in which researchers compete to see whose algorithm is the fastest and most accurate. Teams strive to beat one another at tasks like verifying the identity of faces or retrieving a face from a frame of surveillance video (Grother et al., 2017). The winners celebrate these victories; they can bring fame, job offers, and industrywide recognition (Ever AI, 2018).

Neither the people depicted in the photographs nor their families have any say about how these images are used and likely have no idea that they are part of the test beds of AI. The subjects of the mug shots are rarely considered, and few engineers will ever look at them closely. As the NIST document describes them, they exist purely to "refine tools, techniques, and procedures for face recognition as it supports Next Generation Identification (NGI), forensic comparison, training, analysis, and face image conformance and inter-agency exchange standards" (Founds et al., 2011). The Multiple Encounter Dataset description observes that many people show signs of enduring violence, such as scars, bruises, and bandages. But the document concludes that these signs are "difficult to interpret due to the lack of ground truth for comparison with a 'clean' sample" (Curry et al., 2009). These people are not seen so much as individuals but as part of a shared technical resource – just another data component of the Facial Recognition Verification Testing program, the gold standard for the field.

I've looked at hundreds of datasets over years of research into how AI systems are built, but the NIST mug shot databases are particularly disturbing because they represent the model of what was to come. It's not just the overwhelming pathos of the images themselves. Nor is it solely the invasion of privacy they represent, since suspects and prisoners have no right to refuse being photographed. It's that the NIST databases foreshadow the emergence of a logic that has now thoroughly pervaded the tech sector: the unswerving belief that everything is data and is there for the taking. It doesn't matter where a photograph was taken or whether it reflects a moment of vulnerability or pain or if it represents a form of shaming the subject. It has become so normalized across the industry to take and use whatever is available that few stop to question the underlying politics.

Mug shots, in this sense, are the urtext of the current approach to making AI. The context – and exertion of power – that these images represent is considered irrelevant because they no longer exist as distinct things unto themselves. They are not seen to carry meanings or ethical weight as images of individual people or as representations of structural power in the carceral system. The personal, the social, and the political meanings are all imagined to be neutralized. I argue this represents a shift from *image* to *infrastructure,* where the meaning or care that might be given to the image of an individual person, or the context behind a scene, is presumed to be erased at the moment it becomes part of an aggregate mass that will drive a broader system. It is all treated as data to be run through functions, material to be ingested to improve technical performance. This is a core premise in the ideology of data extraction.

Machine learning systems are trained on images like these every day—images that were taken from the internet or from state institutions without context and without consent. They are anything but neutral. They represent personal histories, structural inequities, and all the injustices that have accompanied the legacies of policing and prison systems in the United States. But the presumption that somehow these images can serve as apolitical, inert material influences how and what a machine learning tool "sees." A computer vision system can detect a face or a building but not why a person was inside a police station or any of the social and historical context surrounding that moment. Ultimately, the specific instances of data – a picture of a face, for example – aren't considered to matter for training an AI model. All that matters is a sufficiently varied aggregate. Any individual image could easily be substituted for another and the system would work the same. According to this worldview, there is always more data to capture from the constantly growing and globally distributed treasure chest of the internet and social media platforms.

A person standing in front of a camera in an orange jumpsuit, then, is dehumanized as just more data. The history of these images, how they were acquired, and their institutional, personal, and political contexts are not considered relevant. The mug shot collections are used like any other practical resource of free, well-lit images of faces, a benchmark to make tools like facial recognition function. And like a tightening ratchet, the faces of deceased persons, suspects, and prisoners are harvested to sharpen the police and border surveillance facial recognition systems that are then used to monitor and detain more people.

The last decade has seen a dramatic capture of digital material for AI production. This data is the basis for sensemaking in AI, not as classical representations of the world with individual meaning, but as a mass collection of data for machine abstractions and operations. This large-scale capture has become so fundamental to the AI field that it is unquestioned. So how did we get here? What ways of conceiving data have facilitated this stripping of context, meaning, and specificity? How is training data acquired, understood, and used in machine learning? In what ways does training data limit *what* and *how* AI interprets the world? What forms of power do these approaches enhance and enable?

In this chapter I show how data has become a driving force in the success of AI and its mythos and how everything that can be readily captured is being acquired. But the deeper implications of this standard approach are rarely addressed, even as it propels further asymmetries of power. The AI industry has

fostered a kind of ruthless pragmatism, with minimal context, caution, or consent-driven data practices while promoting the idea that the mass harvesting of data is necessary and justified for creating systems of profitable computational "intelligence." This has resulted in a profound metamorphosis, where all forms of image, text, sound, and video are just raw data for AI systems and the ends are thought to justify the means. But we should ask: Who has benefited most from this transformation, and why have these dominant narratives of data persisted? The logic of extraction that has shaped the relationship to the earth and to human labor is also a defining feature of how data is used and understood in AI. By looking closely at training data as a central example in the ensemble of machine learning, we can begin to see what is at stake in this transformation.

TRAINING MACHINES TO SEE

It's useful to consider why machine learning systems currently demand massive amounts of data. One example of the problem in action is computer vision, the subfield of AI concerned with teaching machines to detect and interpret images. For reasons that are rarely acknowledged in the field of computer science, the project of interpreting images is a profoundly complex and relational endeavor. Images are remarkably slippery things, laden with multiple potential meanings, irresolvable questions, and contradictions. Yet now it's common practice for the first steps of creating a computer vision system to scrape thousands – or even millions – of images from the internet, create and order them into a series of classifications, and use this as a foundation for how the system will perceive observable reality. These vast collections are called training datasets, and they constitute what AI developers often refer to as "ground truth" (Jaton, 2017). Truth, then, is less about a factual representation or an agreed-upon reality and more commonly about a jumble of images scraped from whatever various online sources were available.

For supervised machine learning, human engineers supply labeled training data to a computer. Two distinct types of algorithms then come into play: *learners* and *classifiers.* The learner is the algorithm that is trained on these labeled data examples; it then informs the classifier how best to analyze the relation between the new inputs and the desired target output (or prediction). It might be predicting whether a face is contained in an image or whether an email is spam. The more examples of correctly labeled data there are, the better the algorithm will be at producing accurate predictions. There are many kinds of machine learning models, including neural networks, logistic regression, and decision trees. Engineers will choose a model based on what they are building – be it a facial recognition system or a means of detecting sentiment on social media – and fit it to their computational resources.

Consider the task of building a machine learning system that can detect the difference between pictures of apples and oranges. First, a developer has to collect, label, and train a neural network on thousands of labeled images of apples and oranges. On the software side, the algorithms conduct a statistical survey of the images and develop a model to recognize the difference between the two classes. If all goes according to plan, the trained model will be able to distinguish the difference between images of apples and oranges that it has never encountered before.

But if, in our example, all of the training images of apples are red and none are green, then a machine learning system might deduce that "all apples are red." This is what is known as an *inductive inference,* an open hypothesis based on available data, rather than a *deductive inference,* which follows logically from a premise (Nilsson, 2009, p. 398). Given how this system was trained, a green apple wouldn't be recognized as an apple at all. Training datasets, then, are at the core of how most machine learning systems make inferences. They serve as the primary source material that AI systems use to form the basis of their predictions.

Training data also defines more than just the features of machine learning algorithms. It is used to assess how they perform over time. Like prized thoroughbreds, machine learning algorithms are constantly raced against one another in competitions all over the world to see which ones perform the best with a given dataset. These benchmark datasets become the alphabet on which a *lingua franca* is based, with many labs from multiple countries converging around canonical sets to try to outperform one another. One of the best-known competitions is the ImageNet Challenge, where researchers compete to see whose methods can most accurately classify and detect objects and scenes[32].

Once training sets have been established as useful benchmarks, they are commonly adapted, built upon, and expanded. A type of genealogy of training sets emerges – they inherit learned logic from earlier examples and then give rise to subsequent ones (Crawford, 2021, ch. 4). For example, ImageNet draws on the taxonomy of words inherited from the influential 1980s lexical database known as WordNet; and WordNet inherits from many sources, including the Brown Corpus of one million words, published in 1961. Training datasets stand on the shoulders of older classifications and collections. Like an expanding encyclopedia, the older forms remain and new items are added over decades.

Training data, then, is the foundation on which contemporary machine learning systems are built[33] (Michalski, 1980). These datasets shape the epistemic boundaries governing how AI operates and, in that sense, create the limits of how AI can "see" the world. But training data is a brittle form of ground truth – and even the largest troves of data cannot escape the fundamental slippages that occur when an infinitely complex world is simplified and sliced into categories.

A BRIEF HISTORY OF THE DEMAND FOR DATA

"The world has arrived at an age of cheap complex devices of great reliability; and something is bound to come of it." So said Vannevar Bush, the inventor and administrator who oversaw the Manhattan Project as director of the Office of Scientific Research and Development and later was integral to the creation of the National Science Foundation. It was July 1945; the bombs were yet to drop on Hiroshima and Nagasaki, and Bush had a theory about a new kind of data-connecting system that was yet to be born. He envisaged the "advanced arithmetical machines of the future" that would perform at extremely fast speed and "select their own data and manipulate it in accordance with the instructions." But the machines would need monumental amounts of data: "Such machines will have enormous appetites. One of them will take instructions and data from a whole roomful of girls armed with simple key board punches, and will deliver sheets of computed results every few minutes. There will always be plenty of things to compute in the detailed affairs of millions of people doing complicated things" (Bush, 1945).

The "roomful of girls" Bush referred to were the keypunch operators doing the day-to-day work of computation. As historians Jennifer Light and Mar Hicks have shown, these women were often dismissed as input devices for intelligible data records. In fact, their role was just as important to crafting data and making systems work as that of the engineers who designed the wartime-era

32. For more information, see: "ImageNet Large Scale Visual Recognition Competition (ILSVRC)." http://image-net.org/challenges/LSVRC/.

33. In the late 1970s, Ryszard Michalski wrote an algorithm based on symbolic variables and logical rules. This language was popular in the 1980s and 1990s, but as the rules of decision-making and qualification became more complex, the language became less usable. At the same moment, the potential of using large training sets triggered a shift from this conceptual clustering to contemporary machine learning approaches

digital computers (Light, 1999). But the relationship between data and processing machinery was already being imagined as one of endless consumption. The machines would be data-hungry, and there would surely be a wide horizon of material to extract from millions of people.

In the 1970s, AI researchers were mainly exploring what's called an expert systems approach: rules-based programming that aims to reduce the field of possible actions by articulating forms of logical reasoning. But it quickly became evident that this approach was fragile and impractical in real-world settings, where a rule set was rarely able to handle uncertainty and complexity (Russell and Norvig, 2010, p. 546). New approaches were needed. By the mid-1980s, research labs were turning toward probabilistic or brute force approaches. In short, they were using lots of computing cycles to calculate as many options as possible to find the optimal result.

One significant example was the speech recognition group at IBM Research. The problem of speech recognition had primarily been dealt with using linguistic methods, but then information theorists Fred Jelinek and Lalit Bahl formed a new group, which included Peter Brown and Robert Mercer (long before Mercer became a billionaire, associated with funding Cambridge Analytica, Breitbart News, and Donald Trump's 2016 presidential campaign). They tried something different. Their techniques ultimately produced precursors for the speech recognition systems underlying Siri and Dragon Dictate, as well as machine translation systems like Google Translate and Microsoft Translator.

They started using statistical methods that focused more on how often words appeared in relation to one another, rather than trying to teach computers a rules-based approach using grammatical principles or linguistic features. Making this statistical approach work required an enormous amount of real speech and text data, or training data. The result, as media scholar Xiaochang Li writes, was that it required "a radical reduction of speech to merely data, which could be modeled and interpreted in the absence of linguistic knowledge or understanding. Speech *as such* ceased to matter." This shift was incredibly significant, and it would become a pattern repeated for decades: the reduction from context to data, from meaning to statistical pattern recognition. Li explains:

> The reliance on data over linguistic principles, however, presented a new set of challenges, for it meant that the statistical models were necessarily determined by the characteristics of training data. As a result, the size of the dataset became a central concern. Larger datasets of observed outcomes not only improved the probability estimates for a random process, but also increased the chance that the data would capture more rarely-occurring outcomes. Training data size, in fact, was so central to IBM's approach that in 1985, Robert Mercer explained the group's outlook by simply proclaiming, "There's no data like more data" (Li, 2017, p. 143).

For several decades, that data was remarkably hard to come by. As Lalit Bahl describes in an interview with Li, "Back in those days… you couldn't even find a million words in computer-readable text very easily. And we looked all over the place for text" (Li, 2017, p. 144). They tried IBM technical manuals, children's novels, patents of laser technology, books for the blind, and even the typed correspondence of IBM Fellow Dick Garwin, who created the first hydrogen bomb design (Brown and Mercer, 2013). Their method strangely echoed a short story by the science fiction author Stanislaw Lem, in which a man called Trurl decides to build a machine that would write poetry. He starts with "eight hundred and twenty tons of books on cybernetics and twelve thousand tons of the finest poetry" (Lem, 2003, p. 199). But Trurl realizes that to program an autonomous poetry machine, one needs "to repeat the entire Universe from the beginning—or at least a good piece of it" (Lem, 2003, p. 199).

Ultimately, the IBM Continuous Speech Recognition group found their "good piece" of the universe from an unlikely source. A major federal antitrust lawsuit was filed against IBM in 1969; the proceedings lasted for thirteen years, and almost a thousand witnesses were called. IBM employed a large staff just

to digitize all of the deposition transcripts onto Hollerith punch cards. This ended up creating a corpus of a hundred million words by the mid-1980s. The notoriously antigovernment Mercer called this a "case of utility accidentally created by the government in spite of itself" (Brown and Mercer, 2013).

IBM wasn't the only group starting to gather words by the ton. From 1989 to 1992, a team of linguists and computer scientists at the University of Pennsylvania worked on the Penn Treebank Project, an annotated database of text. They collected four and a half million words of American English for the purpose of training natural language processing systems. Their sources included Department of Energy abstracts, Dow Jones newswire articles, and Federal News Service reports of "terrorist activity" in South America (Marcus et al., 1993). The emerging text collections borrowed from earlier collections and then contributed new sources. Genealogies of data collections began to emerge, each building on the last – and often importing the same peculiarities, issues, or omissions wholesale.

Another classic corpus of text came from the fraud investigations of Enron Corporation after it declared the largest bankruptcy in American history. The Federal Energy Regulatory Commission seized the emails of 158 employees for the purposes of legal discovery (Klimt and Yang, 2004). It also decided to release these emails online because "the public's right to disclosure outweighs the individual's right to privacy" (Wood III et al., 2003, p. 12). This became an extraordinary collection. Over half a million exchanges in everyday speech could now be used as a linguistic mine: one that nonetheless represented the gender, race, and professional skews of those 158 workers. The Enron corpus has been cited in thousands of academic papers. Despite its popularity, it is rarely looked at closely: the *New Yorker* described it as "a canonic research text that no one has actually read" (Heller, 2017). This construction of and reliance on training data anticipated a new way of doing things. It transformed the field of natural language processing and laid the foundations of what would become normal practice in machine learning.

The seeds of later problems were planted here. Text archives were seen as neutral collections of language, as though there was a general equivalence between the words in a technical manual and how people write to colleagues via email. All text was repurposable and swappable, so long as there was enough of it that it could train a language model to predict with high levels of success what word might follow another. Like images, text corpuses work on the assumption that all training data is interchangeable. But language isn't an inert substance that works the same way regardless of where it is found. Sentences taken from Reddit will be different from those composed by executives at Enron. Skews, gaps, and biases in the collected text are built into the bigger system, and if a language model is based on the kinds of words that are clustered together, it matters where those words come from. There is no neutral ground for language, and all text collections are also accounts of time, place, culture, and politics. Further, languages that have less available data are not served by these approaches and so are often left behind (Baker et al., 2009).

Clearly there are many histories and contexts that combine within IBM's training data, the Enron archive, or the Penn Treebank. How do we unpack what is and is not meaningful to understand these datasets? How does one communicate warnings like, "This dataset likely reflects skews related to its reliance on news stories about South American terrorists in the 1980s"? The origins of the underlying data in a system can be incredibly significant, and yet there are still, thirty years later, no standardized practices to note where all this data came from or how it was acquired—let alone what biases or classificatory politics these datasets contain that will influence all the systems that come to rely on them (Gebru et al., 2021; Mitchell et al., 2019; Raji and Buolamwini, 2019).

CAPTURING THE FACE

While computer-readable text was becoming highly valued for speech recognition, the human face was the core concern for building systems of facial recognition. One central example emerged in the last decade of the twentieth century, funded by the Department of Defense CounterDrug Technology Development Program Office. It sponsored the Face Recognition Technology (FERET) program to develop automatic face recognition for intelligence and law enforcement. Before FERET, little training data of human faces was available, only a few collections of fifty or so faces here and there—not enough to do facial recognition at scale. The U.S. Army Research Laboratory led the technical project of creating a training set of portraits of more than a thousand people, in multiple poses, to make a grand total of 14,126 images. Like NIST's mug shot collections, FERET became a standard benchmark – a shared measuring tool to compare approaches for detecting faces.

The tasks that the FERET infrastructure was created to support included, once again, automated searching of mug shots, as well as monitoring airports and border crossings and searching driver's license databases for "fraud detection" (multiple welfare claims was a particular example mentioned in FERET research papers) (Phillips et al., 1996, p. 9). But there were two primary testing scenarios. In the first, an electronic mug book of known individuals would be presented to an algorithm, which then had to locate the closest matches from a large gallery. The second scenario focused on border and airport control: identifying a known individual – "smugglers, terrorists, or other criminals" – from a large population of unknown people.

These photographs are machine-readable by design, and not meant for human eyes, yet they make for remarkable viewing. The images are surprisingly beautiful – high-resolution photographs captured in the style of formal portraiture. Taken with 35 mm cameras at George Mason University, the tightly framed headshots depict a wide range of people, some of whom seem to have dressed for the occasion with carefully styled hair, jewelry, and makeup. The first set of photographs, taken between 1993 and 1994, are like a time capsule of early nineties haircuts and fashion. The subjects were asked to turn their heads to multiple positions; flicking through the images, you can see profile shots, frontal images, varying levels of illumination, and sometimes different outfits. Some subjects were photographed over several years, in order to begin to study how to track people as they age. Each subject was briefed about the project and signed a release form that had been approved by the university's ethics review board. Subjects knew what they were participating in and gave full consent (Phillips et al., 1996, p. 61). This level of consent would become a rarity in later years.

FERET was the high-water mark of a formal style of "making data," before the internet began offering mass extraction without any permissions or careful camera work. Even at this early stage, though, there were problems with the lack of diversity of the faces collected. The FERET research paper from 1996 admits that "some questions were raised about the age, racial, and sexual distribution of the database" but that "at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals" (Phillips et al., 1996, p. 12). Indeed, FERET was extraordinarily useful for this. As the interest in terrorist detection intensified and funding for facial recognition dramatically increased after 9/11, FERET became the most commonly used benchmark. From that point onward, biometric tracking and automated vision systems would rapidly expand in scale and ambition.

FROM THE INTERNET TO IMAGENET

The internet, in so many ways, changed everything; it came to be seen in the AI research field as something akin to a natural resource, there for the taking. As more people began to upload their images to websites, to photo-sharing services, and ultimately to social media platforms, the pillaging began

in earnest. Suddenly, training sets could reach a size that scientists in the 1980s could never have imagined. Gone was the need to stage photo shoots using multiple lighting conditions, controlled parameters, and devices to position the face. Now there were millions of selfies in every possible lighting condition, position, and depth of field. People began to share their baby photos, family snaps, and images of how they looked a decade ago, an ideal resource for tracking genetic similarity and face aging. Trillions of lines of text, containing both formal and informal forms of speech, were published every day. It was all grist for the mills of machine learning. And it was vast. As an example, on an average day in 2019, approximately 350 million photographs were uploaded to Facebook and 500 million tweets were sent (Aslam, 2020). And that's just two platforms based in the United States. Anything and everything online was primed to become a training set for AI.

The tech industry titans were now in a powerful position: they had a pipeline of endlessly refreshing images and text, and the more people shared their content, the more the tech industry's power grew. People would happily label their photographs with names and locations, free of charge, and that unpaid labor resulted in having more accurate, labeled data for machine vision and language models. Within the industry, these collections are highly valuable. They are proprietary troves that are rarely shared, given both the privacy issues and the competitive advantage they represent. But those outside the industry, such as the leading computer science labs in academia, wanted the same advantages. How could they afford to harvest people's data and have it hand-labeled by willing human participants? That's when new ideas began to emerge: combining images and text extracted from the internet with the labor of low-paid crowdworkers.

One of the most significant training sets in AI is ImageNet. It was first conceptualized in 2006, when Professor Fei-Fei Li decided to build an enormous dataset for object recognition. "We decided we wanted to do something that was completely historically unprecedented," Li said. "We're going to map out the entire world of objects" (Gershgorn, 2017). The breakthrough research poster was published by the ImageNet team at a computer vision conference in 2009. It opened with this description:

> The digital era has brought with it an enormous explosion of data. The latest estimations put a number of more than 3 billion photos on Flickr, a similar number of video clips on YouTube and an even larger number for images in the Google Image Search database. More sophisticated and robust models and algorithms can be proposed by exploiting these images, resulting in better applications for users to index, retrieve, organize and interact with these data (Deng et al., 2009).

From the outset, data was characterized as something voluminous, disorganized, impersonal, and ready to be exploited. According to the authors, "Exactly how such data can be utilized and organized is a problem yet to be solved." By extracting millions of images from the internet, primarily from search engines using the image-search option, the team produced a "large-scale ontology of images" that was meant to serve as a resource for "providing critical training and benchmarking data" for object and image recognition algorithms. Using this approach, ImageNet grew enormous. The team mass-harvested more than fourteen million images from the internet to be organized into more than twenty thousand categories. Ethical concerns about taking people's data were not mentioned in any of the team's research papers, even though many thousands of the images were of a highly personal and compromising nature.

Once the images had been scraped from the internet, a major concern arose: Who would label them all and put them into intelligible categories? As Li describes it, the team's first plan was to hire undergraduate students for ten dollars an hour to find images manually and add them to the dataset (Gershgorn, 2017). But she realized that with their budget, it would take more than ninety years to complete the project. The answer came when a student told Li about a new service: Amazon Mechanical Turk. As presented in chapter 2 of Crawford (2021), this distributed platform meant that it was suddenly possible to access a distributed labor force to do online tasks, like labeling and sorting images, at scale and at low cost. "He showed me the website, and I can tell you literally that day I knew

the ImageNet project was going to happen," Li said. "Suddenly we found a tool that could scale, that we could not possibly dream of by hiring Princeton undergrads" (Gershgorn, 2017). Unsurprisingly, the undergraduates did not get the job.

Instead, ImageNet would become, for a time, the world's largest academic user of Amazon's Mechanical Turk, deploying an army of piecemeal workers to sort an average of fifty images a minute into thousands of categories (Markoff, 2016). There were categories for apples and airplanes, scuba divers and sumo wrestlers. But there were cruel, offensive, and racist labels, too: photographs of people were classified into categories like "alcoholic," "ape-man," "crazy," "hooker," and "slant eye." All of these terms were imported from WordNet's lexical database and given to crowdworkers to pair with images. Over the course of a decade, ImageNet grew into a colossus of object recognition for machine learning and a powerfully important benchmark for the field. The approach of mass data extraction without consent and labeling by underpaid crowdworkers would become standard practice, and hundreds of new training datasets would follow ImageNet's lead. These practices – and the labeled data they generated – eventually came back to haunt the project.

THE END OF CONSENT

The early years of the twenty-first century marked a shift away from consent-driven data collection. In addition to dispensing with the need for staged photo shoots, those responsible for assembling datasets presumed that the contents of the internet were theirs for the taking, beyond the need for agreements, signed releases, and ethics reviews. Now even more troubling practices of extraction began to emerge. For example, at the Colorado Springs campus of the University of Colorado, a professor installed a camera on the main walkway of the campus and secretly captured photos of more than seventeen hundred students and faculty – all to train a facial recognition system of his own (Hernandez, 2019). A similar project at Duke University harvested footage of more than two thousand students without their knowledge as they went between their classes and then published the results on the internet. The dataset, called DukeMTMC (for multitarget, multicamera facial recognition), was funded by the U.S. Army Research Office and the National Science Foundation (Zhang et al., 2017).

The DukeMTMC project was roundly criticized after an investigative project by artists and researchers Adam Harvey and Jules LaPlace showed that the Chinese government was using the images to train systems for the surveillance of ethnic minorities. This spurred an investigation by Duke's institutional review board, which determined that this was a "significant deviation" from acceptable practices. The dataset was removed from the internet (Satisky, 2019).

But what happened at the University of Colorado and Duke were by no means isolated cases. At Stanford University, researchers commandeered a webcam from a popular café in San Francisco to extract almost twelve thousand images of "everyday life of a busy downtown café" without anyone's consent (Harvey and LaPlace, 2015). Over and over, data extracted without permission or consent would be uploaded for machine learning researchers, who would then use it as an infrastructure for automated imaging systems.

Another example is Microsoft's landmark training dataset MS-Celeb, which scraped approximately ten million photos of a hundred thousand celebrities from the internet in 2016. At the time, it was the largest public facial recognition dataset in the world, and the people included were not just famous actors and politicians but also journalists, activists, policymakers, academics, and artists (Locker, 2019). Ironically, several of the people who had been included in the set without consent are known for their

work critiquing surveillance and facial recognition itself, including documentary filmmaker Laura Poitras; digital rights activist Jillian York; critic Evgeny Morozov; and the author of *Surveillance Capitalism,* Shoshana Zuboff (Murgia and Harlow, 2019; Locker, 2019).[34]

Even when datasets are scrubbed of personal information and released with great caution, people have been reidentified or highly sensitive details about them have been revealed. In 2013, for example, the New York City Taxi and Limousine Commission released a dataset of 173 million individual cab rides, and it included pickup and drop-off times, locations, fares, and tip amounts. The taxi drivers' medallion numbers were anonymized, but this was quickly undone, enabling researchers to infer sensitive information like annual incomes and home addresses (Franceschi-Bicchierai, 2015). Once combined with public information from sources like celebrity blogs, some actors and politicians were identified, and it was possible to deduce the addresses of people who visited strip clubs (Tockar, 2014). But beyond individual harms, such datasets also generate "predictive privacy harms" for whole groups or communities (Crawford and Schultz, 2019). For instance, the same New York City taxi dataset was used to suggest which taxi drivers were devout Muslims by observing when they stopped at prayer times (Franceschi-Bicchierai, 2015).

From any seemingly innocuous and anonymized dataset can come many unexpected and highly personal forms of information, but this fact has not hampered the collection of images and text. As success in machine learning has come to rely on ever-larger datasets, more people are seeking to acquire them. But why does the wider AI field accept this practice, despite the ethical, political, and epistemological problems and potential harms? What beliefs, justifications, and economic incentives normalized this mass extraction and general equivalence of data?

MYTHS AND METAPHORS OF DATA

The oft-cited history of artificial intelligence written by AI professor Nils Nilsson outlines several of the founding myths about data in machine learning. He neatly illustrates how data is typically described in the technical disciplines: "The great volume of raw data calls for efficient 'data-mining' techniques for classifying, quantifying, and extracting useful information. Machine learning methods are playing an increasingly important role in data analysis because they can deal with massive amounts of data. In fact, the more data the better" (Nilsson, 2009, p. 495).

Echoing Robert Mercer from decades earlier, Nilsson perceived that data was everywhere for the taking, and all the better for mass classification by machine learning algorithms (Bowker, 2005, 184–85)[35]. It was such a common belief as to have become axiomatic: data is there to be acquired, refined, and made valuable.

But vested interests carefully manufactured and supported this belief over time. As sociologists Marion Fourcade and Kieran Healy note, the injunction always to collect data came not only from the data professions but also from their institutions and the technologies they deploy:

34. When the *Financial Times* exposed the contents of this dataset, Microsoft removed the set from the internet, and a spokesperson for Microsoft claimed simply that it was removed "because the research challenge is over" (Murgia and Harlow, 2019).

35. And, as Geoff Bowker famously reminds us, "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."

> The institutional command coming from technology is the most potent of all: we do these things *because we can*.... Professionals recommend, the institutional environment demands, and technology enables organizations to sweep up as much individual data as possible. It does not matter that the amounts collected may vastly exceed a firm's imaginative reach or analytic grasp. The assumption is that it will eventually be useful, i.e. valuable.... Contemporary organizations are both culturally impelled by the data imperative and powerfully equipped with new tools to enact it (Fourcade and Healy, 2016).

This produced a kind of moral imperative to collect data in order to make systems better, regardless of the negative impacts the data collection might cause at any future point. Behind the questionable belief that "more is better" is the idea that individuals can be completely knowable, once enough disparate pieces of data are collected (Meyer and Jepperson, 2000). But what counts as data? Historian Lisa Gitelman notes that every discipline and institution "has its own norms and standards for the imagination of data" (Gitelman, 2013, p. 3). Data, in the twenty-first century, became whatever could be captured.

Terms like "data mining" and phrases like "data is the new oil" were part of a rhetorical move that shifted the notion of data away from something personal, intimate, or subject to individual ownership and control toward something more inert and nonhuman. Data began to be described as a resource to be consumed, a flow to be controlled, or an investment to be harnessed[36]. The expression "data as oil" became commonplace, and although it suggested a picture of data as a crude material for extraction, it was rarely used to emphasize the costs of the oil and mining industries: indentured labor, geopolitical conflicts, depletion of resources, and consequences stretching beyond human timescales.

Ultimately, "data" has become a bloodless word; it disguises both its material origins and its ends. And if data is seen as abstract and immaterial, then it more easily falls outside of traditional understandings and responsibilities of care, consent, or risk. As researchers Luke Stark and Anna Lauren Hoffman argue, metaphors of data as a "natural resource" just lying in wait to be discovered are a well-established rhetorical trick used for centuries by colonial powers (Stark and Hoffmann, 2019). Extraction is justified if it comes from a primitive and "unrefined" source[37]. If data is framed as oil, just waiting to be extracted, then machine learning has come to be seen as its necessary refinement process.

Data also started to be viewed as capital, in keeping with the broader neoliberal visions of markets as the primary forms of organizing value. Once human activities are expressed through digital traces and then tallied up and ranked within scoring metrics, they function as a way to extract value. As Fourcade and Healy observe, those who have the right data signals gain advantages like discounted insurance and higher standing across markets (Fourcade and Healy, 2016, p. 19)[38]. High achievers in the mainstream economy tend to do well in a data-scoring economy, too, while those who are poorest become targets of the most harmful forms of data surveillance and extraction. When data is considered as a form of capital, then everything is justified if it means collecting more. The sociologist Jathan Sadowski similarly argues that

---

36. Many scholars have looked closely at the work these metaphors do. Media studies professors Cornelius Puschmann and Jean Burgess analyzed the common data metaphors and noted two widespread categories: data "as a natural force to be controlled and [data] as a resource to be consumed" (Puschmann and Burgess, 2014). Researchers Tim Hwang and Karen Levy suggest that describing data as "the new oil" carries connotations of being costly to acquire but also suggests the possibility of "big payoffs for those with the means to extract it" (Hwang and Levy, 2015).

37. Media scholars Nick Couldry and Ulises Mejías call this "data colonialism," which is steeped in the historical, predatory practices of colonial- ism but married to (and obscured by) contemporary computing methods. However, as other scholars have shown, this terminology is double-edged because it can occlude the real and ongoing harms of colonialism (Couldry and Mejías, 2019a; 2019b; Segura and Waisbord, 2019).

38. They refer to this form of capital as "ubercapital".

data now operates as a form of capital. He suggests that once everything is understood as data, it justifies a cycle of ever-increasing data extraction: "Data collection is thus driven by the perpetual cycle of capital accumulation, which in turn drives capital to construct and rely upon a world in which everything is made of data. The supposed universality of data reframes everything as falling under the domain of data capitalism. All spaces must be subjected to datafication. If the universe is conceived of as a potentially infinite reserve of data, then that means the accumulation and circulation of data can be sustained forever" (Sadowski, 2019, p. 8).

This drive to accumulate and circulate is the powerful underlying ideology of data. Mass data extraction is the "new frontier of accumulation and next step in capitalism," Sadowski suggests, and it is the foundational layer that makes AI function (Sadowski, 2019, p. 9). Thus, there are entire industries, institutions, and individuals who don't want this frontier – where data is there for the taking – to be questioned or destabilized.

Machine learning models require ongoing flows of data to become more accurate. But machines are asymptotic, never reaching full precision, which propels the justification for more extraction from as many people as possible to fuel the refineries of AI. This has created a shift away from ideas like "human subjects" – a concept that emerged from the ethics debates of the twentieth century – to the creation of "data subjects," agglomerations of data points without subjectivity or context or clearly defined rights.

ETHICS AT ARM'S LENGTH

The great majority of university-based AI research is done without any ethical review process. But if machine learning techniques are being used to inform decisions in sensitive domains like education and health care, then why are they not subject to greater review? To understand that, we need to look at the precursor disciplines of artificial intelligence. Before the emergence of machine learning and data science, the fields of applied mathematics, statistics, and computer science had not historically been considered forms of research on human subjects.

In the early decades of AI, research using human data was usually seen to be a minimal risk[39]. Even though datasets in machine learning often come from and represent people and their lives, the research that used those datasets was seen more as a form of applied math with few consequences for human subjects. The infrastructures of ethics protections, like university-based institutional review boards (IRBs), had accepted this position for years (Federal Register, 2015). This initially made sense; IRBs had been overwhelmingly focused on the methods common to biomedical and psychological experimentation in which interventions carry clear risks to individual subjects. Computer science was seen as far more abstract.

Once AI moved out of the laboratory contexts of the 1980s and 1990s and into real-world situations – such as attempting to predict which criminals will reoffend or who should receive welfare benefits – the potential harms expanded. Further, those harms affect entire communities as well as individuals. But there is still a strong presumption that publicly available datasets pose minimal risks and therefore should be exempt from ethics review (Metcalf and Crawford, 2016). This idea is the product of an earlier era, when it was harder to move data between locations and very expensive to store it for long periods.

39. Here I'm drawing from a history of human subjects review and largescale data studies coauthored with Jake Metcalf. See Metcalf and Crawford (2016).

Those earlier assumptions are out of step with what is currently going on in machine learning. Now datasets are more easily connectable, indefinitely repurposable, continuously updatable, and frequently removed from the context of collection.

The risk profile of AI is rapidly changing as its tools become more invasive and as researchers are increasingly able to access data without interacting with their subjects. For example, a group of machine learning researchers published a paper in which they claimed to have developed an "automatic system for classifying crimes" (Seo et al., 2018). In particular, their focus was on whether a violent crime was gang-related, which they claimed their neural network could predict with only four pieces of information: the weapon, the number of suspects, the neighborhood, and the location. They did this using a crime dataset from the Los Angeles Police Department, which included thousands of crimes that had been labeled by police as gang-related.

Gang data is notoriously skewed and riddled with errors, yet researchers use this database and others like it as a definitive source for training predictive AI systems. The CalGang database, for example, which is widely used by police in California, has been shown to have major inaccuracies. The state auditor discovered that 23 percent of the hundreds of records it reviewed lacked adequate support for inclusion. The database also contained forty-two infants, twenty-eight of whom were listed for having "admitting to being gang members" (California State Auditor, 2016). Most of the adults on the list had never been charged, but once they were included in the database, there was no way to have their name removed. Reasons for being included might be as simple as chatting with a neighbor while wearing a red shirt; using these trifling justifications, Black and Latinx people have been disproportionately added to the list (Libby, 2016).

When the researchers presented their gang-crime prediction project at a conference, some attendees were troubled. As reported by *Science,* questions from the audience included, "How could the team be sure the training data were not biased to begin with?" and "What happens when someone is mislabeled as a gang member?" Hau Chan, a computer scientist now at Harvard University who presented the work, responded that he couldn't know how the new tool would be used. "[These are the] sort of ethical questions that I don't know how to answer appropriately," he said, being just "a researcher." An audience member replied by quoting a lyric from Tom Lehrer's satiric song about the wartime rocket scientist Wernher von Braun: "Once the rockets are up, who cares where they come down?" (Hutson, 2018).

This separation of ethical questions away from the technical reflects a wider problem in the field, where the responsibility for harm is either not recognized or seen as beyond the scope of the research. As Anna Lauren Hoffman writes: "The problem here isn't only one of biased datasets or unfair algorithms and of unintended consequences. It's also indicative of a more persistent problem of researchers actively reproducing ideas that damage vulnerable communities and reinforce current injustices. Even if the Harvard team's proposed system for identifying gang violence is never implemented, hasn't a kind of damage already been done? Wasn't their project an act of cultural violence in itself?" (Hoffmann, 2018). Sidelining issues of ethics is harmful in itself, and it perpetuates the false idea that scientific research happens in a vacuum, with no responsibility for the ideas it propagates.

The reproduction of harmful ideas is particularly dangerous now that AI has moved from being an experimental discipline used only in laboratories to being tested at scale on millions of people. Technical approaches can move rapidly from conference papers to being deployed in production systems, where harmful assumptions can become ingrained and hard to reverse.

Machine learning and data-science methods can create an abstract relationship between researchers and subjects, where work is being done at a distance, removed from the communities and individuals at risk of harm. This arm's-length relationship of AI researchers to the people whose lives are reflected in datasets is a long-established practice. Back in 1976, when AI scientist Joseph Weizenbaum wrote

his scathing critique of the field, he observed that computer science was already seeking to circumvent all human contexts (Weizenbaum, 1976, p. 266). He argued that data systems allowed scientists during wartime to operate at a psychological distance from the people "who would be maimed and killed by the weapons systems that would result from the ideas they communicated." (Weizenbaum, 1976, p. 275-76). The answer, in Weizenbaum's view, was to directly contend with what data actually represents: "The lesson, therefore, is that the scientist and technologist must, by acts of will and of the imagination, actively strive to reduce such psychological distances, to counter the forces that tend to remove him from the consequences of his actions. He must – it is as simple as this – think of what he is actually doing" (Weizenbaum, 1976, p. 276).

Weizenbaum hoped that scientists and technologists would think more deeply about the consequences of their work – and of who might be at risk. But this would not become the standard of the AI field. Instead, data is more commonly seen as something to be taken at will, used without restriction, and interpreted without context. There is a rapacious international culture of data harvesting that can be exploitative and invasive and can produce lasting forms of harm[40]. And there are many industries, institutions, and individuals who are strongly incentivized to maintain this colonizing attitude – where data is there for the taking – and they do not want it questioned or regulated.

THE CAPTURE OF THE COMMONS

The current widespread culture of data extraction continues to grow despite concerns about privacy, ethics, and safety. By researching the thousands of datasets that are freely available for AI development, I got a glimpse into what technical systems are built to recognize, of how the world is rendered for computers in ways that humans rarely see. There are gigantic datasets full of people's selfies, tattoos, parents walking with their children, hand gestures, people driving their cars, people committing crimes on CCTV, and hundreds of everyday human actions like sitting down, waving, raising a glass, or crying. Every form of biodata – including forensic, biometric, sociometric, and psychometric – is being captured and logged into databases for AI systems to find patterns and make assessments.

Training sets raise complex questions from ethical, methodological, and epistemological perspectives. Many were made without people's knowledge or consent and were harvested from online sources like Flickr, Google image search, and YouTube or were donated by government agencies like the FBI. This data is now used to expand facial recognition systems, modulate health insurance rates, penalize distracted drivers, and fuel predictive policing tools. But the practices of data extraction are extending even deeper into areas of human life that were once off-limits or too expensive to reach. Tech companies have drawn on a range of approaches to gain new ground. Voice data is gathered from devices that sit on kitchen counters or bedroom nightstands; physical data comes from watches on wrists and phones in pockets; data about what books and newspapers are read comes from tablets and laptops; gestures and facial expressions are compiled and assessed in workplaces and classrooms.

The collection of people's data to build AI systems raises clear privacy concerns. Take, for example, the deal that Britain's Royal Free National Health Service Foundation Trust made with Google's subsidiary DeepMind to share the patient data records of 1.6 million people. The National Health Service in Britain is a revered institution, entrusted to provide health care that is primarily free to all while keeping patient data secure. But when the agreement with DeepMind was investigated, the company was found to have

40. For more on the history of extraction of data and insights from marginalized communities, see Costanza-Chock (2020); and D'Ignazio and Klein (2020).

violated data protection laws by not sufficiently informing patients (Revell, 2017). In her findings, the information commissioner observed that "the price of innovation does not need to be the erosion of fundamental privacy rights" (Information Commissioner's Office, 2017).

Yet there are other serious issues that receive less attention than privacy. The practices of data extraction and training dataset construction are premised on a commercialized capture of what was previously part of the commons. This particular form of erosion is a privatization by stealth, an extraction of knowledge value from public goods. A dataset may still be publicly available, but the metavalue of the data – the model created by it – is privately held. Certainly, many good things can be done with public data. But there has been a social and, to some degree, a technical expectation that the value of data shared via public institutions and public spaces online should come back to the public good in other forms of the commons. Instead, we see a handful of privately owned companies that now have enormous power to extract insights and profits from those sources. The new AI gold rush consists of enclosing different fields of human knowing, feeling, and action – every type of available data – all caught in an expansionist logic of never-ending collection. It has become a pillaging of public space.

Fundamentally, the practices of data accumulation over many years have contributed to a powerful extractive logic, a logic that is now a core feature of how the AI field works. This logic has enriched the tech companies with the largest data pipelines, while the spaces free from data collection have dramatically diminished. As Vannevar Bush foresaw, machines have enormous appetites. But how and what they are fed has an enormous impact on how they will interpret the world, and the priorities of their masters will always shape how that vision is monetized. By looking at the layers of training data that shape and inform AI models and algorithms, we can see that gathering and labeling data about the world is a social and political intervention, even as it masquerades as a purely technical one.

The way data is understood, captured, classified, and named is fundamentally an act of world-making and containment. It has enormous ramifications for the way artificial intelligence works in the world and which communities are most affected. The myth of data collection as a benevolent practice in computer science has obscured its operations of power, protecting those who profit most while avoiding responsibility for its consequences.

REFERENCES

Aslam, S. 2020. *Facebook by the Numbers (2019): Stats, Demographics & Fun Facts, Omnicore*. Available at: https://www.omnicoreagency.com/facebook-statistics/(Accessed: 12 May 2022).

Baker, J.M. *et al.* 2009. 'Research Developments and Directions in Speech Recognition and Understanding, Part 1', *IEEE Signal Processing Magazine*, 26(3), pp. 75–80.

Bowker, G.C. 2005. *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Brown, P. and Mercer, R. 2013. 'Oh, Yes, Everything's Right on Schedule, Fred'. *Twenty Years of Bitext Workshop, Empirical Methods in Natural Language Processing Conference*, Seattle, Washington, October. Available at: http://cs.jhu.edu/~post/bitext (Accessed: 13 May 2022).

Bush, V. 1945. 'As We May Think.', July. Available at: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/(Accessed: 11 May 2022).

California State Auditor. 2016. *The CalGang Criminal Intelligence System*. 2015–130. Sacramento, CA. Available at: https://www.auditor.ca.gov/pdfs/reports/2015-130.pdf (Accessed: 12 May 2022).

Costanza-Chock, S. 2020. *Design Justice: Community – Led Practices to Build the World We Need*. Cambridge, MA: MIT Press.

Couldry, N. and Mejías, U.A. 2019a. 'Data Colonialism: Rethinking Big Data's Relation to the Con-temporary Subject', *Television and New Media*, 20(4), pp. 336–349. doi:https://doi.org/10.1177/1527476418796632.

Couldry, N. and Mejías, U.A. 2019b. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford, CA: Stanford University Press.

Crawford, K. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crawford, K. and Schultz, J. 2019. 'AI Systems as State Actors', *Columbia Law Review*, 119(7). Available at: https://columbialawreview.org/content/ai-systems-as-state-actors/.

Curry, S. *et al.* 2009. *NIST Special Database 32: Multiple Encounter Dataset I (MEDS-I)*. NISTIR 7679. National Institute of Standards and Technology. Available at: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir7679.pdf.

Deng, J. *et al.* 2009. 'ImageNet: A Large-Scale Hierarchical Image Database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255. doi:https://doi.org/10.1109/CVPR.2009.5206848.

D'Ignazio, C. and Klein, L.F. 2020. *Data Feminism*. Cambridge, MA: MIT Press.

Ever AI. 2018. 'Ever AI Leads All US Companies on NIST's Prestigious Facial Recognition Vendor Test', *Globe News wire*, 27 November. Available at: https://www.globenewswire.com/news-release/2018/11/27/1657221/0/en/Ever-AI-Leads-All-US-Companies-on-NIST-s-Prestigious-Facial-Recognition-Vendor-Test.html.

Federal Register. 2015. 'Federal Policy for the Protection of Human Subjects'. Available at: https://www.federalregister.gov/documents/2015/09/08/2015-21756/federal-policy-for-the-protection-of-human-subjects.

Founds, A.P. *et al.* 2011. *NIST Special Database 32: Multiple Encounter Dataset II (MEDS-II)*. NISTIR 7807. National Institute of Standards and Technology. Available at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=908383.

Fourcade, M. and Healy, K. 2016. 'Seeing Like a Market', *Socio-Economic Review*, 15(1), pp. 9–29. doi:https://doi.org/10.1093/ser/mww033.

Franceschi-Bicchierai, L. 2015. 'Redditor Cracks Anonymous Data Trove to Pinpoint Muslim Cab Drivers', *Mashable*, 28 January. Available at: https://mashable.com/2015/01/28/redditor-muslim-cab-drivers/.

Garris, M.D. and Wilson, C.L. 2005. *NIST Biometrics Evaluations and Developments*. NISTIR 7204. National Institute of Standards and Technology (NIST). Available at: https://www.govinfo.gov/content/pkg/GOVPUB-C13-1ba4778e3b87bdd6ce660349317d3263/pdf/GOVPUB-C13-1ba4778e3b87bdd6ce660349317d3263.pdf (Accessed: 10 May 2022).

Gebru, T. *et al.* 2021. 'Datasheets for Datasets', *arXiv:1803.09010* [Preprint]. Available at: https://arxiv.org/abs/1803.09010.

Gershgorn, D. 2017. 'The Data That Transformed AI Research—and Possibly the World', *Quartz*, 26 July. Available at: https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.

Gitelman, L. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, MA: MIT Press.

Grother, P. *et al.* 2017. *The 2017 IARPA Face Recognition Prize Challenge (FRPC)*. NISTIR 8197. National Institute of Standards and Technology (NIST), p. 26. Available at: https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8197.pdf.

Harvey, A. and LaPlace, J. 2015. *Brainwash Dataset*, *MegaPixels*. Available at: https://megapixels.cc/brainwash/.

Heller, N. 2017. 'What the Enron Emails Say about Us', *New Yorker*, 17 July. Available at: https://www.newyorker.com/magazine/2017/07/24/what-the-enron-e-mails-say-about-us.

Hernandez, E. 2019. 'CU Colorado Springs Students Secretly Photo-graphed for Government-Backed Facial-Recognition Research', *Denver Post*, 27 May. Available at: https://www.denverpost.com/2019/05/27/cu-colorado-springs-facial-recognition-research/.

Hoffmann, A.L. 2018. 'Data Violence and How Bad Engineering Choices Can Damage Society', *Medium*, 30 April. Available at: https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4.

Hutson, M. 2018. 'Artificial Intelligence Could Identify Gang Crimes—and Ignite an Ethical Firestorm', *Science*, 28 February. Available at: https://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm.

Hwang, T. and Levy, K. 2015. 'Hwang, Tim, and Karen Levy. " 'The Cloud' and Other Dangerous Meta-phors." Atlantic, January 20, 2015.', *Atlantic*, 20 January. Available at: https://www.theatlantic.com/technology/archive/2015/01/the-cloud-and-other-dangerous-metaphors/384518/.

ImageNet Large Scale Visual Recognition Competition (ILSVRC) (no date). Available at: https://www.image-net.org/challenges/LSVRC/.

Information Commissioner's Office. 2017. "Royal Free– Google DeepMind Trial Failed to Comply with Data Protection Law." *Information Commissioner's Office*, 3 July. Available at: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/ (Accessed: 12 May 2022).

Jaton, F. 2017. 'We Get the Algorithms of Our Ground Truths: Designing Referential Databases in Digital Image Processing.', *Social Studies of Science*, 47(6), pp. 811–840. doi:https://doi.org/10.1177/0306312717730428.

Klimt, B. and Yang, Y. 2004. 'The Enron Corpus: A New Dataset for Email Classification Research.', in Boulicat, J.-F. et al. (eds) *Machine Learning: ECML 2004*. Berlin: Springer, pp. 217–226.

Lem, S. 2003. 'The First Sally (A), or Trurl's Electronic Bard', in Gunn, J. (ed.) *The Road to Science Fiction*. Lanham.

Li, X. 2017. *Divination Engines: A Media History of Text Prediction*. Ph.D. New York University.

Libby, S. 2016. 'Scathing Audit Bolsters Critics' Fears about Secretive State Gang Database', *Voice of San Diego*, 11 August. Available at: https://www.voiceofsandiego.org/topics/public-safety/scathing-audit-bolsters-critics-fears-secretive-state-gang-database/.

Light, J.S. 1999. 'Light, Jennifer S. "When Computers Were Women." Technology and Culture40, no. 3 (1999): 455–83.', *Technology and Culture*, 40(3), pp. 455–483.

Locker, M. 2019. 'Microsoft, Duke, and Stanford Quietly Delete Databases with Millions of Faces', *Fast Company*, 6 June. Available at: https://www.fastcompany.com/90360490/ms-celeb-microsoft-deletes-10m-faces-from-face-database.

Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B. 1993. 'Building a Large Annotated Corpus of English: The Penn Treebank', *Computational Linguistics*, 19(2), pp. 313–330.

Markoff, J. 2016. 'Pentagon Turns to Silicon Valley for Edge in Artificial Intelligence', *New York Times*, 11 May. Available at: https://www.nytimes.com/2016/05/12/technology/artificial-intelligence-as-the-pentagons-latest-weapon.html.

Metcalf, J. and Crawford, K. 2016. 'Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide', *Big Data and Society*, 3(1), pp. 1–14. doi:https://doi.org/10.1177/2053951716650211.

Meyer, J.W. and Jepperson, R.L. 2000. 'The "Actors" of Modern Society: The Cultural Construction of Social Agency', *Sociological Theory*, 18(1), pp. 100–120. doi:https://doi.org/10.1111/0735-2751.00090.

Michalski, R.S. 1980. 'Pattern Recognition as Rule-Guided Inductive Infer-ence', *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2(4), pp. 349–361. doi:https://doi.org/10.1109/TPAMI.1980.4767034.

Mitchell, M. *et al.* 2019. 'Model Cards for Model Reporting', in. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta: ACM Press, pp. 220–229. doi:https://doi.org/10.1145/3287560.3287596.

Murgia, M. and Harlow, M. 2019. 'Who's Using Your Face? The Ugly Truth about Facial Recognition', *Financial Times*, 19 April. Available at: https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e.

National Institute of Standards and Technology (NIST). 2010. *Special Database 32—Multiple Encounter Dataset (MEDS)*. Available at: https://www.nist.gov/itl/iad/image-group/special-database-32-multiple-encounter-dataset-meds (Accessed: 10 May 2022).

Nilsson, N.J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York: Cambridge University Press.

Phillips, J.P., Rauss, P.J. and Der, S.Z. 1996. *FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results*. ARL-TL-995. Adelphi, M.D.: Army Research Laboratory. Available at: https://apps.dtic.mil/dtic/tr/fulltext/u2/a315841.pdf.

Puschmann, C. and Burgess, J. 2014. 'Big Data, Big Questions: Metaphors of Big Data', *International Journal of Communication*, 8, pp. 1690–1709.

Raji, I.D. and Buolamwini, J. 2019. 'Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products', in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435.

Revell, T. 2017. 'Google DeepMind's NHS Data Deal "Failed to Comply" with Law', *New Scientist*, 3 July. Available at: https://www.newscientist.com/article/2139395-google-deepminds-nhs-data-deal-failed-to-comply-with-law/.

Russell, A. 2014. *Open Standards and the Digital Age: History, Ideology, and Networks*. New York: Cambridge University Press.

Russell, S.J. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach*. 3rd edn. Upper Saddle River, N.J.: Pearson.

Sadowski, J. 2019. 'When Data Is Capital: Datafication, Accumulation, and Extraction', *Big Data and Society*, 6(1), pp. 1–12. doi:https://doi.org/10.1177/2053951718820549.

Satisky, J. 2019. 'A Duke study recorded thousands of students' faces. Now they're being used all over the world', *Duke Chronicle*, 12 June. Available at: http://shorturl.at/gmq09.

Segura, M.S. and Waisbord, S. 2019. 'Between Data Capitalism and Data Citizenship', *Television and New Media*, 20(4), pp. 412–419.

Sekula, A. 1986. 'The Body and the Archive', *MIT Press*, 39(October), pp. 3–64. doi:https://doi.org/10.2307/778312.

Seo, S. *et al.* 2018. 'Partially Generative Neural Networks for Gang Crime Classification with Partial Information', in *roceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. *roceAAAI/ACM Conference on AI, Ethics, and Society*, pp. 257–263.

Stark, L. and Hoffmann, A.L. 2019. 'Data Is the New What? Popular Metaphors and Professional Ethics in Emerging Data Culture', *Journal of Cultural Analytics*, 1(1). doi:https://doi.org/10.22148/16.036.

Tockar, A. 2014. 'Riding with the Stars: Passenger Privacy in the NYC Taxi-cab Dataset', 15 September. Available at: https://agkn.wordpress.com/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.

Weizenbaum, J. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, CA: W. H. Freeman.

Wood III, P., Massey, W.M. and Brownell, N.M. 2003. *FERC Order Directing Release of Information.* Federal Energy Regulatory Commission. Available at: https://www.caiso.com/Documents/FERCOrderDirectingRelease-InformationinDocketNos_PA02–2-000_etal__Manipulation-ElectricandGasPrices_.pdf.

Zhang, Z. *et al.* 2017. 'Multi-Target, Multi-Camera Tracking by Hierarchical Clustering: Recent Progress on DukeMTMC Project', *arXiv:1712.09531* [Preprint]. Available at: https://arxiv.org/abs/1712.09531.

INNOVATION ECOSYSTEMS FOR SOCIALLY BENEFICIAL AI

YOSHUA BENGIO

Full Professor at Université de Montréal, Founder and Scientific Director of Mila – Québec Artificial Intelligence Institute, Scientific Director of IVADO – l'Institut pour la Valorisation des Données, Co-Director of the CIFAR Learning in Machines & Brains program, member of OBVIA.

ALLISON COHEN

Senior Applied AI Project Manager, AI for Humanity at Mila – Quebec Artificial Intelligence Institute

BENJAMIN PRUD'HOMME

Executive Director, AI for Humanity at Mila – Quebec Artificial Intelligence Institute.

AMANDA LEAL DE LIMA ALVES

Researcher, AI for Humanity at Mila – Quebec Artificial Intelligence Institute.

NOAH ODER

Master's student at McGill University and Sciences Po Paris.

SDG4 - Quality Education
SDG5 - Gender equality
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG17 - Partnerships for the Goals

INNOVATION ECOSYSTEMS FOR SOCIALLY BENEFICIAL AI

ABSTRACT

Governments have been preoccupied with the disruptive potential that AI technology confers, wanting to either maximize the creation of economic growth or minimize the risk of rights-related violations. Consequently, governments and international institutions have focused their efforts either on funding the AI industry broadly or cracking down on adverse applications. However, these approaches result in insufficient attention being paid to the ways in which AI can contribute to socially beneficial discoveries in fields as crucial as drug discovery, climate change, and education. A focus on social impact when investing in, and developing, innovation ecosystems is still a missing link in the AI development and governance landscape and prevents governments from enacting public policies that would otherwise promote socially meaningful innovation in AI.

This chapter is designed to raise awareness about the potential for AI to contribute to meaningful social change. It also provides a series of recommendations, which are built to support an innovation ecosystem that promotes AI for social good projects. The seven recommendations put forward seek to achieve three main objectives: i) enable informed and high-skilled engagement in the field of AI; ii) promote multidisciplinary collaboration across the AI-development value chain; and, iii) reward actors for contributing to this innovation ecosystem.

INTRODUCTION

The advent of artificial intelligence (AI) provides society with numerous opportunities for social benefit, notably by enhancing the speed and lowering the cost of decisions derived from data (not to mention, opening the door to completely new data-driven products and services). However, the current ecosystem for AI development presents several obstacles to the realization of that potential. In the current economic framework, actors' decisions are guided by the pursuit of profits, yielding insufficient innovation in areas of high positive social impact but low economic value, a phenomenon that can be compared to a "tragedy of the commons" (Llyod, 1833) or market failure scenario.

As will be suggested by this article, governments should drive the creation of ecosystems that fill the gap in socially beneficial AI. These ecosystems should be designed to improve the well-being of citizens and reduce the strain on social services ranging from healthcare to education. Governments can develop these ecosystems through the use of guidelines, norms, and incentive frameworks that influence AI's development, catalyzing its application in ways that benefit society more broadly and sustainably. But the public sector will need to play a more active role in guiding the field of AI development to realize this potential.

Major opportunities for involvement, as will be explored throughout this paper, include strategic incentives that foster ethical and socially beneficial AI research and development. To benefit from these incentives, stakeholders would need to abide by certain terms, such as open science, which are designed to accelerate the progress of beneficial technologies and their deployment. It is anticipated that this strategy can reorient the AI industry to improve the likelihood and prevalence of AI-based tools that satisfy pressing social needs. What's more, this strategy can address the current missing link in governments' approach to innovation, specifically, that of a social benefit focus.

The ideas expressed in this chapter should be applied in ways that respect a country's unique context. The recommendations presented aim to guide the public sector, so all stakeholders – including the private sector and civil society – can strive for an ecosystem where socially beneficial AI is actively fostered. We argue that such development depends on meaningful engagement from the public sector, based on the economic theory of the tragedy of the commons and current trends in the AI industry. The following sections will present the meaning of "AI for social good," the relevance of governments' involvement in the early stages of AI development worldwide, and the proposed recommendations to address the current missing link in terms of governments' leadership in the fast-growing AI field.

THE MEANING OF "AI FOR SOCIAL GOOD"

In order to clarify the types of applications being promoted in this chapter, it is important to first define what is meant by AI for social good. The definition that best articulates our notion of the concept states that AI for social good projects involve AI systems that are designed, developed, deployed, monitored and evaluated in order to: "(i) prevent, mitigate and/or resolve problems adversely affecting human life and/or the wellbeing of the natural world, and/or (ii) enable socially preferable or environmentally sustainable developments, while (iii) not introducing new forms of harm and/or amplifying existing disparities and inequities" (Cowls et al., 2021). In this context, one can look to the 17 United Nations Sustainable Development Goals (SDGs) as a useful framework for categorizing domains that are globally considered to be socially beneficial. These goals were agreed upon by all 193 member states of the United Nations and created as a blueprint to catalyze economic, social and environmental progress.

AI'S CURRENT TRAJECTORY

The current economic framework – characterized by the dominance of market forces and the reliance of states on private investors to identify worthwhile R&D projects – has, so far, played a significant role in defining the incentive structure guiding AI's development. Unfortunately, this framework prioritizes the creation of AI tools that are primarily aligned with economic returns, addressing social needs only when it is profitable to do so.

While the last decades have seen a growing trend of private sector engagement with social initiatives (Porter and Kramer, 2006, p.3), this programming largely takes place at the margin of core business activities and generally does not achieve the scale that is needed for addressing several important long-term challenges for our societies. AI tools are currently being developed with three common shortcomings, including: i) unanticipated failures; ii) missed opportunities; and iii) burdensome interventions (Cowls et al., 2021). These shortcomings are manifestations of the AI development incentive structure, which has skewed the technology in the direction of profit to the detriment of social good.

In terms of unanticipated failures, technology which has not placed "social good" at its core is being deployed with unforeseen and sometimes adverse social consequences. A good example is Microsoft's "Tay" chatbot launched to Twitter in 2016. Tay was designed to learn from human users and generate its own content in the style of a teenage girl. The developers' goal was to teach Tay to have conversations with humans about nearly any topic. However, within 24 hours of the launch, the bot was removed from the Internet because it began sharing racist, misogynistic, homophobic and otherwise offensive Tweets (Schwartz, 2019). This example is one in a long list of AI applications that create harms which aren't proactively mitigated by AI developers and therefore represent unanticipated failures.

In addition to negative outcomes, there is a significant number of missed opportunities related to the use and deployment of AI in contexts where social benefit is not prioritized. As an example, AI tools are being developed in the healthcare industry to detect skin cancer in patients. However, given the skew in resources, data, and market incentives, one such tool performed well on light skin but poorly on those with darker skin (Adamson and Smith, 2018). By creating these tools without placing the value of diversity and inclusivity at the center, these tools are missing opportunities to enhance the quality of care being provided to marginalized and vulnerable communities who, in many cases, stand most to benefit from this newfound technology, given the disparity that already exists in healthcare.

Finally, in terms of burdensome interventions, AI tools are being developed with objectives and outcomes that do not provide any clear benefit to society. In fact, these tools can sometimes be the cause of serious social harm. For example, an AI tool was developed to detect, with significant accuracy, whether someone was homosexual (Wang and Kosinski, 2017). This type of tool makes it possible to surveil people using highly personal information that they may or may not choose to disclose publicly. This information can be released to abuse, ostracize or otherwise harm members of the LGBTIQ community. Burdensome interventions are particularly concerning in a scenario where actors developing AI may decide at their own discretion to create, release and maintain problematic AI interventions.

In order to address the market structures that are enabling these types of problematic AI tools to proliferate, many in the industry are turning to regulation, and rightfully so. The risks posed by problematic, unregulated as well as underregulated AI systems are concerning. The policies, programs and initiatives that are being created as a response are critically important and make us hopeful that the industry will be less likely to develop in ways that are harmful moving forward. However, this article is not meant as a contribution to that important body of work. Rather, the intention for this article is to raise awareness about a new approach that governments can take when guiding AI's development. Namely, one in which positive social impact is prioritized to enable particular AI applications to proliferate; specifically, those that benefit society and are currently not appealing to market actors.

Innovation in AI for Social Good is not currently being realized to its full potential. In fact, as the current laissez-faire approach in the AI industry has demonstrated, the market is unlikely to develop socially beneficial technologies with the size and scope that are needed to contribute to addressing some of our most intractable global challenges, ranging from climate change to education and health. Thus, to develop the AI industry faster and in ways that are more robust and socially beneficial, governments must become more actively involved in steering AI development towards addressing meaningful social challenges.

To illustrate how the current system is not built to maximize societal benefit, we can look to the example of antimicrobial resistance (AMR) and the associated drug discovery pipeline. The case of AMR demonstrates how laissez-faire capitalism fails to generate research and development (R&D) in areas that are of critical importance to society; areas where innovation might otherwise save the world from complex and significant challenges.

Antibiotics have revolutionized healthcare since the 1950s, saving countless lives both directly and indirectly (e.g., by enabling safe surgery). However, the bacteria that antibiotics are designed to fight have begun fighting back. In fact, through a process of evolution, bacteria eventually mutate into strains that are resistant to the antimicrobial drug. Those resistant variants can proliferate by virtue of evading antibiotics. In 2019, antimicrobial resistance (AMR) was associated with the death of 4.95 million people. Of those cases, AMR was the direct cause of death for 1.27 million people, a number that is anticipated to climb (Murray et al., 2022; World Health Organization, 2019b). If nothing changes to our use of antibiotics and the current drug discovery pipeline, it is anticipated that antibiotic resistant bacteria will cause 10 million deaths per year by 2050 (Review on Antimicrobial Resistance, 2014). For comparison's sake, the COVID-19 pandemic has so far been responsible for the death of an estimated 5.6 million people (World Health Organization, 2022).

Given the human cost and the consequent economic impact of antibiotic resistant bacteria proliferating, the social value of technologies to prevent and mitigate AMR would also be tremendous. Unmitigated, it is estimated that AMR will cause global GDP to decline by 2 to 3.5% per year, which, when accumulated until 2050, represents a 60 to 100 trillion dollar decrease in the exchange of global goods and services (Review on Antimicrobial Resistance, 2014). Keep in mind that these estimates might be conservative; there is a chance that a variant may be so deadly and transmissible that it could threaten the entire human species, not to mention economic order and social organization.

One obvious question is: why don't pharmaceutical companies just invest in R&D for antibiotics that are effective against the current and future mutated bacteria? Although this research could save an untold number of lives, money, and potentially even the social order as we know it, it is not profitable in the current market conditions. In fact, for pharmaceutical companies to recoup their investments, there must be significant demand for their drugs. However, in the case of antimicrobial resistance, the number of people initially infected with a mutated strain is often only a small percentage of the infected population (Plackett, 2020). Thus, even though new antibiotics would prevent the spread of a new strain of the virus, it would simultaneously stunt the demand for the new drug, by preventing the strain from multiplying.

In addition, doctors rightfully prescribe existing antibiotics as a first line of defense to delay the onset of mutations providing resistance to the new antibiotics, further reducing the market size for new antibiotics. Consequently, this market scenario is not sufficiently interesting for drug developers, whose profit is normally directly proportional to the number of potential consumers, which is thus a major criteria to assess a drug's potential profitability. This is especially the case in the context of antibiotic drug development, which, unlike other drug categories, are sold at very low prices. In fact, traditionally, there has only been room in the marketplace for one profitable drug per bacterial infection (McKenna, 2020). As a result, there is not enough R&D into drugs that would be effective against lethal mutations until it is too late (World Health Organization, 2019a). The irony is that these drugs *would* end up being developed and deployed at scale once the mutated strain proliferates, but only once there has been

a significant human, social, political and economic cost, since the development of a new drug can take a decade. Thus, to pre-emptively address this societal challenge, governments' intervention in drug discovery R&D towards such socially important objectives is critical[41].

The case of antimicrobial resistance demonstrates how a major misalignment can occur between social needs and financial returns. When it does, markets can fail to generate the products that are so desperately needed by society, both in economic and human terms. Thus, the markets should not be solely relied upon to invent the technology that we, as a society, need. Rather, the public sector must be responsible for stimulating R&D in ways that are highly efficient from the perspective of social impact.

When it comes to new AI applications, their potential can be explored in the context of drug discovery, given their promising capacity to accelerate the R&D process – which currently takes an average of 10 years – and reduce the cost of drug discovery, which is currently in the billions of dollars (PhRMA, 2015). Moreover, AI can contribute to the discovery of more effective drugs as this technology can explore a much larger volume of drug candidates in the molecular space.

Still, there are at least three major obstacles to realizing AI's potential. First, there is the issue of data availability. Datasets are often limited in scope and not made publicly available by companies, predominantly in order to protect their investments from the competition. Second, access to AI expertise is still insufficient, notably among start-ups and within the Global South, where a broader range of innovative studies and applications could otherwise be explored. Third, the limited size and scope of most academic research labs are an obstacle, considering that they could otherwise potentially generate meaningful contributions in the drug discovery ecosystem by way of in-house dataset creation, among other capabilities. Unlike pharmaceutical companies, university labs operate in a bottom-up way with significant freedom given to each graduate student and professor to undertake research of their choosing, which is great for basic exploratory research but not as efficient when it comes to mission-oriented R&D. On the other hand, the industrial R&D process is more top-down in order to accommodate companies' strategic objectives, an approach that has been successful in converting early-stage ideas into products. With these examples in mind, the following sections provide a brief overview of how governments can engage with AI development to address each of the barriers that are preventing AI's uptake for socially beneficial use.

THE TRAGEDY OF THE COMMONS FRAMEWORK

The tragedy of the commons is a concept coined in 1833 by the British economist William Forster Lloyd that can shed light on the need for governments' strategic involvement in driving socially beneficial outcomes within the market. Lloyd considered what would happen if every farmer, acting in their own self-interest, allowed their cattle to graze upon a common patch of grassland. Without collectively agreed upon rules for how the farmers would collaborate to maintain the grassland over time, the patch of land would quickly become depleted. This is because, in the absence of common rules, the consumption of grass would become a zero-sum game (Llyod, 1833). As a result, the farmers would be incentivized to continue sending their cattle to graze to the point of depletion, even though depleting common grassland would ultimately lead to the destruction of the resource and is ultimately in no one's interest.

41. Governments have invested in research organizations including the Antimicrobial Resistance Multi Partner Fund (AMR MPTF), the Global Antibiotic Research & Development Partnership (GARDP), and the AMR Action Fund. Furthermore, governments like Sweden, Germany and the US are piloting reimbursement models to fund innovation in AMR research.

The tragedy of the commons could be solved with collectively agreed upon rules, which would influence individuals' decisions to the extent that they begin operating in ways that are more aligned with the interests of the group.

In the AI market, where the rules of the game are not well defined (LaCroix and Mohseni, 2021; Benkler, 2019), corporations' inherently profit-driven interests prevail. As a consequence, some public goods are not only depleted but their very creation and maintenance are disincentivized, leading to suboptimal social outcomes. This is particularly troublesome given AI's unbelievable potential to achieve social good. For key stakeholders to start producing more socially beneficial AI, governments must re-write the incentive structure governing stakeholders' decision-making in this field. Ultimately, the state is the only actor with sufficient influence to affect the practices of industry at the pace and scale that is needed.

Creating new incentives to propel innovation is not a new concept. In fact, the success of industries such as IT, biotechnology, and nanotechnology has depended on government investment well before private actors entered the field, sustaining the early R&D needed to bring the technology to maturation and profitability. It was therefore only after the initial risks were absorbed by the government through substantial investments in fundamental research and infrastructure that a market was generated for these and other breakthrough innovations (Mazzucato, 2013).

While it is true that many governments have already invested heavily in the AI market, these investments are generally directed towards commercially viable applications of AI. For example, governments have been funding industrial research in AI by paying a portion of the R&D costs (Government of Canada, 2018; Wiggers, 2021). This incentive structure requires that the research be sufficiently commercially appealing such that companies are motivated to incur the cost of the other portion of this work. Hence, to change the current trajectory of AI, governments must recognize their role in investing in, and promoting the uptake of, socially responsible AI since, as we have seen, this will not be achieved through the markets alone.

It is recommended that governments take a long-term approach when designing the incentives that will govern the field. This is because profit calculations involving short-term returns often exponentially discount the longer-term results that are expected. In practice, this long-term approach consists of strategically directing returns on investment back into future common good initiatives rather than focusing on short-term, profit-focused cycles, which is the direction currently pursued by many private sector actors (Lazonick and Mazzucato, 2013). Indeed, funding provided by "venture capital funds tend to be concentrated in areas of high potential growth, low technological complexity and low capital intensity, since the latter raises the cost significantly" (Mazzucato, 2013, p. 55). Unfortunately, this incentive structure is misaligned with the large-scale investments that are needed to develop the field of AI for social good. Rather, innovation in this space should be considered a cumulative process that will lead to higher quality, lower cost products only after *years* of research and industry development (Lazonick and Mazzucato, 2013).

GOVERNMENT INCENTIVES

Governments must take a leadership role in shaping the incentives governing the field of AI to drive positive change among the types of AI projects being developed. The primary levers of change can be put into practice through both positive and negative incentives. Negative incentives can include financial and non-financial penalties on socially problematic developments while positive incentives can include financial and non-financial rewards for socially beneficial behavior. Whether through positive or negative means, the incentives must be scaled at a sufficient level to appropriately influence the decisions of private sector actors.

The recommendations below are categorized according to the stakeholder group that they target (from the individual to the institutional to the societal). Each of these stakeholders plays a unique and valuable role in the innovation ecosystem and must be mobilized to generate desirable change. The recommended points of intervention are built to achieve three main objectives governments should pursue:

- Enable informed and high-skilled engagement in the field of AI (*Recommendations 1-3*);
- Promote multidisciplinary collaboration across the value chain (*Recommendations 4-5*);
- Reward actors for contributing to an ecosystem that promotes socially beneficial AI (*Recommendations 6-7*).

- Recommendation 1: Train talent and expertise at all levels, from basic digital literacy to highly qualified AI personnel in universities, combining both social awareness and technical education.

Globally, expertise in AI is scarce. Not to mention, those who are skilled in AI are concentrated among particular countries, sectors, industries and demographic groups, which has implications for the types of AI applications that are being developed (World Bank Group, 2021). A starting point to address the scarcity in the talent pipeline, and to foster a global community that can leverage AI for social development, is investing in digital literacy. As defined by UNESCO (2018), "digital literacy is the ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately through digital technologies for employment, decent jobs and entrepreneurship. It includes competences that are variously referred to as computer literacy, ICT literacy, information literacy and media literacy." There have been efforts to enhance AI literacy among various groups within the population; namely those who do not have a technical background (Kong et al., 2021), are members of underrepresented groups in the industry (Office for Students, 2020), and would not otherwise learn AI as part of the standard curriculum (Lee et al., 2021). Each of those initiatives has seen encouraging results that should be explored further.

AI skills are also highly concentrated along geographic lines. One can observe this phenomenon when analyzing the concentration of AI outputs among the small number of countries that host the overwhelming majority of AI talent (World Bank Group, 2021). What's more, often, skilled workers from the Global South move to find work in the Global North, resulting in a brain drain that has been acutely felt in countries without AI research and industry hubs (McKinsey Global Institute, 2020). The trends also indicate that AI skills are overwhelmingly concentrated among men. According to the World Economic Forum's (2020) Global Gender Gap Index, women constitute only 26 percent of the data and AI workforce globally. In Canada, the gender disparity among data and AI professionals is 70 percent men and 30 percent women. In academia, the gap is even wider. In fact, according to the Global AI Talent Report (Hudson and Mantha, 2020), women have only authored 15 percent of AI papers on arXiv, an open-access archive widely used within the AI community.

The scarcity and skew of AI talent is resulting in AI applications that are designed by and for some and not others; as well as a growing discrepancy in the concentration of wealth and power (Crawford, 2021). According to the International Data Corporation (IDC), worldwide revenues from the AI market are expected to surpass $300 billion in 2024 (Savage, 2020). For greater numbers of people to benefit from this economic opportunity, skills development is key (OECD, 2015).

It is recommended that governments make digital literacy and AI training more widely available and accessible across demographic groups *and* ensure people are equipped to engage with the downstream ethical and social consequences of the tools they create. It is important when training AI talent to raise

awareness of downstream consequences in order to dissuade problematic applications and promote socially beneficial ones. Furthermore, this training would allow AI practitioners to more systematically consider the potential misuses of their tools and take measures to mitigate those risks beforehand[42].

- **Recommendation 2:** Feed the knowledge discovery pipeline in socially important applications, from exploratory fundamental research in AI to its adoption in the industry.

While the opportunity for AI-fueled growth is widely understood within the industry, adoption rates are relatively slow-moving over concerns regarding the technology's lack of maturity and fast-paced development (Deloitte, 2019). However, by integrating the knowledge discovery pipeline, from exploration to implementation, governments can catalyze industrial investment in internal AI capabilities and thereby bolster the technology's uptake.

In order to integrate stakeholder groups, governments should incentivize the use of AI solutions across the value chain (i.e., in research, development and deployment). This can be done by creating an ecosystem wherein independent researchers are given grants to explore, test and develop novel and socially beneficial algorithms that can be leveraged within industry.

For example, in order to integrate stakeholders throughout the AI value chain, Canada established the Pan-Canadian AI Strategy, a $125 million initiative that seeks to drive Canadian leadership in the field of AI. The major lever of change involves building local and regional AI ecosystems that support AI talent, foster industry uptake, and build a broader understanding of the social implications of AI across the value chain (CIFAR, 2017). Mila – Québec Artificial Intelligence Institute is an example of a research hub that contributes to a wider AI ecosystem in Montréal and beyond by developing fundamental and applied research in AI, with more than 900 researchers and a host of industry partners from startups to well-established technology companies[43].

- **Recommendation 3:** Embolden the AI ecosystem through poles of excellence in AI research and training.

In order to build innovation ecosystems, countries must attract high quality talent, provide them with the resources and support their needs so they can sustain themselves over time. The European Commission (2020) articulated a strategy in this regard by setting out a series of recommended measures to achieve an "ecosystem of excellence" along the entire knowledge and value chain. They argue that unlike the fragmented landscape that currently characterizes centers of excellence around Europe, a pan-European approach could achieve the scale that is needed to compete with leading institutes globally. According to the Commission, a centralized approach would enable stronger training and attraction of researchers, which would lead to the development of high-quality technology and unlock significant investments in AI (European Commission, 2020). Such a model could serve as inspiration for other regions of the world, in which resources might be scarce and a cross-national effort could enable a stronger ecosystem that develops socially beneficial AI to tackle common challenges.

42. Canada's Algorithmic Impact Assessment is a useful risk assessment tool that can help AI developers uncover areas of risk that should be mitigated before deployment. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

43. For more information about Mila, visit the website: https://mila.quebec/en/

- **Recommendation 4:** Finance and integrate different parts of the pipeline to minimize leakages across the value chain, especially in pursuit of relevant societal goals.

If governments want to successfully build market-generating endeavors, their funding must go beyond early-stage research. Throughout history, public investments that have generated technological revolutions (in fields such as IT, biotech, and nanotech) engaged with stakeholders across the entire innovation chain. However, in general, there is still a need to address leakages – of brains, untapped ideas and promising startups, for instance – across the value chain (Spicer et al., 2018). An example of integrated process can be observed in the United States, where the National Science Foundation (NSF) performed fundamental research, which was sent to DARPA and the National Institutes of Health (NIH) to be further developed, tested and applied. Then, through agencies like Small Business Innovation Research (SBIR), companies seeking to bring the research into production received the early-stage funding that was needed to do so (Block and Keller, 2011). Thus, innovation must be strategically encouraged across the R&D life cycle in order for governments to succeed in fostering new markets.

However, without proper measurement and tracking, it is very difficult to ensure that governments' engagement has been impactful. As such, governments should measure the success of their intervention according to the economic, social and environmental benefits that are generated as a result. While it can be politically challenging to emphasize the long-term benefits of the investment, it is a worthwhile exercise to ensure that the markets being created are ones that provide the most benefit to society.

- **Recommendation 5:** Facilitate the growth of the AI for social good startup ecosystem and foster its connection to the industry[44].

The private sector plays a prominent role in the AI innovation ecosystem. However, private sector actors could be doing more to achieve greater success in this domain. Specifically, stakeholders should engage in a symbiotic relationship with one another that leverages their respective strengths for optimal industrial growth. The strengths among large companies include their ability to afford top talent, computing facilities, and experimental labs, all of which produce high quality data and generate exciting AI products. Startups are also critical as they focus on niche areas of AI development, which often contain the most cutting-edge AI technology. Furthermore, they are incredibly dynamic and can respond quickly to changing needs, whether in processes, talent, or operations.

Organizations such as the European AI Startup Landscape provide an interesting example of how to foster connections between start-ups, large enterprises and venture capital (European AI Startup Landscape, n.d.). The promise of these partnerships is that they can embolden the ecosystem for innovation and drive technological diffusion in new areas (World Bank Group, 2021). In addition to helping with the general development of the AI startup ecosystem, governments could leverage their financial contributions to encourage socially beneficial innovation based on their governmental priorities.

- **Recommendation 6:** Stimulate research and innovation in fields where there is great societal value but too little commercial value for companies.

As was mentioned earlier, there are fields of research in AI that are being underdeveloped, not because they do not offer clear societal benefit but because they are not sufficiently commercially appealing for companies. To ensure that the AI for social good ecosystem is robust, governments must be responsible for identifying these fields of research and incentivizing relevant stakeholders to engage.

44. The European Startup Landscape is an interesting example of developing a network of AI startups and establishing a dynamic ecosystem with other stakeholders such as industry. See: https://www.ai-startups-europe.eu/

Often, being able to identify promising avenues for R&D requires the expertise of those who specialize in AI and the application domain. Thus, it is recommended that governments partner with centers of excellence whose teams of in-house experts can identify underexplored areas of research with socially relevant implications. It is also recommended that governments leverage those teams to evaluate the merits of grant applications[45]. A partnership of this nature has the potential to spur greater economic growth and development than governments might otherwise realize on their own.

For fields of research that require particularly sizable investments, it is recommended that government-funded organizations take form at the international level to independently determine the most strategic directions of AI innovation. These organizations would be responsible for drafting innovation procurement contracts and defining metrics of success. The independence of these organizations would provide them with the freedom to plan according to a longer-term time horizon with fewer political demands. Nonetheless, these organizations would be responsible for consistent and transparent reporting on their activities to ensure they are held accountable for their decisions.

- **Recommendation 7:** Establish a framework to promote the sharing of knowledge and data between actors while maintaining data privacy.

Even if governments were to achieve the first six recommendations, there would still be a whole host of missed opportunities caused by limited access to important datasets. That is because, without access to appropriate data, it can be impossible to train machine learning models to perform accurately.

Unfortunately, actors that collect data often attempt to keep it a secret or retain a monopoly over it. As a result, these actors "create an artificial scarcity in knowledge in exactly the same way that a baker's cartel creates an artificial scarcity in bread" (Maurer, 2003, p. 175). When the owner of an intellectual property restricts how it may be used, a whole host of inefficiencies to innovation might occur. For example, without data sharing, stakeholders who have the expertise, imagination, and material facilities needed to create innovative AI products might be unable to do so or achieve optimal results without access to the datasets needed. This bottleneck stunts further work and can have spillover effects on the ecosystem.

It is essential to improve the access to and the management of data to enable the development of AI and other digital applications. In Europe, for instance, a report investigated the extent of an opportunity cost from the lack of interoperable data. By looking into seven indicators – time spent, cost of storage, license costs, research retraction, double funding, interdisciplinarity and potential economic growth – the study revealed that the estimated cost of not sharing data reached 10 billion euros annually (European Commission, 2020). It is for this reason that the OECD's Committee for Scientific and Technological Policy has been arguing that access to data should become a major policy priority within the OECD (OECD, 2021).

One lever that the government can use to enable data sharing is through Requests for Proposals (RFPs). In awarding RFPs to AI companies, governments can require that, as a condition of receiving a bid, the recipients make all datasets that the project generates openly available. While this condition may result in the need to better compensate the grant recipient, the long-term benefits of these policies are likely to outweigh the cost. It is estimated that open data can unlock $3 trillion globally each year in economic value by contributing to innovation in every sector of the economy (McKinsey, 2014, p.10). That is because greater access to data lowers the barriers to working in AI and increases competition with new market entrants joining the industry. The financial benefits of this policy manifest in new revenue

45. The National Institute of Health (NIH) (medical research agency in the United States) developed a National Center of Excellence under their "Bridge2AI" program in order to help them catalyze promising research.

sources, savings and economic surplus in domains ranging from education to transportation. With new market entrants and greater innovation, governments are creating an enabling environment for the development of new, socially beneficial AI products.

CONCLUSION

AI is an incredibly powerful tool that has the potential to generate socially beneficial discoveries in critically important fields, from education and the environment to healthcare. However, for these opportunities to be realized, governments must actively shape the trajectory of AI research and development by engaging all stakeholders within the AI ecosystem. Otherwise, industrial stakeholders are left to decide for themselves how the field of AI will develop, which tends to marginalize innovations that have great societal value when they are not sufficiently commercially attractive. Governments should act to re-orient this industry: they should invest in AI literacy and education, set up a well-integrated, multi-stakeholder ecosystem, create sufficient incentives along the pipeline to engage and maintain talent, inspire AI for social good applications and promote data sharing. With this approach, society can begin to harness the promise of AI as a tool for social *as well as* economic development.

REFERENCES

Adamson, A. and Smith, A. 2018. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology.* Vol. 154, No. 11, pp. 1247–1248. DOI:10.1001/jamadermatol.2018.2348

Benkler, Y. 2019. *Don't let industry write the rules for AI*. https://www.nature.com/articles/d41586-019-01413-1

Block, F. L., and M. R. Keller. 2011. *State of innovation: The U.S. government's role in technology development*. Boulder, CO: Paradigm Publishers.

CIFAR. 2017. *Pan-Canadian AI Strategy*. https://cifar.ca/ai/

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. 2021. A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, *3*(2), 111-115.

Deloitte. 2019. *Future in the balance? How countries are pursuing an AI advantage*. https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/ai-investment-by-country.html

European AI Startup Landscape. n.d. *Motivation*. https://www.ai-startups-europe.eu/

European Commission. 2020. *On artificial intelligence – A European approach to excellence and trust*. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Economic and Social Committee. 2021. How the Digital Transformation can put Humans at the Centre of Robotics and Automation: collaboration between humans and machines for better quality products and services. April 2021. doi: 10.2864/733324

Government of Canada. 2018. *Government of Canada invests in artificial intelligence and start-up innovation across Canada*. Ottawa, Innovation, Science and Economic Development Canada. https://www.canada.ca/en/innovation-science-economic-development/news/2018/10/government-of-canada-invests-in-artificial-intelligence-and-start-up-innovation-across-canada.html

Hudson, S. and Mantha, Y. 2020. *Global AI Talent Report 2020*. https://jfgagne.ai/global-ai-talent-report-2020/

Kong, S. C., Cheung, W. M. Y. and Zhang, G. 2021. Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence*, 100026.

LaCroix, T., and Mohseni, A. 2020. *The Tragedy of the AI Commons*. arXiv preprint arXiv:2006.05203.

Lazonick, W. and Mazzucato, M. 2013. The risk-reward nexus in the innovation-inequality relationship: who takes the risks? Who gets the rewards?. *Industrial and Corporate Change*, Vol. 22, No. 4, pp. 1093-1128.

Lee, I. et al. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pp. 191-197.

Lloyd, W. F. 1833. *Two lectures on the checks to population*. JH Parker.

Maurer, S. 2003. Designing Public–Private Transactions that Foster Innovation. Esanu, J.M. and Uhlir, P.F. (eds). *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*. National Academies Press, pp. 175-79.

Mazzucato, M. 2013. *The entrepreneurial state: Debunking the public vs. private myth in risk and innovation*. London: Anthem Press.

McKenna, M. 2020. *The antibiotic paradox: why companies can't afford to create life-saving drugs*. https://www.nature.com/articles/d41586-020-02418-x

McKinsey Global Institute. 2020. *How to Ensure Artificial Intelligence Benefits Society: A Conversation with Stuart Russell and James Manyika*. https://www.mckinsey.com/featured-insights/artificial-intelligence/how-to-ensure-artificial-intelligence-benefits-society-a-conversation-with-stuart-russell-and-james-manyika

Murray, C. J. L., et al. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, Vol. 399, No. 10325, pp. 625-655. DOI:https://doi.org/10.1016/S0140-6736(21)02724-0

OECD. 2015. Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en

OECD. 2021. *Recommendation of the OECD Council concerning Access to Research Data from Public Funding*. https://www.oecd.org/sti/recommendation-access-to-research-data-from-public-funding.htm

Office for Students. 2020. *Apply now – new courses in artificial intelligence and data science*. https://www.officeforstudents.org.uk/news-blog-and-events/press-and-media/apply-now-new-courses-in-artificial-intelligence-and-data-science/

PhRMA. 2015. *Biopharmaceutical Research & Development: The Process Behind New Medicines*. http://phrma-docs.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf

Plackett, Benjamin. 2020. Why Big Pharma Has Abandoned Antibiotics. *Nature Outlook: Antimicrobial Resistance*, 21 October. https://www.nature.com/articles/d41586-020-02884-3.

Porter, M. E. and Kramer, M. R. 2006. Strategy & Society: The Link between Competitive Advantage and Corporate Social Responsibility. *Harvard Business Review*, December 2006. https://hazrevista.org/wp-content/uploads/strategy-society.pdf.

Review on Antimicrobial Resistance. 2014. *Antimicrobial resistance: tackling a crisis for the health and wealth of nations*. Review on Antimicrobial Resistance.

Savage, N. 2020. The Race to the Top among the World's Leaders in Artificial Intelligence. *Nature*, December. https://www.nature.com/articles/d41586-020-03409-8.

Schwartz, O. 2019. In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. *IEEE Spectrum*, 25 November. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

Spicer, Z. et al. 2018. *Reversing the Brain Drain: Where is Canadian STEM Talent Going?*. https://brocku.ca/social-sciences/political-science/wp-content/uploads/sites/153/Reversing-the-Brain-Drain.pdf

UN General Assembly. 2015. *Transforming our world: the 2030 Agenda for Sustainable Development*. https://www.refworld.org/docid/57b6e3e44.html

Wang, Y. and Kosinski, M. 2022. *Deep neural networks are more accurate than humans at detecting sexual orientation from facial images*. OSF, 23 June. https://doi.org/10.17605/OSF.IO/ZN79K

Wiggers. K. 2021. *U.S. agencies are increasing their AI investments*. AI Weekly. San Francisco, VentureBeat. https://venturebeat.com/2021/09/11/ai-weekly-u-s-agencies-are-increasing-their-investments-in-ai/#:~:text=R%26D%20spending%20reached%20%241.2%20billion,by%20a%20combined%20%2481%20million.

World Bank Group. 2021. *Harnessing Artificial Intelligence for Development in the Post-COVID-19 Era. A Review of National AI Strategies and Policies*. https://thedocs.worldbank.org/en/doc/2e658ef2144a05f30e254221ccaf7a42-0200022021/original/DD-Analytical-Insights-Note-4.pdf

World Economic Forum. 2020. *Data Explorer: Global Gender Gap Index*. http://reports.weforum.org/global-gender-gap-report-2020/dataexplorer/

World Health Organization. *2019a. 2019 Antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline*. https://apps.who.int/iris/bitstream/handle/10665/330420/9789240000193-eng.pdf

——. 2019*b*. *New report calls for urgent action to avert antimicrobial resistance crisis.* https://www.who.int/news/item/29-04-2019-new-report-calls-for-urgent-action-to-avert-antimicrobial-resistance-crisis

——. 2021*a*. *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer

——. 2021*b*. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

——. 2022. *WHO Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/

A MANIFESTO CONCERNING ARTIFICIAL INTELLIGENCE FOR MONITORING SUSTAINABLE DEVELOPMENT: THE MISSING LINK BETWEEN SDGS, INVESTMENT AND TRUST

JOHN SHAWE-TAYLOR

Professor of Computational Statistics and Machine Learning and UNESCO Chair of Artificial Intelligence at University College London and Director of the International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO at the Jožef Stefan Institute in Slovenia.

DANIEL MIODOVNIK

Director at Social Finance and co-founder of their Digital Labs, Advisor to the International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO.

DAVOR ORLIC

Honorary Research Assistant at UCL Centre for Artificial Intelligence and Chief Operations Officer at the International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO at the Jožef Stefan Institute in Slovenia.

SDG6 - Clean Water and Sanitation
SDG7 - Affordable and Clean Energy
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG12 - Responsible Consumption and Production
SDG13 - Climate Action
SDG15 - Life on Land
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

A MANIFESTO CONCERNING ARTIFICIAL INTELLIGENCE FOR MONITORING SUSTAINABLE DEVELOPMENT: THE MISSING LINK BETWEEN SDGS, INVESTMENT AND TRUST

ABSTRACT

We have seen an immense surge of interest in tackling the United Nations Sustainable Development Goals (SDGs). Countries, companies and investors around the world are committed to addressing the global economic, social and environmental crises. Investors have already committed US$89 trillion in assets to investments targeting SDG outcomes as part of the Principles for Responsible Investment (PRI) program.

However, there is the danger that without objective and reliable ways of assessing progress the momentum will be lost. We've seen an erosion of trust between citizens and governments, tech companies and industry alike. The lack of a consistent framework and the current subjectivity of data and ratings are holding us back.

We believe that artificial intelligence (AI) and data are fundamental to building the trustworthiness and evidence of measurable progress against the SDGs. We are already seeing examples of how clearly defined and measurable outcomes can unlock investment to solve the SDGs. For example, clear outcome metrics and data collection underpinned a $10 million outcomes contract to address rural sanitation in Cambodia (SDG Goal 6: Ensure availability and sustainable management of water and sanitation for all).

Therefore, we have designed a manifesto that calls on NGOs, the UN, companies, investors and countries to collaboratively build a robust, accessible and transparent system for measuring and certifying attainment of the SDGs. Together, we can build the AI and data ecosystem to create trust and enable investors, companies and governments to demonstrate progress, secure investment, and ultimately, change the world.

INTRODUCTION

At the moment, there is a lack of objective understanding of where exactly we stand in terms of progress towards reaching the SDGs. We believe that a universally acknowledged innovative technical mechanism, as well as a mechanism for finance and investment, would create trust between all stakeholders. The missing link is the convergence of both the usage of AI for measurement of the progress of the SDGs and social impact bonds that together can be used by governments to finance such technical endeavors. In this chapter, we propose a narrative that could potentially solve this missing link.

SOCIAL IMPACT BONDS

Poor sanitation, especially in places where open defecation routinely occurs, is linked to poor health outcomes, from spreading diseases to contaminating drinking water. To help the Royal Government of Cambodia bring safe sanitation to some of the poorest and most vulnerable households in Cambodia, Social Finance partnered with Stone Family Foundation, International Development Enterprises (iDE) and the United States Agency for International Development (USAID) to design the world's first impact bond for sanitation.

The goal of the impact bond was to reach 85% rural sanitation coverage in target areas by 2023, with 1,600 villages achieving open-defecation-free (ODF) status. Reaching this milestone would accelerate Cambodia's efforts to reach universal sanitation ahead of the 2030 SDG targets (Social Finance, 2021).

After an impressive decade of growth in sanitation coverage in rural Cambodia, remaining households tended to be in the poorest and hardest-to-reach areas. To help the Cambodian government realize its ambitious target by 2023, iDE, a leading rural sanitation provider, needed to access funding to innovate. The impact bond provided funding to innovate. Stone Family Foundation contributed the upfront funding to iDE, which gave iDE the resources to develop and deliver a rural sanitation program to reach the poorest and most vulnerable households. USAID agreed to deliver up to £10 million in outcome funding to Stone Family Foundation if iDE's program enabled these villages to achieve ODF status.

The impact bond was launched in November 2019. USAID last reported that 500 villages had achieved ODF status, with 88,738 households now having confirmed access to sanitation, in line with the Cambodian government's ODF guidelines. USAID has paid $3,125,000 in outcomes to date (USAID, 2019). This is an example of how data can enable innovative financing to drive progress towards the SDGs.

Challenges and importance of verification

The previous example confirms our belief that there is a significant appetite among investors to commit their money to companies that are able to contribute to sustainable development. In other words, such investors are willing to potentially settle for a lower or longer financial return on their investment if they can be reassured that the money will be used to further specific or general SDGs. This should be no surprise, given the interest in ethical investment that has over many years seen investors remove their support for companies whose actions are seen as unethical, such as promoting smoking, using cheap labor in sweatshops, and so on. The key difference between the constraint to investing in unethical businesses and investing in sustainable development is that the former is derived from evidence that the company has performed unethically in a specific way that is relatively easy to verify. On the other hand, proving a company is consistently contributing to sustainable development requires a very different level of evidence.

An example of this difficulty is illustrated by a recent article analyzing green bonds in the Brazilian forestry sector (Ferrando et al., 2021). We quote from the abstract:

> Through the study of recent green bond issuances realized by private companies active in the forestry sector in Brazil, we discuss how green bonds as a "new" form of "green" debt put nature at work and transform the territories and natural elements in the Global South into "temporal and spatial fixes" for the needs of global financial capital.

This is just one example of how demonstrating one's green credentials is difficult to underpin with objective and verifiable data. A recent scientific study has investigated the extent to which carbon offsets are generating the promised effects and found evidence of overestimation: "Results suggest that the accepted methodologies for quantifying carbon credits overstate impacts on avoided deforestation and climate-change mitigation" (West et al., 2020).

At the core of the difficulties with verification is the question of whom we can trust to provide objective and accurate information. Indeed, the whole ESG Initiative (Environment, Social and Governance) has been called into question by Tariq Fancy, BlackRock's first global chief investment officer for sustainable investing: "But there are other issues with ESG investing, including its subjectivity and the unreliability of data and ratings" (Amaro, 2021).

The key issue here is that the people generating the ratings and data are those that will potentially profit from a positive assessment, creating a conflict of interest and consequent erosion of trust at the heart of the initiative.

There are certainly very encouraging reports, such as the work in Costa Rica that received Prince William's environmental Earthshot Prize or a recent submission to the IRCAI Global Top 100 Outstanding Project list[46] based on using computer vision to detect carbon emission in Zambia's forests:

> Our project is based on detecting and reducing carbon emissions in forests using computer vision. We intended to collect data using satellite and also data science, machine learning and artificial intelligence. After collecting the data, we are going to pre-process it, and it will be ready for training and metrics and performance evaluation using Keras software for analysis. The impact of this project is on about 300 people within and near the national parks near the forest that will benefit from this project (Zamculture, 2019).

The surge of support and interest in tackling the UN SDGs is currently at record high levels. While this is an extremely positive development, there is a real danger of disillusionment setting in if companies and countries are found to short-change on the truth, and as we have seen, there is already evidence that this is happening. There is also the danger that without objective and reliable ways of assessing progress, for example, social media could be used to stain a company's image by spreading unfounded rumors that the credentials they claim are not true. Such developments could significantly undermine the interest in and support for SDG investment.

An example of the scale of support for SDG investment is given by the Principles for Responsible Investments (PRI) program of investing with SDG outcomes (UNPRI, 2022) to which investors have committed a combined US$89 trillion in assets under management. Their framework is summarized in their diagram (UNPRI, 2020), which includes the following steps:

46. IRCAI is a center under the auspices of UNESCO. Website at ircai.org

1. Identify outcomes
2. Set policies and targets
3. Investors shape outcomes
4. Financial system shapes collective outcomes
5. Global stakeholders collaborate to achieve outcomes in line with the SDGs

The framework is well-constructed and identifies the aims of the program to direct investment to address the UN SDGs. At the heart of this approach is the need for "investors [to] individually seek to increase positive outcomes, decrease negative outcomes and measure progress toward established targets."

While the question of measurement is highlighted, the broader question of trust is also important to capture. Again, in the words of this report:

> With more objective assessment of SDG Key Performance Indicators (KPI) there is greater opportunity for stakeholders to support initiatives that are making verifiable impact: these could be individual investors, governments, other companies making informed choices about collaboration, etc.

However, we are living in a time of widespread mistrust of institutions and leaders, with most people believing government and business leaders are seeking to mislead them (UNESCO, 2020). Set against this backdrop of the erosion of trust, we believe that this missing piece of the jigsaw is crucial for the role of AI in sustainable development. Therefore, we propose the following manifesto:

There is an urgent need to create a robust system for measuring and certifying the attainment of SDG KPIs, where possible giving evidence for the interventions that were responsible for any changes (positive or negative). The system and its operation need to earn the trust of all stakeholders: citizens, governments, tech companies and industry.

REALIZING THE MANIFESTO

We now turn our attention to the question of how this manifesto can be brought to life. Here we will argue that trust can be created if the conclusions are based on collected and verifiable data and that there is an even-handed presentation of the strengths and weaknesses of the inferences that are drawn from the data.

The role of data

All types of datasets can form the basis for assessing several aspects of the realization of different KPIs of the SDGs. Data has the potential:

- to measure whether an outcome has occurred;
- to record that outcome in a manner that is trusted by all;
- to ensure verifiability and attributability of the outcome to that service or product;
- to use that data to make a payment and to analyze how to improve services, as we shouldn't be satisfied until the SDGs have been fully delivered.

Data are being collected at an unprecedented rate using local and remote sensors. There is also a well-established movement that is arguing for such data collections and science more generally to be made open. For example, UNESCO has established a Recommendation on Open Science:

> The idea behind Open Science is to allow scientific information, data and outputs to be more widely accessible (open access) and more reliably harnessed (open data) with the active engagement of all the stakeholders (open to society) (Masakhane, 2022).

Open Science captures perfectly the potential role and approach that can engender trust in data, but also encourage broader participation in scientific exploration. This is an important part of building trust, namely that all groups should feel that they can participate, in terms of collecting data but also in verifying and contributing to its analysis. By groups here we could be referring to different regions of the world, different sections of society, different scientific disciplines, different governments, NGOs or corporations. The model of developed nations bringing ready-made solutions to bear on remote problems can very easily result in solving the wrong problem or overlooking critical local conditions, resulting in a poor solution or, even worse, no solution at all, with the consequent erosion of trust in both the collaboration and science in general.

An important part of open science and open data is a recognition that local challenges need local participation, in defining the challenge, collecting the data, and collaborating in developing solutions. The Masakhane initiative is an excellent example of an organization trying to do this for African languages with considerable success:

> Masakhane is a grassroots organization whose mission is to strengthen and spur NLP (natural language processing) research in African languages, for Africans, by Africans. Despite the fact that 2,000 of the world's languages are African, African languages are barely represented in technology. The tragic past of colonialism has been devastating for African languages in terms of their support, preservation and integration. This has resulted in technological space that does not understand our names, our cultures, our places, our history (Fairtrade Foundation, n.a.).

The technologies required to certify validity of data are well studied and are being increasingly deployed. In some cases, this can be relatively straightforward, for example for data collected remotely by satellite. The Fairtrade brand has a more challenging problem of tracking its products and producers to ensure that their standards are maintained, but this is an example of a trusted brand that has succeeded in managing this complex task:

> FLOCERT, an independent organization, checks that the Fairtrade standards have been met by the farmers, workers and companies that are part of the product supply chains. In order to reassure consumers that this has happened, we license the use of the FAIRTRADE Mark on products and packaging to signal the standards have been met (VideoLectures.NET, 2020).

Hence, while we do not want to underestimate the challenge, we believe that there is reason for optimism that the Open Science initiative can provide a framework within which the task of collecting and certifying relevant data can be developed and realized. However, collecting and certifying data in itself is not sufficient to attest to the achievement of the KPIs, let alone attribute responsibility. For this, we need to extract insights and knowledge from the data, and it is here that AI can play a vital role.

The role of AI

AI and machine learning are technologies that can be used to extract useful information from data in a verifiable and transparent way: hence they have an increasingly key role to play. As an example, Aidan O'Sullivan has used AI to analyze multispectral satellite imagery to assess water quality in lakes anywhere in the world (Schölkopf, 2019). While this might at first sight only appear to require access

to satellite imagery, there is a vital role of some "ground-truth" data concerning the quality of the water taken from different lakes in order to provide the training data that enables the AI to correctly identify the quality from the multispectral measurements and generalize from a small number of ground-truth measurements. This is an example of the need for local data collection requiring appropriate validation and certification, while there may also be a need for further refinement of the AI methods in order to quantify the accuracy of the predictions in specific cases.

This example again illustrates the variety of contributions that are needed and how a collaboration of the willing can potentially create an ecosystem that will inspire trust through transparency, openness and connectivity. We return to this theme below, but first we should discuss a critical technological component that is required, but which has yet to reach the necessary level of maturity: AI digital twins and mathematical modeling that allow for complex models to track KPIs and provide causal evidence between actions and their outcomes.

The challenge is the need to assign credit or responsibility for changes in the KPIs to the various actors involved. This could be evidence of continued exploitation of a resource such as in deforestation or evidence of interventions that address the issues causing the negative trend, such as for example interventions to improve water quality. The analysis of causality in machine learning is well-established (Schölkopf, 2019) but needs to be scaled to what is often now referred to as digital twins. These are computer models of a particular phenomenon or ecosystem that can be used to test how various interventions have influenced, or could influence, the different KPIs. Hence, through building a complex model of a particular environment we are able to answer "what if" questions and apportion responsibility for the observed and documented changes. As indicated above, a complex model will require advances in AI and mathematical modeling, in particular building on recent advances such as the data-centric engineering program at the Alan Turing Institute (ATI, 2021).

WHAT ARE THE BARRIERS TO REALIZING THE MANIFESTO?

There are a number of issues that may hinder implementation of the manifesto and it is sensible to assess the risks they might pose to its realization. Here we list them briefly.

The first is a lack of common definitions of outcomes and ways to measure them that are trusted by the public, companies and NGOs. The KPIs of the SDGs developed by the United Nations provide a starting point, but this issue will require careful attention, coupled with technical and public engagement, in order to build the necessary level of agreement and trust.

This naturally leads to the second concern that there is a collective action problem around who is, and should be, responsible for developing the definitions of outcomes and the technology solutions that capture and record them. This topic of building solutions that measure and verify outcomes does not represent an obviously attractive focus for funding, because given its nature we are not sure what would be the ideal funding body or the timeline for the return on investment for such a type of initiative. Our manifesto is designed to make the case for this funding by arguing that it makes sense to invest in such an initiative, but leaves open the question of the potential sources of that funding.

A third area of concern is the issue of data ethics and privacy and what is appropriate and ethical to collect and store. This concern needs to be addressed in collaboration with the people and communities affected in order to build trust in how data is being used, following the guidelines of the UNESCO Recommendation on the Ethics of Artificial Intelligence.

Governance is the final issue that we want to highlight: the question of who is responsible for "approving" an outcome definition or the AI for measuring SDGs. It is, of course, critical that the governance be made accountable and transparent in order to engender the necessary trust. This last component builds on the previous ones and is essentially the linchpin for making the manifesto credible and effective.

We need to overcome all these barriers if we are to unlock the potential of data and AI to measure progress against the SDGs, create accountability, and enable investment in companies focused on them.

HOW MIGHT WE OVERCOME THE BARRIERS?

Perhaps the one most important guiding principle for addressing all of the barriers and risk factors is to work collectively: the public, investors, companies, governments, international organizations and NGOs need to come together to define standards around outcome definition, collection, verification, attributability, etc. It is only by ensuring a consensus that the agreed methodology will not become discredited by the criticisms of one or more stakeholders.

The second guiding principle is to start small to build trustworthiness: building trust does not happen overnight. Instead of trying to tackle all 17 SDGs from the beginning, we should rather start with a small number of SDGs in order to demonstrate the potential for data and AI in overcoming trust barriers and building credibility in the approach. This will help test key assumptions around building trust, perception of risks, and whether it results in the unlocking of more investment into tackling the SDGs.

The third guiding principle is to leverage existing tools and applications that scale. There are so many emerging AI solutions that could support this ambition. We should understand what exists and what can be used and scaled without reinventing the wheel.

We have already stressed the need to create transparent governance. We believe this can be achieved by establishing agreed methodologies for determining what outcome definitions, approaches to recording data, and other mechanisms are acceptable to all stakeholders, and how this approval process happens. We need an accounting framework for SDG outcomes that enables organizations to audit what has been achieved. The accounting framework needs to be developed in partnership with the public, investors, companies, governments and NGOs in order to build trust and utility. Furthermore, engagement and partnership efforts need to include the affected people and communities so that the efforts reflect their experiences and expectations. We cannot let companies or NGOs detached from the day-to-day experiences of people determine what is an outcome for them and how it should be measured.

Overall, it will be essential that companies become involved but equally we need to ensure that the methodology is defined by a broader group of stakeholders with the interests of all societies being represented at the international level. It is natural to assume that international organizations such as the United Nations and UNESCO should take a leading role in this process, with UNESCO Category 2 centers such as IRCAI providing technical assistance.

TOWARDS A GLOBAL PARTNERSHIP

The range of expertise and geographies involved make the challenge of measuring SDGs a truly global one that requires the engagement of local teams of researchers in every region that can respond to the call to action. In this sense, we believe that bottom-up funding will be the most effective. This means distributed funding not coming from one funding source or body but diverse sources and scoping, including the size and amounts of contributions, ranging from open calls for technical solutions to micro projects at AI research institutions. The key to success is building trust in the approach and this cannot be imposed, but rather can be achieved only by creating a broad coalition including NGOs, companies, governments and international organizations. It is only by ensuring citizens everywhere feel represented that the support and trust of all societies can be commanded.

For this to be successful, a vital feature will be transparency in terms of what any given technology or solution can deliver. In other words, the description of the pluses and potential minuses, so that criticisms cannot create a narrative of "You are being misled." It is also vital that we create the common language of data to facilitate cross-partisan discussion and agreement on appropriate strategies – in other words, depoliticization of the discussion. It may be easier to achieve all of these desiderata if there is an initial focus on a single or small subset of SDGs, where perhaps the views are less polarized. By building trust in this setting, the opportunity would be created to extend to other more challenging SDGs.

This global partnership could initially be piloted by building a research community in sustainability and AI, via a network that strengthens AI research excellence centers across the world and facilitates collaboration and networking. The objective of this vibrant global network of AI excellence centers in sustainable development would be to boost the research capacity in this domain, and make it attractive for scientists and investors – both social impact and venture capital – and policymakers. This initiative is also expected to contribute to the development of ethical and trustworthy AI, as described in the UNESCO recommendations.

CONCLUSION

> We have argued for the need for the manifesto earlier in this chapter, but it is worth exploring what additional benefits might accrue from its successful realization. One useful analogy is the view that financial markets offer a very sophisticated machinery for ensuring that invested resources deliver the biggest financial return. The sustainable development agenda challenges the belief that this should be the only way in which investments should be measured, and we have argued that there is growing support for this view. However, there is no corresponding mechanism for measuring performance of companies against these new criteria. If we are to literally "put our money where our mouth is," we urgently need to create such mechanisms as our manifesto has urged. Only through the more effective use of data and AI can we avoid the "greenwashing" effect, where companies, via marketing and PR, spin claims to the public and their customers that they are delivering against the SDGs, when in reality they are not. More importantly, this will open up a robust and verifiable route for investors to support sustainable development and for companies to make the case of their products' value for society. It will also allow for companies to showcase their products' added value and potential savings they can bring to governments in terms of quantifiable improvements to SDGs, hence informing social impact bonds.

REFERENCES

Amaro, S. 2021. Blackrock's former sustainable investing chief says ESG is a dangerous placebo. August 24. CNBC. https://www.cnbc.com/2021/08/24/blackrocks-former-sustainable-investing-chief-says-esg-is-a-dangerous-placebo.html

ATI, 2021. Data-centric engineering. The Alan Turing Institute. https://www.turing.ac.uk/research/research-programmes/data-centric-engineering

Edelman. 2020. *21st Annual Edelman Trust Barometer*. https://www.edelman.com/sites/g/files/aatuss191/files/2021-01/2021-edelman-trust-barometer.pdf

Fairtrade Foundation. n.d. What Fairtrade does. https://www.fairtrade.org.uk/what-is-fairtrade/what-fairtrade-does/

Ferrando, T., Miola, I., Junqueira, G. O., Prol, G. M., Vecchione-Goncalves, M. and Herrera, H. 2021. Capitalizing on green debt. *Journal of World-Systems Research*, Vol. 27, No. 2, pp. 410–437.

Schölkopf, B. 2019. Causality for machine learning. *arXiv:1911.10500.* https://arxiv.org/abs/1911.10500

Social Finance. 2021. Cambodia rural sanitation: Bringing safe sanitation to rural communities in Cambodia. https://www.socialfinance.org.uk/projects/cambodia-rural-sanitation

Masakhane. 2022. Masakhane home page. https://www.masakhane.io/

UNESCO. 2020. Open Science home page https://www.unesco.org/en/natural-sciences/open-science

UNPRI. 2020. *Investing with SDG Outcomes: A Five-Part Framework*. https://www.unpri.org/sustainable-development-goals/investing-with-sdg-outcomes-a-five-part-framework/5895.article

UNPRI. 2022. About the PRI. https://www.unpri.org/about-us/about-the-pri

USAID. 2019. *The Cambodia Rural Sanitation DIB: Lessons Learnt from the First Year.* https://www.thesff.com/wp-content/uploads/2022/02/Development-Impact-Bond-lessons-learnt-March-2021-1.pdf

VideoLectures.NET. 2020. AI and Climate: Water Quality Measurement System. http://videolectures.net/IRCAILaunch2021_sullivan_ai_climate/

West, T. A. P., Börner, J., Sills, E. O. and Kontoleon, A. 2020. Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon. *PNAS*, Vol. 117, No. 39, pp. 24188–24194. https://doi.org/10.1073/pnas.2004334117

Zamculture. 2019. Zambia school mapping project. https://github.com/Zamculture/Zambia-School-Mapping-Project-

AI FOR THE SDGS—AND BEYOND? TOWARDS A HUMAN AI CULTURE FOR DEVELOPMENT AND DEMOCRACY

EMMANUEL LETOUZÉ

Marie-Curie Fellow at Universitat Pompeu Fabra and Director and Co-Founder of the Data-Pop Alliance

NURIA OLIVER

Scientific Director and Co-Founder of the ELLIS Unit Alicante Foundation and Chief Data Scientist at the Data-Pop Alliance

BRUNO LEPRI

Senior Researcher at Fondazione Bruno Kessler and Senior Research Affiliate at the Data-Pop Alliance

PATRICK VINCK

Assistant Professor at the Harvard Medical School and Harvard T.H. Chan School of Public Health, Co-Founder and Co-Director of the Data-Pop Alliance

---

SDG1 - No Poverty
SDG2 - Zero Hunger
SDG3 - Good Health and Wellbeing
SDG5 - Gender Equality
SDG8 - Decent Work and Economic Growth
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG12 - Responsible Consumption and Production
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

AI FOR THE SDGS—AND BEYOND? TOWARDS A HUMAN AI CULTURE FOR DEVELOPMENT AND DEMOCRACY

ABSTRACT

Artificial intelligence (AI) can contribute to the United Nations Sustainable Development Goals (SDGs) and 2030 Agenda to end extreme poverty, advance gender equality, protect natural ecosystems, and promote inclusive societies, among others. One channel involves using AI and new digital "crumbs" to estimate SDG indicators to inform better decisions. Yet, in a world where democracy is increasingly tested, including by the influence of AI on inequalities and polarization, using AI to advance human progress and the SDGs calls for more profound changes than providing better fuel to old engines. The primary pitfalls and potential of AI are not technological, they are political and cultural.

Our chapter critically assesses the key tenets and gaps of the "AI for SDGs" narrative and initiatives. It also discusses the contours and conditions of a human AI culture where societies learn and improve using AI as an inspiration and as an instrument controlled by humans. This requires developing awareness, skills and systems for monitoring all SDGs— including the most politically sensitive ones related to press freedom—as well as considering new goals and fostering the participation and collaboration of all data subject-citizens in AI-enabled and AI-inspired initiatives.

To that end, we call on citizens, policymakers, scientists, educators, donors, journalists, civil society members and employees to read and reflect on the perspectives shared in this chapter, hoping they will help shape and leverage AI to promote and protect human development and democracy by 2030 and beyond.

INTRODUCTION

In September 2021, *Wired* magazine published an article entitled "How Valencia crushed COVID with AI" (Marx, 2021). Describing an award-winning initiative led by Nuria Oliver, one of the co-authors of this contribution, the article described an instance where artificial intelligence (AI), using cell-phone metadata combined with epidemiological and online survey data, was used by the government to inform policy decisions with direct effects on public health and economic activity. It exemplified a positive vision where AI, the new epicenter of the data revolution, could help humanity's march towards shared objectives, including the 17 United Nations (UN) Sustainable Development Goals (SDGs) and their underlying agenda, formally adopted by 193 Member States in September 2015.

In its simple version, the line of argumentation underpinning the mainstream "AI for SDGs" discourse is that the explosion in the quantity and diversity of data related to human actions and interactions collected by digital devices and services (i.e. Big Data), and the parallel improvements in algorithmic systems able to learn from these data (e.g., machine learning) may help policymakers, researchers, non-governmental organizations (NGOs), companies and other relevant groups to better measure, and in turn affect, processes and outcomes that are reflected in or relevant to the SDGs. Many initiatives and publications suggest that there is partial truth in this value proposition: AI-powered indicators, insights and initiatives can of course inform decisions and actions that contribute to the SDGs. But it is time to recognize that this argument and most of its surrounding discussions fail to delve into specifics, nuances, caveats and grey zones (Letouzé, 2015b).

For instance, a major problem with such discussions is the assumption that good intentions from decision-makers or global leaders are primarily hindered by insufficient or inadequate information and that simply alleviating that constraint, thanks to AI methods, would have a major impact. The reality is that the main bottlenecks to making data and AI work for the human development and the SDGs are not fundamentally technological. The main bottlenecks are incentives, power dynamics and imbalances that determine the control and use of key resources. For this reason and more, we believe that the "AI for SDGs" vision needs a clearer, bolder theory of change, and a better plan, based on firm conceptual and contextual grounds.

The present contribution focuses on two topics: (1) the neglected discussion about the role that politics, power, and ultimately culture play in the context of "AI for SDGs" efforts; and (2) the paradigmatic changes and ingredients that we think are required in order for AI to fulfill its expectations and defeat the most ominous predictions.

Our key proposition is to create the conditions for a human AI culture where AI will be used as an instrument controlled by humans and as an inspiration for nurturing learning societies.

To do so, we use an analytical framework referred to as "the Four Cs of AI," or 4Cs, that helps describe and discuss the core constituting elements and requirements of AI in a systematic and structured manner. We also propose a taxonomy of contribution channels—including the "measurement channel"—considering current use cases to unpack the theory of change linking AI applications and human development outcomes in an explicit way. We then use the 4Cs as a framework to summarize the main roadblocks and risks that current efforts face. Last, considering the political and economic resistance to change, we sketch the features of a new theory of change and vision that we call a human AI culture, which we argue may support the SDG and democratic agendas in the next decade and beyond, including the most politically sensitive SDG targets and other objectives.

AI AND THE SDGS: CONCEPTUAL AND CONTEXTUAL CLARIFICATIONS

AI is a discipline within computer science or engineering that encompasses a variety of methods and fields (Vinuesa et al., 2020), such as machine learning, computer vision, natural language processing and speech recognition, applied in a wide range of areas with varying levels of societal impacts. While AI as a discipline has existed since the 1950s, several interconnected factors have given it a boost and reboot in the past fifteen years (Lazer et al., 2009). First, the availability of large and rich sets of digital data provides the fuel of data-driven AI methods. Second, we have seen improvements in computing capacities and the development of sophisticated machine learning algorithms, called deep learning, that can learn from large-scale data by leveraging high-performance computing (King, 2013). Third, we have seen the emergence and growth of ecosystems of companies, research groups, public and international organizations and citizen-customers. Finally, the fourth factor that has boosted AI is the advent of a mindset and culture that values efficiency, predictability, and to some extent accountability, cost-effectiveness and measurement, rooted in the adage "you cannot manage what you cannot measure" (Weigend, 2013). A good example of the power of these factors working together is the improved performance of real-time language translation systems. Accordingly, building on past work (King, 2013; Weigend, 2013; Letouzé 2014; Letouzé 2015a), we propose that rather than a mere technological discipline, AI should be conceptualized and discussed as a socio-technological phenomenon made up of four key elements (Figure 1):

1. Crumbs: the pieces of digital data that humans leave behind (Pentland, 2012) as by-products of actions and interactions involving digital devices and services (Letouzé et al., 2013) (see Table 1 in the Annex). These constitute the raw input to data-driven AI methods.

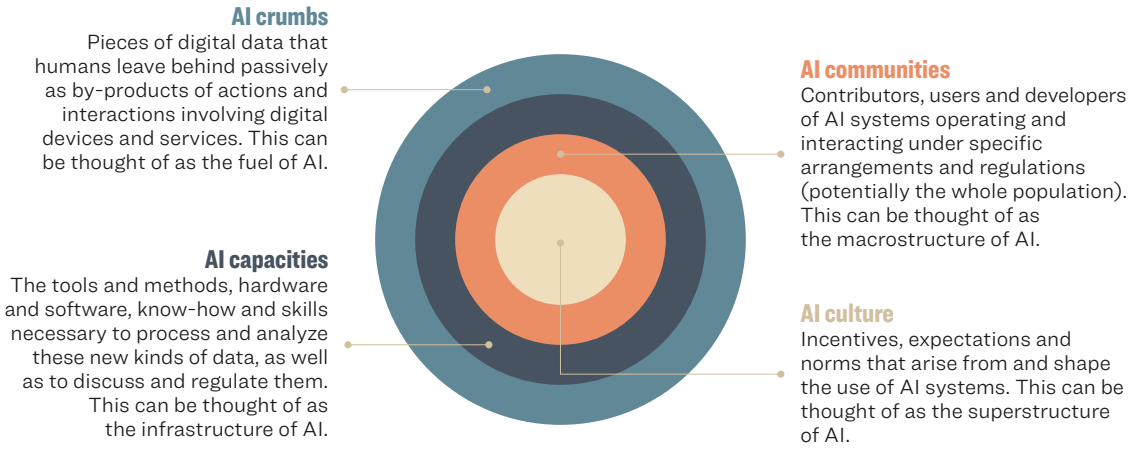
2. Capacities: the tools and methods, hardware and software, know-how and skills necessary to process and analyze these new kinds of data. They can be thought of as AI's infrastructure.

3. Communities: contributors, users and developers of AI systems operating and interacting under specific arrangements and regulations, including UN agencies and other stakeholders of the larger data revolution movement. They may be considered as AI's macrostructure.

4. Culture: the set of incentives, expectations, ideologies, and norms that shape and stem from the use of AI systems, i.e., AI's superstructure, in a Marxist sense.

| FIGURE 1 |

The four Cs of AI as a socio-technological phenomenon, based on Letouzé (2015).



The conceptual framework presented in Figure 1 helps assess and discuss the features and requirements of current and future AI in a structured and holistic manner, as part of a complex ecosystem. It is also useful to describe the genesis and context of the "AI for SDGs" and data revolution narratives and initiatives.

One of the first reports focused on the nexus of AI and SDGs actually predates both. In 2012, UN Global Pulse published a white paper entitled "Big Data for Development: Challenges and Opportunities" (UN Global Pulse, 2012), which laid the foundations of most discussions that have taken place since. In 2013, the High-Level Panel on the Post-2015 Development Agenda called for "a data revolution for sustainable development" (see Figure 2). A year later, an Independent Expert Advisory Group appointed by the UN Secretary General published a report titled "A World that Counts: Mobilizing the data revolution for sustainable development" (IEAG, 2014). The expectation was, and remains, that AI could help fight the dearth of official statistics in developing countries (Letouzé and Jütting, 2015), referred to a "statistical tragedy" (Devarajan, 2013) or "data drought" (*The Economist*, 2014), which would then improve development outcomes, as reflected in the phrases "better data for better decisions and better lives" (Melamed, 2018) and "data are the lifeblood of decision-making and the raw material for accountability" (IEAG, 2014).

| FIGURE 2 |

A New Data Revolution (United Nations, 2013).

> "Too often, development efforts have been hampered by a lack of the most basic data about the social and economic circumstances in which people live... Stronger monitoring and evaluation at all levels, and in all processes of development (from planning to implementation) will help guide decision making, update priorities and ensure accountability. This will require substantial investments in building capacity in advance of 2015. A regularly updated registry of commitments is one idea to ensure accountability and monitor delivery gaps. We must also take advantage of new technologies and access to open data for all people."
>
> *Bali Communiqué of the High-Level Panel, March 28, 2013*

Many groups and efforts have argued they are leveraging AI for the SDGs (Vinuesa et al., 2020; Tomašev et al., 2020).[47] Yet, the fundamental question of how exactly AI is or may be affecting the SDGs—i.e., the underlying theory (or theories) of change at play—has not been sufficiently investigated and articulated. Authors of this contribution have proposed to examine various functions of AI, such as prediction and prescription (Letouzé et al., 2013), while others have proposed to structure analysis by sectors of impact (Vinuesa et al., 2020). In this contribution, the taxonomy built around four contribution channels and modalities is used with the aim of making the possible causal relationships between AI applications and real-world outcomes explicit: measurement and monitoring; precision and smartness; design, monitoring and evaluation; and all other business.

AI for the SDGs: Four contribution channels

The four main contribution channels that we identify are as follows:

1. A measurement and monitoring channel that aims to fill data gaps and improve situational awareness about specific SDG indicators or closely related indicators.
2. A precision and smartness channel via AI-based products and services that are explicitly designed to have an impact on one or more areas covered by the SDGs.
3. A design, monitoring and evaluation channel with the nascent development of AI-powered approaches that seek to design and deploy evidence-based policies and programs.
4. A channel covering all other business, which includes every other AI system not purposely designed with the SDGs in mind; their developers may never have heard of the SDGs, but these systems affect them down the road.

The list is far from exhaustive but aims to give a summary of the state of play in a structured manner.

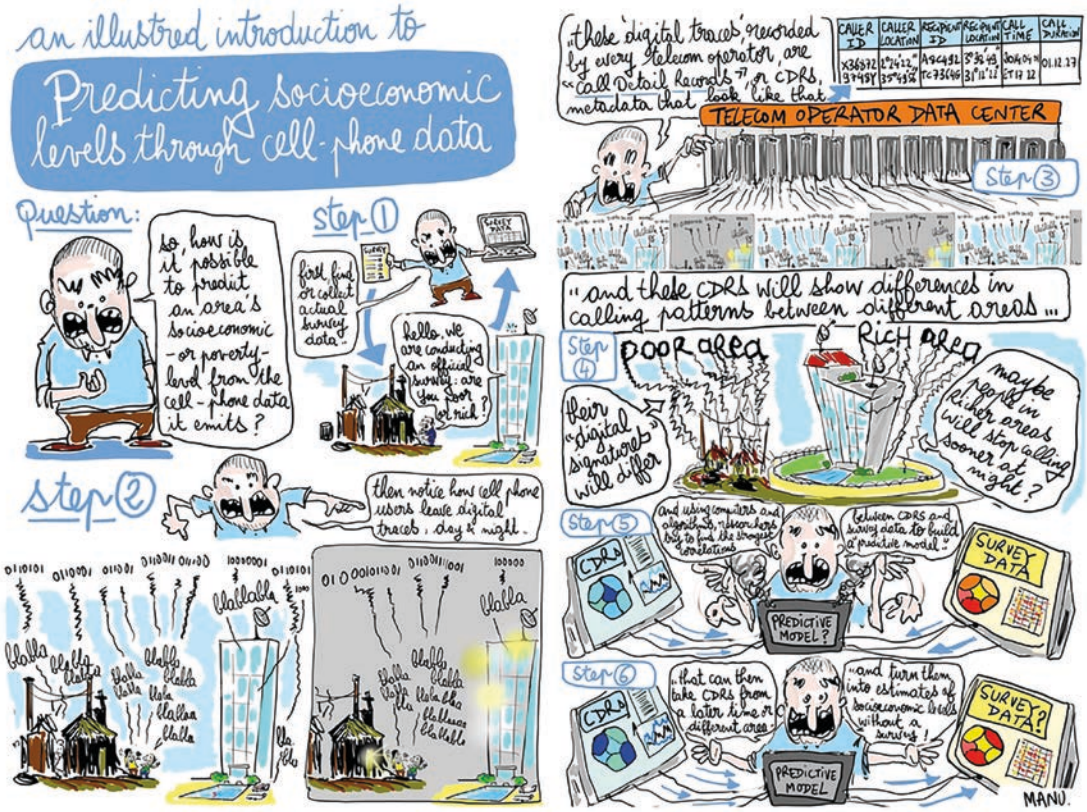
The "measurement and monitoring" contribution channel

As suggested above, it has now long been argued that AI could help promote the SDGs by helping measure and monitor them. Goals and related SDG indicators that have been measured or estimated by AI approaches are typically those that show up in digital crumbs (e.g., electricity consumption tells a lot about socioeconomic status) and are currently monitored through traditional data that provide ground truth. The basic tenets and steps of these approaches are described in Figure 3.

47. Lists of relevant efforts to leverage AI for the SDGs have been compiled in several repositories. For example, the ITU's SDG AI Repository (2021), the database of the AI4SDGs Think Tank (2021) and the database of University of Oxford's Research Initiative AIxSDGs (Saïd Business School, 2021), which lists over 100 projects.

| FIGURE 3 |

Predicting socioeconomic levels through cell phone data (Emmanuel Letouzé, 2013).



Several problems with the "measure and monitoring" channel can be noted. One is the risk of state and corporate surveillance. Another is the scientific validity of some measures. For example, it is conceivable to develop social cohesion monitoring systems based on the frequency of physical and digital contacts derived from records of call details, but whether such interaction constitutes a meaningful and valid measure of social cohesion remains to be determined. Furthermore, such measurements are limited by and often reflect bias and structural inequalities, as discussed further in the next sections. Furthermore, there is a key question of whether and how better measurements of development outcomes such as the SDGs might affect these very outcomes.

The following section provides selected examples of the many studies and pilots that have used AI to estimate indicators falling under the 17 SDGs (Letouzé, 2015a; Oliver, 2021).

Examples of measurement and monitoring efforts by SDG

SDG1 has been covered by numerous efforts, leveraging Earth observation data such as light emissions and rooftop features (Jean et al., 2016), cell-phone metadata (Sundsoy et al., 2016; Soto et al., 2011), digital bank transactions and online ads (Cruz et al., 2019).

SDG2 has been covered by AI techniques that analyze weather data (USAID, 2010), satellite data, demographic data (Quinn et al., 2010) and socio-economic data (Okori and Obua, 2011) to detect hunger and crop yield in developing countries (Zhu et al., 2018; Ghandi and Armstrong, 2016).

SDG3 has been covered by AI methods through the monitoring of social media data to identify epidemics and outbreaks of various diseases as well as vaccine concerns (Letouzé, 2015b). Affordable wearable devices have also enabled the collection of large-scale longitudinal data (Clifton et al., 2014).

SDG4 has been covered by AI through machine learning methods that have aimed to measure students' attendance and performance levels, for example, through the use of socioeconomic and internet-based data to predict dropout rates (Freitas et al., 2020).

SDG5 has been covered by AI using social media data to identify domestic violence hotspots, as well as using other AI methods to identify gender bias and the participation of women in meetings through speech recognition, natural language processing and conversation analysis (Fedor et al., 2009).

SDG6 has been mapped by AI through different measures to detect and track major sources of water contamination (Wu et al., 2021), including drinking water networks (Dogo et al., 2019), as well as to estimate water consumption in rural and urban areas (Brentan et al., 2017).

SDG7 has been covered by AI through techniques that can estimate energy access for electrification and clean cooking fuel through highly frequent Earth observation (EO) (Pokhriyal et al., 2021).

SDG8 has been mapped by AI using satellite data to estimate GDP at national and sub-national levels, as well as through the use of internet-based data to estimate inflation rates (Letouzé, 2015b).

SDG9 has been covered by AI through techniques that can monitor existing infrastructures by analyzing aerial images (Bao et al., 2019; Ren et al., 2020; Xu et al., 2019), as well as detecting the construction of infrastructures, the production of pollutants in industry (Xu et al., 2015), and energy consumption anomalies.

SDG10 has been covered by analyses using airtime credit and mobile phone datasets to evaluate socioeconomic status (Gutierrez et al., 2013), as well as using mobility data and survey data to assess the inequity of access to urban spaces by different socio-economic groups (Letouzé et al., 2022).

SDG11 has been covered by AI techniques focused on urban planning, estimating urban density from aerial images (Lu et al., 2010), and studying transport use through transport cards data and identifying crime hotspots (Bogomoloy, 2014) and illegal drug trafficking (Li et al., 2019).

SDG12 has been covered by AI through the creation of land-use maps to provide an accurate picture of the state and use of natural resources (Talkudar et al., 2020), as well as inferring socially responsible consumption and disposal behavior (Talkudar et al., 2020).

SDG13 has been mapped by AI through satellite data to measure net primary production, make methane observations and monitor population- and energy-related greenhouse gas emissions (Letouzé, 2015b).

SDG14 has been covered by AI through projects that monitor the quality of oceans using deep learning methods, as well as aerial and satellite image analysis and classification that have enabled the estimation of the volume of plastic debris (Martin et al., 2018), estimate the CO_2 flux (Chen et al., 2019) and detect oil spills (Jiao et al., 2019).

SDG15 has been mapped by AI methods through the monitoring of deforestation (de Bem et al., 2020), forest quality (Zhao et al., 2019) and aboveground biomass (Madhab Ghosh and Behera, 2018), as well as the classification of wildlife (Tabak et al., 2018) and detection of illegal wildlife trade (Di Minin et al, 2019).

SDG16 has been covered by AI focused on corruption, through applying AI algorithms to government corruption (Adam and Fazekas, 2018) and financial transactions (West and Bhattacharya, 2016) and on extremism through language processing of social media content (Johansson et al., 2017).

"Precision and smartness" channel and efforts

Efforts in this channel that use AI do not seek to measure any SDG, but to optimize systems and processes that inform decision-making in areas covered by one or more of the 17 SDGs. They are typically described with the qualifier "precision" or "smart," applied to fields such as agriculture, medicine and healthcare, urban development and more. One example is the Famine Action Mechanism (FAM), which supports risk analysis, financing and programming to fight famine (SDG 2) (Badr et al., 2016). AI can also improve child welfare through the early detection of needs (Schwartz et al., 2017), which impacts inequalities (SDG 10). Other initiatives assist in clinical and public health decision-making, including by offering predictions of cancer, (Esteva et al., 2017), tuberculosis (Doshi, 2017), the probability of intensive care (Kaji et al., 2019) and mental health support needs (Walsh et al., 2017).

Other systems relevant to SDGs 9 and 11 aim to optimize garbage collection and recycling as well as predict solid waste patterns (Kannangara, 2018). Efforts to promote responsible consumption and production and climate action (SDGs 12 and 13) focus on the optimization of production systems, such as the estimation of the impact of logging in forests (Hethcoat et al., 2019) and predicting the occurrence and impact of extreme weather events (Lee et al., 2020; Radke et al., 2019; Wong et al., 2020, Pastor-Escuredo et al, 2014), such as the Artificial Intelligence for Disaster Response project that uses social media data (Ong et al., 2020). Still others include Intelligent Tutoring Systems (ITS) and educational interfaces to help design adequate learning tools for students with disabilities (Abdul Hamid, 2018), which is relevant for SDGs 4 and 10. Another example is *Bob Emploi* (Marion, 2018), a project that promised to help better connect job seekers and opportunities (SDG 8). Concerns associated with this channel are often centered around the fairness and governance of automated systems (Lepri et al.,2017).

"Policy design, monitoring and evaluation" channel and efforts

The possibility of using AI to improve policies and programs throughout their life cycles, from design to evaluation, has received much attention in recent years (Bamberger et al., 2016; Letouzé et al., 2019). One argument is that AI and new data sources offer the possibility to capture a target population's behavioral responses and perceptions using social media and other data sources in almost real-time. This feature helps answering the holy-grail question of policymaking: "Has this intervention worked?" or, better, "Is it working now?", thereby allowing a faster course correction. This line of thinking is summarized by a shift from proving to improving in the field of monitoring and evaluation (Letouzé et al., 2019). However, there are still few real-world applications. One example is the use of AI to better target social assistance (Noriega-Campero et al., 2020) by predicting false positives (i.e., people who benefit but should not according to the rules) and false negatives (i.e., people who do not benefit but should). Another is the use of AI to help detect government fraud (West, 2021).

But AI has contrasting effects on the "evaluability challenge." For instance, it is difficult to know the extent to which causality can be assigned between interventions and outcomes (Bamberger et al., 2016) because AI can create many feedback loops and echoes that further complicate causal inference and predictive power, as in the famous example of the "epic failure" of Google Flu Trends (Lazer et al., 2014). AI is poised to affect policymaking in fundamental ways in the future, including by helping identify new concerns and questions of interest. But it should not mean bypassing careful scientific design based on mixed methods, as guidelines developed to that effect have pointed out (Bamberger et al., 2016), and they cannot be a substitute for well-functioning democratic systems.

"All other businesses" channel and efforts

This final channel includes all AI approaches that are used and impact the SDGs daily in positive or negative ways without having been designed with them in mind (or while considering them only very remotely). Although this may be the single most powerful way in which AI affects the SDGs, it is impossible to say whether overall, and for whom, the net impact is positive or negative, both because of the multitude of effects on different people and groups and because these systems are still very new (Vinuesa et al., 2021). For example, Google Maps may reduce pollution and stress by incentivizing people to avoid driving when traffic is bad, but it can lead to fatalities if drivers are fiddling with their phones. Whether the AI-powered services that Amazon provides are overall positive or negative for people and the planet can be argued endlessly either way depending on perspectives and metrics. An important point is that AI effects must be assessed and discussed much more thoroughly, transparently and respectfully based on available data to maximize their positive impacts (Vinuesa et al., 2021), bearing in mind that there is hardly ever a definitive truth.

Key challenges and limitations in data, capacities, communities and culture

The challenges and limitations of current "AI for the SDGs" initiatives have been the subject of a large body of literature (Letouzé and Oliver, 2019). We summarize these challenges and limitations below using the 4Cs of AI as our framework: crumbs (data), capacities, communities and culture.

Crumbs: Locked, biased, messy and sensitive

We may be swimming in data, yet accessing and using these digital crumbs systematically and safely to train AI systems is a major challenge. Most AI crumbs are controlled—legally, practically or both—by private corporations that are often reluctant to share or facilitate access to them and that frequently collect such data with limited consent or control on the part of those whose data are being collected. One reason is commercial considerations: some companies are or may soon be developing their own commercial data-driven services as part of data monetization strategies, so they fear that sharing data may provide insights to competitors. In addition, some of these datasets contain personally identifiable information, which also raises significant reputational and legal risks that companies may not be willing to take. These concerns are especially salient for companies subject to the European General Data Protection Regulation (GDPR), given what we now know about the limits of data anonymization (de Montjoye et al., 2013; 2015) and even differential privacy in practice (de Montjoye et al., 2019). Some social media platforms have developed APIs (application program interfaces) enabling the automated sharing and standardization of data. However, many only allow the querying of archives of past messages. Although satellite data are usually less expensive than ground mapping—for instance, those provided for free by the United States' *National* Aeronautics and *Space Administration* (NASA) and the European Space Agency (ESA)—some remote sensing products are costly, creating a barrier to access.

A next challenge to data is stability and predictability of access to these data, given that many projects and pilots are yet one-offs, which limits the feasibility and desirability of using AI-based measurement and monitoring of human development indicators over the long run. Irrespective of the size and richness of any dataset, and perhaps especially with large complex ones, one must ask what information they really contain and convey. AI crumbs are typically non-representative of the entire population of interest and may reflect and exacerbate existing biases and structural inequalities (Bradley et al., 2021). As discussed in other contributions in this volume, models trained on such data will typically be irrelevant and in some cases unfair or dangerous to segments of the population that were not represented in the training datasets. These biases will tend to be greater with technologies that have lower penetration rates due to a lack of representativeness. This undermines interpretation and actionability as captured by the concepts of internal and external validities as well as the legitimacy of these systems (Flashcard Machines, 2011).

While all statistics shrink the human experience, leaving aside many of its facets, AI crumbs come from much less controlled collection processes than official statistics do. Many are unstructured and user-generated text, so information might be produced by fake profiles or by real people sharing information that may not accurately reflect their own perceptions or acts. A final challenge is the need to combine crumbs with official statistics in many cases for training and ground-truthing. This requires statistics to be easily available and accessible, which often collides with technical and trust levels (Letouzé and Jütting, 2015).

Capacities: HAIves vs hAIves-not

The second set of challenges and limitations to SDGs is the current extent of AI capacities. These encompass human, technological, scientific and financial aspects. A clear key message is that AI capacities are very unevenly distributed across the globe, with implications that are not yet fully grasped and, even less, addressed. Many nations, institutions and communities neither have nor can afford the kinds of equipment and human resources required to create and run the types of AI systems developed and used by top global universities and corporations. Despite progress in the past decade, Global South countries still lag far behind rich countries in all measures of technological capacities, and it is unclear whether the divide is shrinking or widening as a result of the COVID-19 pandemic (UNCTAD, 2021).

Human capacities are another obvious key limiting factor. An example is the lack or shortage of skilled staff in statistical offices in Global South countries, where young computer science graduates are more likely to be working in a local or global technology company than for an underfunded government agency. Popular analytics software such as Python and R may be free, but local staff may not be equipped or incentivized to use them. In general, the diversity of data sources and techniques involved in developing or using AI implies significant training and retraining needs (Dondi et al., 2021; Brown et al., 2019).

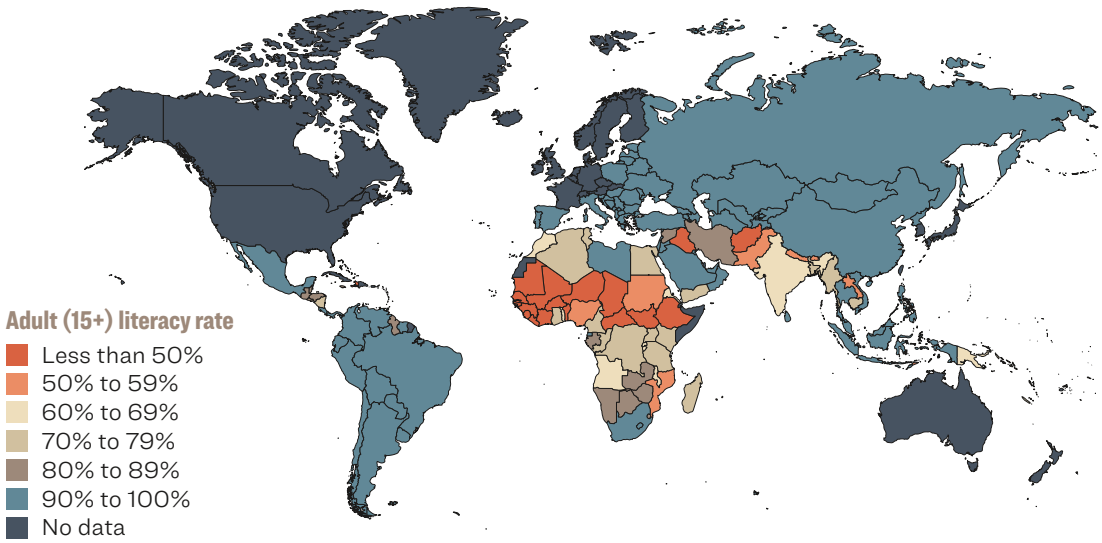
Beyond advanced techno-scientific capacities, key stakeholders generally lack the relevant skills, especially in developing countries—a situation which can be proxied by adult literacy levels (Figure 4). Calls to promote data literacy are welcomed, but these efforts must go beyond simply training students and professionals on how to code (Letouzé et al., 2015). Capacity constraints also include limited standardization of methodologies and technologies to access data in a privacy-conscious manner (de Montjoye et al., 2018), despite the promise of differential privacy[48] (Dwork and Roth, 2014) and attempts such as the Open Algorithm (OPAL) project (Roca and Letouzé, 2016). Techniques to correct for sampling bias using standard statistical techniques and sources are being developed (Zagheni and Weber, 2012; Letouzé et al., 2019), but more needs to be done to ensure that biases are systematically assessed and addressed in the original datasets.

Another capacity issue is the massive energy requirements and carbon footprint of AI-related data storage and processing. According to one study, energy consumption of data centers in Europe may grow 28% between 2018 and 2030 (Montevecchi et al., 2020), while another estimated that training one state-of-the-art Natural Language Processing (NLP) deep-learning model led to an emission of carbon dioxide equivalent to that of the average American in two years (Strubell et al. 2019). On the upside, energy-efficient infrastructures are being developed (Lei and Masanet, 2020), AI may help optimize energy consumption (Gao, 2014), and research is being conducted to better measure the carbon emissions of AI (Lacoste et al., 2019; Henderson et al., 2020; Cowls et al., 2021). However, these trends may still simply be unsustainable.

48. Differential privacy consists of performing a statistical analysis of the datasets that may contain personal data, such that when observing the output of the data analysis, it is impossible to determine whether any specific individual's data was included or not in the original dataset.

| FIGURE 4 |

Adult literacy rates by country (UNESCO, 2017).



Communities: Poor connections and inclusion

As in the case of the Valencian initiative, successful AI efforts require the participation of many stakeholders from the private sector, governments, academia, international organizations and civil society organizations (CSOs), even though their incentives, constraints, and priorities often do not match up well (Letouzé and Oliver, 2019). Some progress has been made in recent years to strengthen connections and trust between stakeholders, including through "data for good" challenges, such as the Data for Refugees Challenge, and other pilots and initiatives, including the European Commission's recent setup of an Expert Group on facilitating the use of new data sources for official statistics, following similar initiatives (Salah et al., 2018; Skibinski, 2020; European Commission, 2022). Collaboration modalities have been proposed to help develop projects within the AI community, such as Data Collaboratives and possible collaboration guidelines and goals (Tomašev et al., 2020). But key obstacles to such initiatives remain, such as the absence of clear business models for data-sharing, as well as regulatory uncertainties, ethical concerns and political calculus (Letouzé et al., 2015; Letouzé and Oliver, 2019).

The woefully inadequate inclusion and participation of marginalized, vulnerable and minority groups—not just in datasets but even (or especially) at the different steps of AI processes and projects—is still a major limitation to applying AI for SDGs. Data and AI systems are neither neutral nor objective; they reflect the questions and preferences of the groups that have the power to put them on the table. Ensuring data protection and individual privacy to mitigate potential harms is of paramount importance, but privacy should also be conceptualized to include group privacy (Kammourieh et al., 2017). Privacy should also include agency, i.e., the capacity of people represented in or affected by AI systems to have a say well beyond simply providing consent when prompted (Letouzé et al., 2015). One attempt at offering a medium for greater local inclusion and representation is the Council for the Orientation of Development and Ethics (CODE) set up by Data-Pop Alliance for all its projects (Letouzé and Yáñez, 2021). But much more needs to be done to promote the appropriate inclusion and participation of data subjects in AI systems.

Culture: When fears, distrust and greed get in the mix

Despite the enthusiasm for AI in some circles, the broad mood in the public space, and to some extent within the "AI for good" community, is one of distrust and fear (Ford, 2015; Ikkatai et al., 2022; Schmelzer, 2019). Mistrust in AI or in AI partners may limit the positive impact of AI on the SDGs and presents a great challenge because it is rooted in legitimate concerns fueled by repeated failure, public scandals and inter-state competition. At the same time, reining in the worst excesses of AI applications may result in overly restrictive legal and regulatory measures that may impede innovation.

Beyond legitimate concerns and grievances, resistance to change is fueled by habits and well-perceived interests. For example, early attempts at leveraging non-traditional data were met with deep skepticism in the official statistical community and government circles, both on scientific grounds and out of fear of losing relevance (Letouzé and Jütting, 2015). At the same time, there are limited incentives for some decision-makers to push for fundamental changes and investments in AI. Even assuming a high-performing AI system, decision-makers may decide to ignore the resulting insights. This decision gap, well known in the humanitarian sector, refers to the disconnect between information and action, which results in part from a lack of a habit of using data for quick decision-making or from a mistrust in such data, and from other political factors, as further discussed in the following section.

The apparent irrelevance of facts could be partly attributed to an overload of data that have "killed facts and truth" (Lepore, 2020). Also, as psychology has shown, it is very difficult for humans to change their minds and actions when such change is at odds with deeply rooted religious, political, economic and other cultural determinants of our identities, or when the behavior stems from an addiction (Kolbert, 2017). For example, over many decades, scientific evidence has proven the detrimental effects of our ways of life on carbon emission and biodiversity, and of alcohol consumption on our own health, but altering hard-wired beliefs and behaviors is very hard.

Trust is a key requirement in order for AI projects to function and for people to slowly come to terms with facts backed by science, which is typically better served by rational and respectful discussions. However, trust is often not strong enough between key stakeholders. An important conclusion drawn from experience and numerous studies is that intangible factors, unrelated to data, technology, skills or regulations, have a significant impact on whether and how AI is used for the purposes of public good (West, 2021).

Towards a human AI culture for human development, learning and democracy in the 21st century

In this section, we aim to propose a longer-term and innovative vision of how AI could contribute to human development objectives, including all the SDGs and beyond, and to democratic principles and processes. We question some of the basic tenets of the standard SDG agenda and discourse in an age of growing distrust and inequality, which are in part fueled by the ubiquity of AI in our lives. In doing this, we sketch the contours and requirements of a vision of a human AI culture.

Restating our problems with the standard "AI for SDGs" narrative

As mentioned above, the argument that AI can help promote human development through the SDGs is weakened by several hard world realities, of which we highlight two.

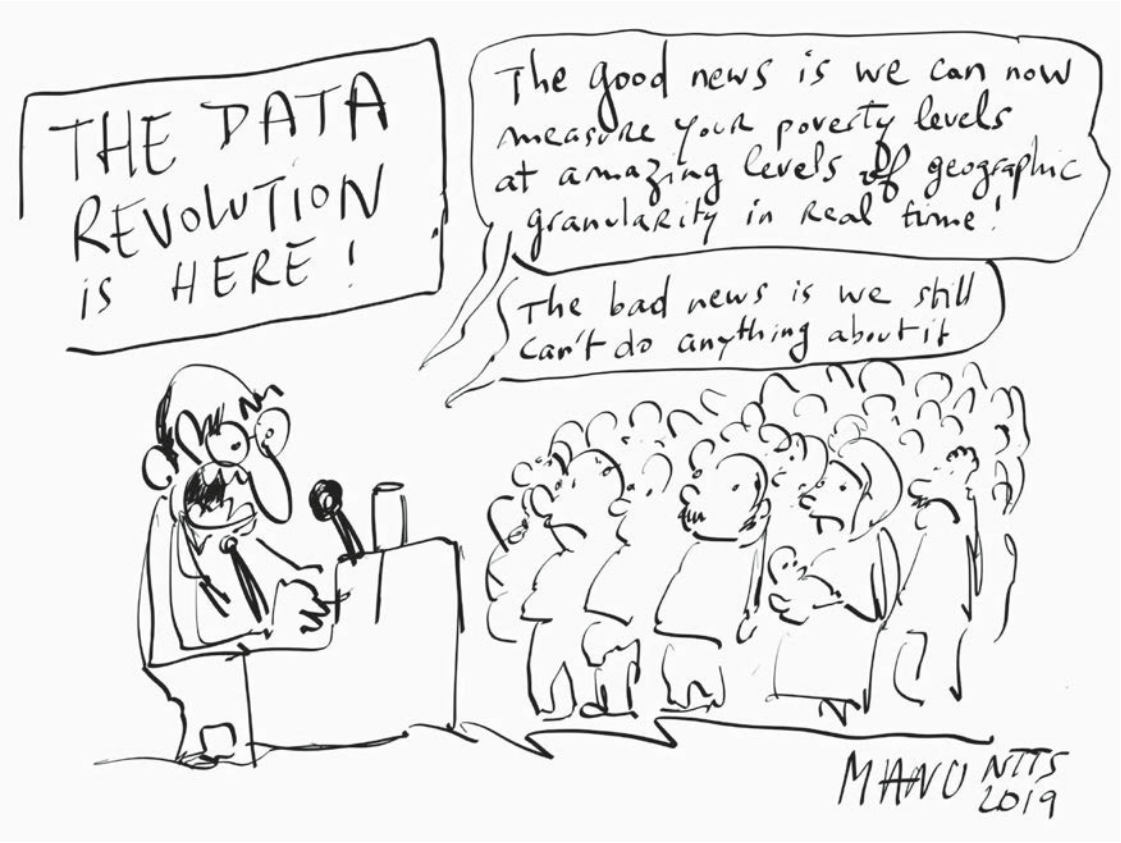
One is the nature and functioning of political regimes around the globe. Indeed, the SDG rationale and the common discourse of the "AI for good" community hinge strongly on the assumption that those making consequential decisions care about the wellbeing of citizens, and that all they lack is high-quality, timely and relevant data to make better decisions. It follows that measurements in this context matter the same way institutions are believed to matter, i.e., they are seen to have a causal effect on outcomes (Przeworski, 2004; Acemoglu and Robinson, 2012; Letouzé, 2018). In contrast, we argue that in the real world, some such leaders have little to no incentive to implement evidence-based policies, especially when the evidence suggests they should implement policies contrary to their political interests or simply leave office. At the same time, they have major incentives to leverage new technologies such as AI for population surveillance and control (Lillis, 2021).

The fact that the SDGs were signed by all 193 heads of governments of UN Members States at the time they were created is both their greatest strength and their greatest flaw. Strength, because, although they are not legally binding, the SDGs help societies hold these signatories accountable regarding commonly set and clearly stated developmental objectives. Flaw, because the nature of many of the signatories' political regimes are such that if any SDG or the whole enterprise had posed a threat to the status quo, they most likely would not have signed them. It has even been argued that the SDGs "undermine[d] democracy" by "pushing an agenda carefully calibrated to avoid upsetting the world's dictators, kleptocrats, and human rights offenders" (Smith and Gladstein, 2019). Although this statement may seem radical, it is not entirely without merit. Democracy appears to be retreating and autocrats have been emboldened by the COVID-19 pandemic. According to the Economist Intelligence Unit (2021), "across the world in 2020, citizens experienced the biggest rollback of individual freedoms ever undertaken by governments during peacetime (and perhaps even in wartime)" and "global democracy continued its precipitous decline in 2021." Income inequality and other forms of inequality continue to widen (Ferreira, 2021; Oxfam, 2022) and, at the time of this writing, the Pandora Papers scandal had just broken (ICIJ, 2021). With all these events combined, it seems naive to argue that the primary obstacle to poverty eradication, gender equality and environmental preservation, among others, is the lack of relevant and timely data or AI algorithms available to political and economic leaders.

The reality is that political and economic interests typically trump scientific evidence and official statistics in determining the priorities and policies that shape real-world outcomes (Figure 5). In this context, the standard "AI or data for good" and "data revolution" narratives may not only be inoperative, but also counterproductive, by providing arguments for development practitioners and politicians to evade accountability. By placing the focus on the dearth of data and the marvels that better AI-powered insights could enable, it is easy for them, especially those who are corrupt, incompetent or both, to claim they failed to improve X because they didn't have the right data on X. To be clear, in our view, poor countries and communities are not poor because their leaders lack good poverty data about them; they are poor and their poverty is not adequately captured because they do not count. When an engine is broken, improving its fuel won't do the trick. The question is, how can it be repaired?

| FIGURE 5 |

The Data Revolution is here! (will it improve all lives?), taken from Emmanuel Letouzé, illustration at the Eurostat NTTS event, March 13, 2019.



In this endeavor, AI can certainly help, though it presents certain challenges. In addition to the barriers to truly advancing AI for the SDGs posed by governments' conflicting political and economic interests, the second major issue is the role of AI-powered platforms in breaking down trust in experts, institutions, neighbors, and, ultimately, facts. A growing body of research suggests that social media platforms and technology giants that are effectively data companies with near complete market dominance are contributing to political polarization, and some fear that they may threaten the very survival of democratic practices and systems (Helbing et al., 2017; Bergstrom and West, 2020; Risse, 2021). This would also mean that objective benefits from AI such as the ability to detect cancer or fraud may be considered suspicious. The result is that AI can hardly be expected to seamlessly help "build back better" after the COVID-19 pandemic, amid multiple compounding ecological and socio-political crises under current conditions, without a fundamental change in how and by whom AI systems are developed, used and regulated—for whom and with what goals.

New legal and regulatory frameworks are emerging around the world to guide the use of data and AI. These developments, however, are largely region- or country-specific and fall short of effectively creating new global rights. Some examples include the right to be forgotten and the European General Data Protection Regulation (GDPR), which are not global norms and in effect result in unequal digital treatment of people. As our physical and digital lives become intertwined, there may be a more fundamental need to rethink our human rights and an equally fundamental need to formalize the rights and responsibilities of AI systems. The Asilomar AI principles[49] are an important first step in that direction (Future of Life Institute, 2017). However, they are limited to AI research and development and are not internationally agreed-upon rules and global norms subjected to enforcement and accountability, which are urgently needed to reduce the risk of a dystopian AI future, including the potential for AI warfare.

A question that is getting more attention is whether AI regulations should focus on ex-ante requirements or ex-post accountability. While the focus is currently on the former, the latter may be more realistic given the distributed nature of AI systems.

Features, requirements and expected benefits of a human AI vision and culture

Despite these worrying trends and growing concerns, we believe that AI can help promote human development and democratic goals. Fundamentally, AI systems are not just powerful tools that can help achieve specific tasks; they also show how data nodes and feedback together enable systems to learn to get better at reaching a set of shared objectives. Somewhat ironically, while AI was inspired by the human brain, we argue that AI could and should now serve as an inspirational analogy for better human systems and societies based on learning, provided the right ingredients are available, nurtured and used.

Following previous contributions, this idea of considering and using AI as both an instrument (narrow AI systems that excel at specific tasks) and an inspiration for human societies based on a renewed desire and ability for collective learning is referred to as "human AI culture" (Pentland, 2017; Letouzé and Pentland, 2018). The human AI culture fosters a vision of how the various parts (nodes) that make up human societies collaborate to learn and reinforce our progress towards shared goals, for which AI could be used as a tool. Such culture would, for example, question whether the goal of building a safer, more peaceful society is best served by the "war on drugs" and related mass incarceration policies that have been taking place in parts of North America over the past decades, or by other means (Pearl, 2018). In doing so, it may leverage AI to help suggest and test alternative approaches, but it may also prefer low-tech solutions.

A human AI culture would also consist of a vision under which the desirability and legitimacy of certain objectives—such as boosting GDP or maximizing profits—would be reassessed in a systematic and continuous fashion based on their effects, as in a learning system. The key requirements and ingredients of such a culture are relatively well known. For instance, it requires nurturing a culture of reasoned and rational discussion, cooperation and, therefore, trust between the nodes far beyond what is observable today between groups, such that measurement has a chance to matter the way it does in AI. In addition, it requires having accurate and timely input data and feedback information from which the system can constantly learn. Furthermore, it requires broad data literacy in societies (Letouzé and Bhargava, 2015), greater control from data subjects over data about themselves—for instance, through the development of data cooperatives or other data-sharing and access mechanisms (Pentland and Hardjono, 2020)—and free press (UNESCO, 2022).

49. See https://futureoflife.org/ai-principles/

The way towards a human AI culture would entail reviving or reinventing democratic principles of participation, self-governance and government by means of discussions based on rational compassion (Bloom, 2016), including and increasingly at local levels. It also requires developing incentives, means and habits for all stakeholders to demand that collective decisions be evaluated systematically. This evaluation should be conducted using the best available data and methodologies in order to adjust future iterations and contribute to a body of evidence on what actions yield which results. In this sense, to avoid deepening the inequalities that the digital economy seems prone to producing, such incentives, means and habits should involve a reconsideration of how different forms of capital—including digital capital—are shared (Gardels, 2022).

It will not be easy to build a human AI culture that places rational respectful discussions based on trust and facts at the core of a new social contract among humans and between humans and machines in 21st-century societies. This is true mostly because it implies addressing the excesses and abuses of powerful actors that are at the root of most humanity's ills and considering dissident voices and the complexities of human realities. As suggested above, it is not just about using AI to provide a better fuel to old machineries; it means and requires upgrading these systems, using AI as an instrument when and as needed, but also as an inspiration.

New indicators and the next SDGs agenda?

One concrete way to start drafting new indicators and the next SDGs agenda is to promote AI efforts that seek to monitor all SDGs' targets, notably the politically sensitive Tier 3 indicators under SDG 16, which seeks to "promote just, peaceful and inclusive societies." These include SDG indicator 16.6.2, "proportion of population satisfied with their last experience of public services' analyzing social media data" (Data-Pop Alliance, 2018) and indicator 16.10.1, "number of verified cases of killing, kidnapping, enforced disappearance, arbitrary detention and torture of journalists, associated media personnel, trade unionists and human rights advocates in the previous 12 months" (Muñoz et al., 2021). These efforts could garner support from international research and advocacy organizations as well as like-minded companies willing to put pressure on governments that are most reluctant to discuss and address these phenomena.

New goals that reflect new societal realizations and priorities should also be considered. Some groups are already suggesting new priorities, such as animal health, welfare and rights (Visseren-Hamakers, 2020), sustainable space (ITU News, 2021) or space for all (National Space Society, 2020), meaningful and safe digital life (Jespersen, n.d), ensuring the Digital Age supports people, the planet, prosperity and peace (Luers, 2020), development and disability (Le Marrec, 2016).

AI may also assist in identifying the SDGs that should be prioritized based on expressed public interests and feasibility studies. Such efforts should take place under human supervision through a carefully participatory design to ensure that they do not reflect structural biases present in datasets. The way to mitigate structural biases could follow a similar line to what has been argued for identifying research priorities in AI (Vinuesa et al., 2020) or for reflecting ethical values in AI systems (Rahwan, 2017).

CONCLUSION

The Data and AI Revolution need to be politicized. The COVID-19 pandemic has exposed and exacerbated pre-existing structural fault lines in our society. Our world is increasingly digital and unequal; while digitalization is steadily increasing, democracy and equality seem to be retreating. In this context, the rise of AI seems to be a perfect case of a Promethean fire. It can certainly help better measure and promote the SDGs and other human development objectives, despite challenges and obstacles in the way, which can be addressed with appropriate investments in data, capacities, collaborations and initiatives. But AI can also further fuel inequities, polarization and the breakdown of trust.

Fundamentally, we argue that the problems to address are not primarily technological. They are primarily political and cultural, rooted in personal greed, elite capture, power hunger and societal distrust. It follows that their solutions must be primarily political and cultural.

Thus, unless there is a recognition that the current standard "AI for SDGs" discourse—according to which the primary constraint is lack of indicators on the dashboards of global leaders—errs on the side of complacency or naivety, AI will not deliver on its promise. In a business-as-usual scenario, where AI remains controlled by individuals and groups driven by power and profit motives, AI is more likely to yield and fuel a future of technological control of citizens, with reduced choices and freedoms and lowered living standards for those on the losing side of rising economic, social, political and environmental inequalities.

But we are not giving up on AI. Paradoxically, while AI mimics the human brain, human societies could now try and take inspiration from AI systems by valuing and nurturing learning capacities and cooperation. We call this a human AI culture, and we describe this culture as using AI as both an inspirational analogy and a set of instruments to measure, monitor and reach commonly set objectives. The most critical objective is to uphold and protect democratic principles and processes. In particular, by giving all people much greater control and transparency over the design and use of AI systems that impact their lives. This must be coupled with clear and firm accountability and compliance mechanisms regarding the design and use of such systems. Perhaps the case of Valencia, Spain, mentioned in the beginning of this chapter, shows that a human AI culture can be achieved.

REFERENCES

Acemoglu, D. and Robinson, J. 2012. *Why Nations Fail: The Origins of Power, Prosperity and Poverty.* New York: Crown Business.

Avendano, R., Jütting, J. and Kuhm, M. 2020. *The Palgrave Handbook of Development Cooperation for Achieving the 2030 Agenda*. London: Palgrave Macmillan.

Barret, P., Hendrix, J. and Sims, G. 2021. Fueling the fire: How social media intensifies U.S. political polarization – and what can be done about it. *NYU Stern Center for Business and Human Rights*, September 21.

Big Data UN Global Working Group. 2019. Training, Skills and Capacity-Building — UN GWG for Big Data.

Bloom, P. 2016. *Against Empathy: The Case for Rational Compassion*. London: The Bodley Head.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. and Flaxman, S. 2021. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, No. 600, pp. 695–700.

Cowls, J., Tsamados, A., Taddeo, M. and Floridi, L. 2021. The AI Gambit — Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations. https://ssrn.com/abstract=3804983

Devarajan, S. 2013. Africa's statistical tragedy. *Review of Income and Wealth*, No. 59, pp. 9–15. https://doi.org/10.1111/roiw.12013

Dickinson, E. 2011. GDP: A Brief History. *Foreign Policy*. January 3. https://foreignpolicy.com/2011/01/03/gdp-a-brief-history/

Economist. 2015. How to catch the overfishermen. January 22. https://www.economist.com/leaders/2015/01/22/how-to-catch-the-overfishermen

Economist. 2022. Daily chart: A new low for global democracy. https://www.economist.com/graphic-detail/2022/02/09/a-new-low-for-global-democracy

European Commission. 2022. High-Level Expert Group on Artificial Intelligence: shaping Europe's digital future. June 13. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Ferreira, F. 2021. Inequality in the times of COVID-19. IMF Finance and Development. https://www.imf.org/external/pubs/ft/fandd/2021/06/inequality-and-covid-19-ferreira.htm

Flashcard Machine. 2011. Internal v. External Validity. June 13. https://www.flashcardmachine.com/internal-vsexternal-validity.html

Ford, P. 2015. Our fear of artificial intelligence. MIT Technology Review. February 11. https://www.technologyreview.com/2015/02/11/169210/our-fear-of-artificial-intelligence/

Grameen Foundation. 2011. Lessons learned from AppLab's first three years in Uganda. January 21. https://grameenfoundation.wordpress.com/2011/01/21/lessons-learned-from-applab%E2%80%99s-first-three-years-in-uganda/

Griswold, W. 2008. *Cultures and societies in a changing world*. 3rd edition. Thousand Oaks, CA: Pine Forge.

Gutierrez, T., Krings, G. and Blondel, V. 2013. *Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets.* https://doi.org/10.48550/arXiv.1309.4496

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, No. 21, pp. 1–43.

ICIJ. 2021. Pandora Papers. International Consortium of Investigative Journalists. https://www.icij.org/investigations/pandora-papers/

Ikkatai, Y., Hartwig, T., Takanashi, N. and Yokoyama, H. M. 2022. Octagon measurement: public attitudes toward AI ethics. *International Journal of Human–Computer Interaction,* pp. 1–18. January 10. https://doi.org/10.1080/10447318.2021.2009669

Independent Expert Advisory Group (IEAG). 2014. *A World that Counts: Mobilising the Data Revolution for Sustainable Development.* https://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf

Jespersen, S. n.d. Advocating for an 18th Sustainable Development Goal: A meaningful and safe digital life. VERTIC. https://www.vertic.com/our-thinking/advocating-for-an-18th-sustainable-development-goal-a-meaningful-and-safe-digital-life

King, G. 2013. Big Data is not about the data! Talk presented at the Golden Seeds Innovation Summit, New York City. Institute for Quantitative Social Science, January 30. http://gking.harvard.edu/files/gking/files/evbase-gs.pdf

Kolbert, E. 2017. Why facts don't change our minds. *New Yorker Magazine*. February 27. https://www.newyorker.com/magazine/2017/02/27/why-facts-dont-change-our-minds

Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T. 2019. Quantifying the carbon emissions of machine learning. *arXiv:1910.09700* [cs.CY]

Le Marrec, J. 2016. Where is Sustainable Development Goal 18? Development and Disability blog. http://developmentanddisability.blogspot.com/2016/04/where-is-sustainable-development-goal-18.html

Lepore, J. 2020. The End of Knowledge: How Data Killed Facts. Lecture given at the Fox Center for Humanistic Inquiry, Emory University, April 8.

Letouzé, E. 2013. Could Big Data provide alternative measures of poverty and welfare? Development Progress blog, June 11. https://www.oecdbetterlifeindex.org/fr/blogue/could-big-data-provide-alternative-measures-of-poverty-and-welfare.htm

Letouzé, E. 2014. Big Data for development: Facts and figures. SciDev.net, April 15. http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html

Letouzé, E. 2015a. *Big Data and Development: An Overview.* Data-Pop Alliance. http://datapopalliance.org/item/white-paper-series-official-statistics-big-data-and-human-development/

Letouzé, E. 2015b. Thoughts on Big Data and the SDGs. Data-Pop Alliance. February 18. https://datapopalliance.org/wp-content/uploads/2020/09/7798BigData-Data-Pop-Alliance-Emmanuel-Letouze.pdf

Letouzé, E., del Villar Z., Molina, R. L., Nieto B. F., Romero, G., Ricard, J., Vazquez, D. and Maya, L. A. C. 2022. Parallel Worlds: Revealing the Inequity of Access to Urban Spaces in Mexico City Through Mobility Data. In: *Measuring the City: The Power of Urban Metrics*. Edited by Ahn, C., Ignaccolo, C. and Salazar-Miranda, A. https://projections.pubpub.org/pub/01kebgos/release/1

Letouzé, E., Meier, P. and Vinck, P. 2013. Big Data for conflict prevention: New oil and old fires. In *New Technology and the Prevention of Violence and Conflict*, pp. 4–27. International Peace Institute. https://www.ipinst.org/images/pdfs/IPI_Epub-New_Technology-final.pdf

Letouzé, E. Noonan, A., Bhargava, R., Deahl, E., Sangokoya, D. and Shoup, N. 2015. Beyond data literacy: reinventing community engagement and empowerment in the age of data. MIT Media Lab, September. https://www.media.mit.edu/publications/beyond-data-literacy-reinventing-community-engagement-and-empowerment-in-the-age-of-data/

Letouzé, E. and Pentland, A. 2018. Human AI for human development. *ITU Journal: ICT Discoveries* Special Issue, No. 2 (December). https://www.itu.int/en/journal/002/Pages/15.aspx

Letouzé, E., Pestre, G. and Zagheni, E. 2019. The ABCDE of Big Data: Assessing biases in call-detail records for development estimates. *The World Bank Economic Review*, December 3. https://doi.org/10.1093/wber/lhz039

Letouzé, E. and Oliver, N. 2019. Paper sharing is caring: Four key requirements for sustainable private data sharing and use for public good. Data-Pop Alliance; Vodafone Institute for Society and Communications, November. https://datapopalliance.org/paper-sharing-is-caring-four-key-requirements-for-sustainable-private-data-sharing-and-use-for-public-good/

Letouzé, E., Vinck, P. and Kammourieh, L. 2015. The law, politics and ethics of cell phone data analytics. Data-Pop Alliance Big Data and Development Primer Series, April. http://datapopalliance.org/item/white-paper-the-law-politics-and-ethics-of-cell-phone-data-analytics/

Letouzé, E. and Yáñez Soria, I. 2021. The CODE for building participatory and ethical data projects. Data-Pop Alliance. https://datapopalliance.org/the-code-for-building-participatory-and-ethical-data-projects/

Lillis, K. B. 2022. NSA watchdog finds "concerns" with searches of Americans' communications. *CNN*. February 1. https://edition.cnn.com/2022/01/31/politics/nsa-watchdog-concerns-searches-american-communications/index.html

Luers, A. 2020. The missing SDG: Ensure the Digital Age supports people, planet, prosperity & peace. Inter Press Service News Agency, July 6. http://www.ipsnews.net/2020/07/missing-sdg-ensure-digital-age-supports-people-planet-prosperity-peace/

Marx, W. 2021. How Valencia crushed COVID with AI. *Wired*. September 9. https://www.wired.co.uk/article/valencia-ai-covid-data

Montjoye, Y.-A., L. Radaelli, V. K. Singh, and A. S. Pentland. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, no. 6221, pp. 536-39. https://doi.org/10.1126/science.1256297

Montjoye, Y-A, Hidalgo, C. A., Verleysen, M. and Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, Vol. 3, No. 1, pp. 1–5. https://doi.org/10.1038/srep01376

Oliver, N. 2021. *Artificial Intelligence for Social Good: The Way Forward.* European Commission. Forthcoming.

Pearl, B. 2018. Ending the War on Drugs: By the numbers. Fact sheet. Center for American Progress. June 27. https://www.americanprogress.org/issues/criminal-justice/reports/2018/06/27/452819/ending-war-drugs-numbers/

Pentland, A. 2014. *Social Physics: How Good Ideas Spread—The Lessons from a New Science*. New York: Penguin Press.

Pentland, A. 2012. Reinventing society in the wake of Big Data: A conversation with Alex (Sandy) Pentland. *Edge*. August 30. https://www.edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data

Przeworski, A. 2004. Institutions matter? *Government and Opposition*, Vol. 39, No. 4, pp. 527–40. https://doi.org/DOI:10.1111/j.1477-7053.2004.00134.x

Rasmussen, K. and McArthur, J. 2017. How successful were the Millennium Development Goals? *Brookings* blog, January 11. https://www.brookings.edu/blog/future-development/2017/01/11/how-successful-were-the-millennium-development-goals/

Roca, T. and Letouzé, E. 2016. Open algorithms: A new paradigm for using private data for social good. Data-Pop Alliance. https://datapopalliance.org/open-algorithms-a-new-paradigm-for-using-private-data-for-social-good/

Salah, A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y-A., Dong, X. and Dagdelen, O. 2018. Data for refugees: the D4R challenge on mobility of Syrian refugees in Turkey." *arXiv*. October 14. http://arxiv.org/abs/1807.00523.

Schmelzer, R. 2019. Should we be afraid of AI? Forbes. October 31. https://www.forbes.com/sites/cognitiveworld/2019/10/31/should-we-be-afraid-of-ai/.

Skibinski, A. 2020. Expert Group on Facilitating the Use of New Data Sources for Official Statistics. *CROS – European Commission*. December 4. https://ec.europa.eu/eurostat/cros/content/expert-group-facilitating-use-new-data-sources-official-statistics_en.

Smith, J. and Gladstein, A. 2018. How the UN's Sustainable Development Goals undermine democracy. *Quartz Africa*, June 7. https://qz.com/africa/1299149/how-the-uns-sustainable-development-goals-undermine-democracy/

Tomašev, N., Cornebise, J., Hutter, F. et al. 2020. AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, Vol. 11, article no. 2468. https://doi.org/10.1038/s41467-020-15871-z

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M. and Nerini, F. F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, Vol. 11, article no. 233. https://doi.org/10.1038/s41467-019-14108-y

UNCTAD. 2021. *Technology and Innovation Report 2021: Catching Technological Waves: Innovation with Equity.* https://unctad.org/system/files/official-document/tir2020_en.pdf

UN Global Pulse. 2012. *Big Data for Development: Challenges and Opportunities.* https://www.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay2012.pdf

UN Global Pulse. 2015. Big Data and Development: An Overview. https://www.unglobalpulse.org/document/big-data-for-development-in-action-un-global-pulse-project-series/

UNESCO. 2017. *Literacy Continues to Rise from One Generation to the Next*. UNESCO Fact sheet No. 45. September. https://uis.unesco.org/sites/default/files/documents/fs45-literacy-rates-continue-rise-generation-to-next-en-2017_0.pdf

United Nations. 2013. *A New Global Partnership: Eradicate Poverty and Transform Economies Through Sustainable Development: The Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda.* https://sustainabledevelopment.un.org/content/documents/8932013-05%20-%20HLP%20Report%20-%20A%20New%20Global%20Partnership.pdf

United Nations Development Programme. 2019. Human Development Index. http://hdr.undp.org/en/content/human-development-index-hdi

Visseren-Hamakers, I. J. 2020. The 18th Sustainable Development Goal. *Earth System Governance*, Vol. 3. https://doi.org/10.1016/j.esg.2020.100047. https://www.sciencedirect.com/science/article/pii/S2589811620300069

West, D. M. 2021. Using AI and machine learning to reduce government fraud. Brookings. September 10. https://www.brookings.edu/research/using-ai-and-machine-learning-to-reduce-government-fraud/

Zagheni, E. and Weber I. 2012. You are where you e-mail: Using e-mail data to estimate international migration rates. In *WebSci '12: Proceedings of the 4th Annual ACM Web Science Conference*, pp. 348–351. ACM Digital Library. https://doi.org/10.1145/2380718.2380764

Zheng, X. and Lee, M. K. O. 2016. Excessive use of mobile social networking sites: Negative consequences on individuals. *Computers in Human Behavior*, Vol. 65, pp. 65–76. https://doi.org/10.1016/j.chb.2016.08.011

| ANNEX | Taxonomy and examples of Big Data sources

Types	Examples	Opportunities
CATEGORY 1: EXHAUST DATA		
Mobile-based	Call details records (CDRs) GPS (fleet tracking, bus AVL)	Estimate population distribution and socioeconomic status in places as diverse as the UK and Rwanda.
Financial transactions	Electronic ID E-licenses (e.g., insurance) Transportation cards (including airplane fidelity cards) Credit and debit cards	Provide critical information on population movements and behavioral response after a disaster.
Transportation	GPS (fleet tracking, bus AVL) EZ passes	Provide early assessment of damage caused by hurricanes and earthquakes.
Online traces	Cookies IP addresses	Mitigate impacts of infectious diseases through more timely monitoring using access logs from the online encyclopedia Wikipedia.
CATEGORY 2: DIGITAL CONTENT		
Social media	Tweets (Twitter API) Check-ins (Foursquare) Facebook content YouTube videos	Provide early warning on threats ranging from disease outbreaks to food insecurity.
Crowd-sourced and online content	Mapping (Open Street Map, Google Maps, Yelp) Monitoring and reporting (uReport)	Empower volunteers to add ground-level data that are useful notably for verification purposes.
CATEGORY 3: SENSING DATA		
Physical	Smart meters Speed and weight trackers USGS seismometers	Sensors have been used to assess the demand for using sensors to estimate demand for high-efficiency cookstoves at different price points in Uganda or willingness to pay for chlorine dispensers in Kenya.
Remote	Satellite imagery (NASA TRMM, LandSat) Unmanned aerial vehicles (UAVs)	Satellite images revealing changes in, for example, soil quality or water availability have been used to inform agricultural interventions in developing countries.

An overview of initiatives addressing SDGs

SDG/impact field	Project or initiative	Organization	Data sources and tools	What is monitored or studied?	Description	Country or region	Implications of using data-driven approaches	Years	Tiers	Type of organization
Goal 16: Peace, Justice and Strong Institutions	FollowTheMoney.org	National Institute on Money in Politics	Campaign finance reports	Campaign financing	Compilation and categorization of **campaign finance reports** made open to the public	USA	Promote transparency in campaign financing, as well as promote open access to large body of cross-jurisdictional reports	2010–present	Tier III	Government
Goal 12: Responsible Consumption and Production, **Goal 8:** Decent Work and Economic Growth	Scanner data in the Swiss CPI: An alternative to price collection in the field	Swiss Federal Statistical Office (FSO)	Price scanner data	Consumer price index	Use **price scanner data** to calculate consumer price index for food and near-food groups	Switzerland	Improve the price collection of the consumer price index: improved quality, reduced costs and reduced administrative burden	2018–present	Not classi-fied	Government
Goal 11: Sustainable Cities and Communities, **Goal 12:** Responsible Consumption and Production	Using satellite imagery and geo-spatial data for the census of agriculture and the census of building and housing	Mongolia NSO	Satellite imagery, geospatial data	Crop production	Use of **satellite imagery and geospatial data** to identify crop types and estimate production to create a first agricultural by-census	Mongolia	Supplement existing data with satellite images	2017	Not classi-fied	Government
Goal 3: Good Health and Wellbeing	Assessment of the Potential for International Dissemination of Ebola Virus through Commercial Air Travel During the 2014 West African Outbreak	Flowminder	International Air Transport Association data, historic traveler flight itinerary	Ebola epidemic	Model the expected number of internationally exported Ebola virus infections, the potential effect of air travel restrictions, and the efficiency of airport-based traveler screening at international ports of entry and exit using **international air transportation data and historic traveler flight itineraries**	Guinea, Liberia, and Sierra Leone	Inform decision-makers on the potential harms of travel restrictions and most efficient screening sites	2014	Not classi-fied	Academic
Goal 2: Zero Hunger, **Goal 3:** Good Health and Well-Being, **Goal 5:** Gender Equality	Big Data and the Cloud – Piloting "eHealth" for Community Reporting of Community Performance-Based Financing in Ghana	World Bank Group	Mobile-based surveys	Effectiveness of Maternal Child Health Nutrition Improvement Project	Report performance of community-level health teams by using **Android-based software survey tools**	Ghana	Circumvent the time delay, capacity constraints and data quality challenges associated with paper-based reporting	NA	Tier III	International, government

SDG/impact field	Project or initiative	Organization	Data sources and tools	What is monitored or studied?	Description	Country or region	Implications of using data-driven approaches	Years	Tiers	Type of organization
Goal 1: No Poverty	Forecasting Poverty and Shared Prosperity Using Cell Phone Data	World Bank Group	Call-detail records (CDR)	Estimate and forecast poverty and shared prosperity	Measure population "digital footprints" by analyzing **cell phone records** using data mining and computer-learning techniques to estimate and forecast poverty and shared prosperity	Guatemala	Provide an affordable, practical and scalable solution for mapping poverty	2019	Tier III	International, government, private organization
Goal 1: No Poverty, **Goal 2:** Zero Hunger, **Goal 11:** Sustainable Cities and Communities, **Goal 13:** Climate Action	Predicting vulnerability to flooding and enhancing resilience using big data	World Bank Group	Google cloud data (elevation, satellite imagery, census data)	Flooding risk	Use of **Google cloud data, census data and satellite imagery** to refine surface risk predictions of flooding in Bangladesh	Bangladesh	Identify and define at-risk populations as well as improve DRM planning	2019	Tier III	International
Goal 11: Sustainable Cities and Communities	Fragile Cities	Igarape Institute	Structured and unstructured sources	Fragility	Rate cities on a fragility index using **structured and unstructured sources**	Worldwide	Understand the dimensions of city fragility through a data visualization platform	2010–2017	Tier I	Academic, NGO, international
Goal 5: Gender Equality	Chega de FiuFiu	Chega de FiuFiu	Crowd-sourced reports on harassment and gender-based discrimination	Gender discrimination, violence against women	Geolocate citizen reports to create a map that informs hotspots for dangerous and uncomfortable places for women using **crowd-sourced and geo-located reports of harassment incidents**	Brazil	Render visible gender-based street harassment hotspots	2013–present	Not classi-fied	NGO
Goal 5: Gender Equality	Mapping eVAW	Hamara Internet	Crowd-sourced reports on electronic harassment	Gender discrimination, violence against women	Geolocate **citizen reports** of Electronic Violence Against Women (eVAW) to map incidents of gender violence in different cities of Pakistan	Pakistan	Render visible gender-based street harassment hotspots	2014–2016	Not classi-fied	NGO, International

SDG/impact field	Project or initiative	Organization	Data sources and tools	What is monitored or studied?	Description	Country or region	Implications of using data-driven approaches	Years	Tiers	Type of organization
Goal 16: Peace, Justice and Strong Institutions	Ibrahim Index of African Governance	Mo Ibrahim Foundation	International agency information, data projects, surveys	Governance performance	Measure and monitor governance performance using **data aggregated, clustered and weighted from multiple sources**, including international agencies, data projects and surveys	Africa	Enhance the transparency and accountability of governance by joining multiple sources of data	2016–present	Tier II	International
Goal 5: Gender Equality	Hollaback!	Knight Foundation	Crowd-sourced reports on harassment	Harassment	Collect and track **crowd-sourced reports of online, street and other forms of harassmen**t	USA, Bosnia and Herzegovina, Canada, Colombia, and 12 other countries	Render visible rarely reported and culturally accepted harassment	2019	Not classi-fied	NGO

THE WESTMINSTER PARLIAMENT'S IMPACT ON UK AI STRATEGY

LORD CLEMENT-JONES CBE

former Chair of the House of Lords Select Committee on AI, Co-Chair of the All-Party Parliamentary Group on AI, founding member of the OECD Parliamentary Group on AI, member of the Council of Europe's Ad-hoc Committee on AI, and House of Lords Liberal Democrat Spokesperson for Digital.

SDG8 - Decent Work and Economic Growth

SDG9 - Industry, Innovation and Infrastructure

SDG16 - Peace and Justice Strong Institutions

THE WESTMINSTER PARLIAMENT'S IMPACT ON UK AI STRATEGY

ABSTRACT

We have the right foundations for a UK AI Strategy and Governance Framework. Now we must build on them. In many ways the UK has been in the vanguard in its understanding and appreciation of the impact and implications of AI on society. Both the UK Parliament, through its select committees and all-party groups, and Government, with a series of policy initiatives and the setting up and developing of a number of key institutions—such as the Office for AI, the Centre for Data Ethics and Innovation, the AI Council and the Alan Turing Institute—have demonstrated they understand the challenges. But the task now is to coordinate the many stakeholders in the future of AI in the UK to agree on a risk-based approach to AI governance that broadly conforms to initiatives from the EU, the Council of Europe and the OECD, as well as a set of common standards for a range of audit and risk assessment tools. That way, developers and those procuring and deploying AI will get the regulatory certainty they now badly need for the UK to retain a leading role in AI development.

INTRODUCTION

At the World Economic Forum meeting in Davos in January 2018 the then Prime Minister, Theresa May (2018), in her keynote speech, focused on Britain's Strategy for the development of AI and how she wanted the UK to lead the world in deciding how AI can be deployed in a safe and ethical manner:

> ...In a global digital age we need the norms and rules we establish to be shared by all.
>
> This includes establishing the rules and standards that can make the most of Artificial Intelligence in a responsible way, such as by ensuring that algorithms don't perpetuate the human biases of their developers.
>
> So we want our new world-leading Centre for Data Ethics and Innovation to work closely with international partners to build a common understanding of how to ensure the safe, ethical and innovative deployment of Artificial Intelligence.

There could have been no stronger demonstration of the emphasis and importance the UK placed and still places on the development of an internationally competitive, indeed world-leading, AI strategy.

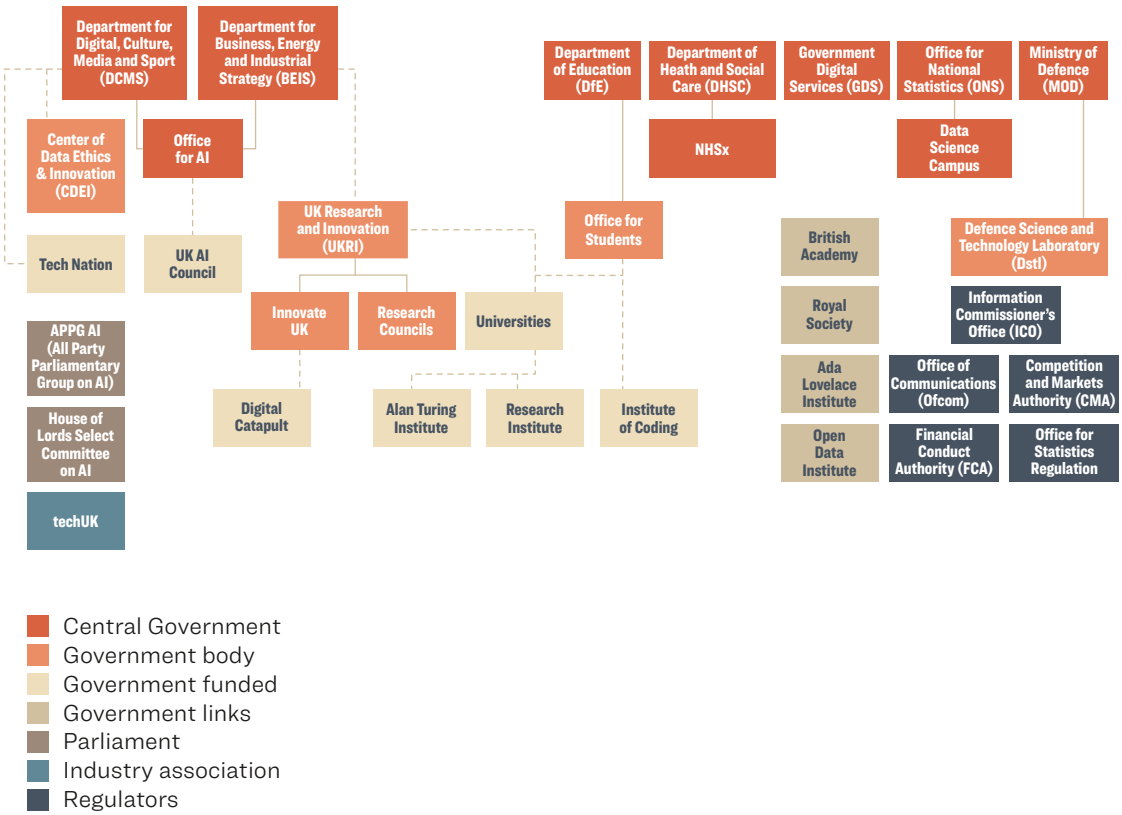
THE UK AI ECOSYSTEM

Despite the priority placed on it, however, the ecology of AI policymaking in the UK, as it has developed, has become complicated in that no single institution has been given responsibility for developing and implementing a national AI strategy. At present:

- On the government side, we have the Office for AI, with oversight split between the Business (BEIS) and Culture (DCMS) Departments. Within government, deployment of AI comes under the new Central Digital and Data Office (CDDO) and its technical arm, the Government Digital Service, although in healthcare, the National Healthcare Service user experience unit, NHSX, has a particular remit to develop digital and AI solutions.
- On the regulatory front, we have the Information Commissioner's Office (ICO) overseeing the crucial area of data governance, the Centre for Data Ethics and Innovation (CDEI) advising on the ethical underpinning of AI use and deployment, and a whole collection of regulators—Ofcom (the telecoms regulator), the Competition and Markets Authority (CMA) and the Financial Conduct Authority (FCA)—having an interest in algorithmic operation in their sectors.
- On the research and innovation side, we have UK Research and Innovation (UKRI) as the oversight body and the Alan Turing Institute as the center of excellence for AI research. Through its Fellows, the Institute has relationships with a whole host of universities, institutions and catapults, such as the Engineering and Physical Sciences Research Council (EPSRC), Innovate UK and the Digital Catapult, which are tasked with helping to commercialize R&D in this space.
- On the non-governmental front, the key instrument has been the AI Council, comprised of the key AI innovators, developers and users in business, academia and the public and third sectors, that advises government on policy and research.
- Other institutions, such as the Royal Society, the British Academy, the Open Data Institute, NESTA and the Ada Lovelace Institutes, and CogX, the extraordinary AI conference community, have all had a major influence on the direction of travel. Big Brother Watch and Liberty too are NGOs which have campaigned on the impact of intrusive AI surveillance.

| FIGURE 1 |

"AI public policy and regulation in the UK"
by PricewatershouseCoopers LLP (2021) in (Axente, 2021).



When it comes to international relationships with bodies such as the UN, OECD, Council of Europe and Global Partnership on AI (GPAI) which relate to the development of international policy on AI, the expertise of the Turing, CDEI and Office for AI are variously brought into play.

This complexity is a contrast with many countries—such as Canada and Germany, where the landscape is a great deal simpler—and is both a strength and weakness, as I hope to explain in this chapter. Despite what has been described as a collaborative ethos which "links government departments with academia, think tanks and businesses to co-create and iterate the story of AI in the UK" (Axente, 2021), there have been criticisms of the pace and focus of policy and strategy development in a number of critical areas, such as algorithmic decision-making in the public sector and the deployment of live facial recognition technology in public spaces.

My task and that of my Parliamentary colleagues—through a variety of reports from the Commons Science and Technology Committee, the House of Lords AI Select Committee and the All-Party Parliamentary Group on AI in particular, but also including our Committee on Standards in Public Life,

and parliamentary debates and questions—has been, and continues to be, to understand these complications and the roles played by the various institutions, to keep up the pressure for strategic coordination on AI policy, and to influence AI policy formation and implementation, not least in the assessment of opportunity for and risk to society.

Both in the Select Committee and the All-Party Parliamentary AI Group, we have tracked the development of AI solutions and systems in a variety of areas, in education, smart cities, health and energy management in particular and we have examined the potential of individual AI applications. But as parliamentarians, we have also been concerned to ensure the mitigation of the risks of AI in terms of its ethical implications and societal impact.

The formation of an initial UK AI Strategy: The Hall Pesenti Review, the Industrial Policy and the AI APPG

To return to the narrative, however, and the origins of Prime Minister May's speech, the genesis of much of the AI policy contained in it and some of the institutional architecture, was the Hall Pesenti Review. This was an independent review commissioned in March 2017 by the UK Government from Professor Dame Wendy Hall, Regius Professor of Computer Science at the University of Southampton, and Jérôme Pesenti, then CEO of Benevolent Tech, tasked with reporting on the potential impact of AI on the UK economy. Their review, "Growing the Artificial Intelligence Industry in the UK," was published in October 2017 (Hall and Pesenti, 2017).

Hall and Pesenti (2017) made a number of key recommendations which set a clear course for UK AI strategy:

- Given the importance of data sets to the training and operation of AI systems, data trusts should be developed to provide proven and trusted frameworks to facilitate the sharing of data between organizations holding data and organizations looking to use data to develop AI.
- The supply of skills should be improved by embracing the value and importance of a diverse workforce for AI, with a major program of students to pursue Master's-level courses in AI, with an initial cohort of 300 students; one-year conversion Master's degrees in AI for graduates in subjects other than computing and data science; and the creation of a minimum additional 200 PhD places dedicated to AI at leading universities.
- To maximize research, the Alan Turing Institute should become the national institute for AI and data science with the creation of an International Turing AI fellowship program for AI in the UK.
- A UK AI Council should be established to help coordinate and grow AI in the UK.

The UK Government's subsequent "Industrial Strategy: Building a Britain fit for the future," published in November 2017, listed putting AI "at the forefront of the UK's AI and data revolution" as one of four "Grand Challenges" identified as key to Britain's future. The Industrial Strategy recognized that ethics would be key to the successful adoption of AI in the UK, which led to the establishment of the Centre for Data Ethics and Innovation in late 2018 with the remit "to make sure that data and AI deliver the best possible outcomes for society, in support of their ethical and innovative use" (United Kingdom Government, 2017).

The Industrial Strategy then led, in early 2018, to the £950m AI Sector Deal, which incorporated nearly all the recommendations of the Hall Pesenti review and established a new Government Office for AI designed to coordinate their implementation (United Kingdom Government, 2019a).

UK parliamentary activity: "AI in the UK: Ready, Willing and Able?"

In the same month as the Hall Pesenti review was announced, Stephen Metcalfe and I held the first meeting of the new All-Party Parliamentary Group on AI (APPG AI), founded with the assistance of Justin Anderson, then of the Hypercat Alliance, and the Big Innovation Centre, to meet our concerns about the lack of parliamentary oversight over the future of AI. The APPG AI was also intended as a means of helping Peers and MPs engage with the AI community and was designed to help shape future AI policy in the UK, particularly as regards ethical, moral and societal issues and the governance and regulatory implications.

At the time, in the context of a discussion with the new All-Party Parliamentary Group, I illustrated these ethical and moral questions by reference to the example of Tay, the public-facing AI chatbot from Microsoft which opened and closed within a week in March 2016 due to the racist and sexist content it was producing (Taylor, 2017):

> Are we really going to instil human values in our AI? Do we want to?... If we want to instil the worst aspects of human behaviour, which we seem to be able to do in cases like Tay, or indeed inflict violent behaviour on military robots...we should be thinking about values in a rather different way.

The area of AI ethics and regulation was not part of the Hall Pesenti Review's remit, but shortly after the House of Lords Select Committee of Enquiry into AI, which I was asked to chair, was set up in June 2017, it was appointed "to consider the economic, ethical and social implications of advances in artificial intelligence." From the outset of the inquiry, we asked ourselves, and our witnesses, five key questions (United Kingdom Parliament, 2018a, p. 153):

1. How does AI affect people in their everyday lives, and how is this likely to change?
2. What are the potential opportunities presented by AI to the UK? How can these be realized?
3. What are the possible risks and implications of AI? How can these be avoided?
4. How should the public be engaged with in a responsible manner about AI?
5. What are the ethical issues presented by the development and use of AI?

With the key AI institutions and ambitions in place, the subsequent report of the House of Lords Select Committee which was published in April 2018 under the title "AI in the UK: Ready, Willing and Able?" had a great deal to say about the AI strategy which was taking shape, whether it was the right one and the need for coordination in delivering it. We also focused heavily on the need for ethical deployment and use of AI systems.

Our inquiry concluded that the UK was in a strong position to be among the world leaders in the development of AI given that it is home to leading AI companies, a dynamic academic research culture, a vigorous start-up ecosystem, and a constellation of legal, ethical, financial, and linguistic strengths located in close proximity to each other. AI, handled carefully, could be a great opportunity for the British economy (United Kingdom Parliament, 2018a).

Our recommendations were designed to support the Government and the UK in realizing the potential of AI for our society and our economy, and to protect society from potential threats and risks. But we emphasized that if poorly handled, public confidence in AI could be undermined. The UK had a unique opportunity to forge a distinctive role for itself as a pioneer in ethical AI.

We recommended in particular that the Government needed to draw up a national policy framework, in lockstep with the Industrial Strategy, to ensure the coordination and successful delivery of AI policy in the UK as part of this: "The UK must seek to actively shape AI's development and utilisation, or risk passively acquiescing to its many likely consequences" (UK Parliament, House of Lords, 2018).

In anticipation of the OECD's subsequent digital AI principles (OECD, 2019), we proposed five principles that could form the basis of a cross-sector AI code and which could be adopted nationally and internationally. We did not at that point recommend any new regulatory body for AI-specific regulation but said that such a framework of principles could underpin regulation, should it prove to be necessary in the future, and that existing regulators would be best placed to regulate AI in their respective sectors.

We were particularly concerned about ensuring that the prejudices of the past would not be unwittingly built into automated systems, that such systems should be carefully designed from the beginning and that, as Hall Pesenti (2017) had recommended, we should see the development of new frameworks and mechanisms, such as data trusts. To ensure that our use of AI did not inadvertently prejudice the treatment of particular groups in society, we called for the Government to incentivize the development of new approaches to the auditing of datasets used in AI, and for greater diversity in the training and recruitment of AI specialists. Given the huge potential disruption in employment, we also advocated a significant Government investment in skills and training. Retraining would become a lifelong necessity. All this added up to a package which we believed would ensure that the UK could remain competitive in this space whilst retaining public trust. It remains an influential public policy document on AI in that it took a holistic approach in framing AI covering opportunities alongside societal impact, risks, ethics, and public engagement.

The Government Response 1

The test of any parliamentary report, however, is whether the Government has accepted its recommendations. In that respect, it was a mixed scorecard (United Kingdom Government, 2018).

On the plus side there was:

- Acceptance by the Government of the need to retain and develop public trust through an ethical approach both nationally and internationally.
- The appointment of the new Chair of the Centre for Data Ethics and Innovation and the start of a consultation on its role and objectives including exploration of governance arrangements for data trusts and access to public datasets.
- Recognition by the Competition and Markets Authority (CMA) of competition issues around data monopoly.
- Recognition of the need for multiple perspectives and insights during the development, deployment and operation of algorithms, as well as diversity in the AI workforce.
- Commitment to a National Retraining Scheme.

On the other hand:

- The AI Sector deal was a good start, but only a start, towards a national policy framework.
- It was unclear whether the new Government Office for AI would deliver greater coordination with the new Council for AI and whether the Centre for Data Ethics and Innovation would have the resources and status it needed to deliver on a national ethical framework.
- There was only qualified acceptance by the Department for Health of the need for transparency particularly in healthcare applications.

- The Department for Education was defensive on its record on apprenticeships and the need to reform the Apprenticeship Levy and appeared to have limited understanding of the need for creative and critical thinking skills as well as computer skills.
- The Ministry of Defence, in its section of the response, continued to rely on a definition of "autonomous" in relation to Lethal Autonomous Weapons Systems (LAWS) which no other country shared.

So, some omens from the Government were good, others less so. We did accept at that stage, however, that AI policy was in its infancy in the UK and that the Government had made a good start in policymaking.

"AI in the UK: No Room for Complacency"

In autumn 2020, the House of Lords Liaison Committee, which coordinates the work of Lords' Select Committees, asked me and a number of my former colleagues on the AI Select Committee to follow up with a review of progress made since our previous report.

In December 2020, our new report, "AI in the UK: No Room for Complacency," examined the progress made by the UK government (United Kingdom Parliament, 2020a). After interviews with government ministers, regulators, and other key players, our new report made a number of key recommendations:

Public trust and data governance

Greater public understanding is essential for the wider adoption of AI, and also to enable challenge to any organization deploying AI in an ethically unsound manner. Active steps must be taken by the government to explain to the general public the use of their personal data by AI. In addition, the development of policy to safeguard the use of data, such as data trusts, needed to pick up pace. Otherwise, it risked being left behind by technological developments.

A code of ethics

Since our original report, a clear consensus had emerged that ethical AI is the only sustainable way forward. The UK had in the meantime become a signatory of the OECD Recommendation on AI embodying five principles for responsible stewardship of trustworthy AI (OECD, 2019) and the G20 non-binding principles on AI. We said that the time has come for the UK Government to move from deciding what the ethics are to how to instill them in the development and deployment of AI systems. The Government must lead the way on making ethical AI a reality. To not do so would be to waste the progress it had made to date and to squander the opportunities AI presents for everyone in the UK. We called for the CDEI to establish and publish national standards for the ethical development and deployment of AI.

Risk and regulation

In this regard, we said users and policymakers needed to develop a better understanding of risks and how they can be assessed and mitigated, in terms of the context in which AI is applied. The report recommended that the ICO—with input from the CDEI, the Office for AI, and the Alan Turing Institute—develop a training course for regulators.

Skills and upskilling

As regards skills, we considered that government inertia was a major concern. It was clear that the pace, scale, and ambition of government action did not match the challenge facing many people working in the UK. As and when the COVID-19 pandemic receded and the UK Government had to address the economic impact of it, the nature of work would change. AI would not necessarily make huge numbers of people redundant but there would be a need for different jobs and skills. The Government and industry needed to take steps to ensure that the digital skills of the UK are brought up to speed, as well as to ensure that people have the opportunity to reskill and retrain to be able to adapt to the evolving labor market. A specific training scheme should be designed to support people to work alongside AI and automation, and to be able to maximize its potential. There was an urgent need too for diversity and inclusion in the AI workforce and for greater digital literacy.

Strategic coordination

Our conclusion was that the UK Government had done well to establish a range of bodies to advise it on AI over the long term. However, we cautioned against complacency. Coordination between the various bodies involved with the development of AI, including the various regulators, is essential. The UK Government needed to better coordinate its AI policy and the use of data and technology by national and local governments. We said that a Cabinet Committee must be created whose first task should be to commission and approve a five-year strategy for AI. The strategy should prepare society to take advantage of AI rather than be taken advantage of by it.

International engagement

A final conclusion in our new report was that the UK should show global leadership on shared challenges through bodies such as the Global Partnership on AI. As regards LAWS, however, we were as concerned as previously about the lack of action, especially in the light of the creation of a new Autonomy Development Centre within the Ministry of Defence which we believed would be inhibited by the failure to align the UK's definition of autonomous weapons with those of international partners.

The AI Roadmap

The AI Roadmap from the AI Council came out soon afterwards, in January 2021, and significantly shared a number of themes with the latest Lords report (United Kingdom Government, 2021a). In particular, the AI Roadmap recommended the creation of a national AI strategy in the UK and stressed that the UK should lead in developing appropriate standards on data governance and enact clear and flexible regulation building on guidance from regulators such as the ICO.

The AI Roadmap noted that "the public should be reassured that the use of AI is safe, secure, fair, ethical and overseen by independent entities and the ability for regulators to enforce sanctions." In addition to the continuous development of industry standards and suitable regulations and frameworks for algorithmic accountability, the Roadmap emphasized the need to be world-leading in respect of responsive regulation and governance and suggested what it called an "independent entity" to advise on "the next steps in the evolution of governance mechanisms, including impact and risk assessments, best practice principles, ethical processes and institutional mechanisms that will increase and sustain public trust" (United Kingdom Government, 2021a).

The Government Response 2

The Government response to the latest Lords report, published in February 2021, was again a mixed bag, especially in the area of skills, but the central suggestion of a National AI Strategy was taken up and is expected to be delivered in autumn 2021 at the time of writing, no doubt with a great deal of input from the AI Council's Roadmap—and I hope from Parliament too.

The Government expressed both its welcome for the report's positive recommendation and its message that there was no room for complacency. They noted the messages in common with the AI Council's Roadmap, in particular that the Government's approach needs to focus on establishing the right arrangements between institutions: across Government and the public sector, between regulators, and with academia and industry to "ensure that momentum gained over the past few years is not lost, but instead reinvigorated to drive economic recovery and prosperity across the union, and allow us to use our lead in AI to solve global challenges" (United Kingdom Government, 2021e).

As regards public understanding and data, the Government fully recognized the critical importance of furthering this by accelerating work on actionable legal frameworks for data governance, in particular for public health data and on addressing issues of data competition, as recommended by the Digital Markets Taskforce report and the Furman Review (United Kingdom Government, 2019c, 2020a).

As regards ethics and recommendations about establishment of national standards for the ethical development and deployment of AI, the Government's response undertook that the Government Digital Service (GDS) would explore the development of an appropriate and effective mechanism to deliver more transparency on the use of algorithmic assisted decision-making within the public sector and it was considering what the Centre for Data Ethics' future functions should be.

In terms of jobs and the criticism of inadequate action by government on predicting the skills and retraining that will be needed, they asserted that the future of work is a key policy area for a number of departments across Government. The Government's planned major expansion of post-18 education and training to level up and prepare workers for the post-COVID-19 economy to include a Lifetime Skills Guarantee. They highlighted the announcement in 2020 of AI apprenticeships. They expressed agreement with both our report and the AI Roadmap about the need for diversity and underlined progress on the delivery of a thousand more PhDs at 16 Centres for Doctoral Training, a hundred industry-funded Master's courses, and 2,500 AI conversion courses with a thousand scholarships for people from underrepresented groups.

As regards the Committee's recommendations on public trust and regulation, the Government noted the formation of the Digital Regulation Cooperation Forum (DRCF) by the CMA, the ICO and Ofcom to support regulatory coordination in digital markets and cooperation on areas of mutual importance, which could result in an AI regulatory training course being developed. The Government drew attention to its ambitions for an online media literacy strategy which would "ensure a coordinated and strategic approach to online media literacy education and awareness for children, young people and adults" (United Kingdom Government, 2020g).

On strategic coordination, the Government acknowledged that responsibility for AI policy and driving uptake across the economy is split across ministers in both the DCMS and BEIS, but insisted that this meant that the benefits of AI were realized across wider government and agencies.

Perhaps the most surprising and heartening response was on LAWS:

> We agree that the UK must be able to participate in international debates on autonomous weapons, taking an active role as moral and ethical leader on the global stage, and we further agree the importance of ensuring that official definitions do not undermine our arguments or diverge from our allies.

Although an operative definition for LAWS themselves had not yet been agreed upon, the UK had recently accepted NATO's latest definitions of "autonomous" and "autonomy." The Government pointed out that the UK had a prominent voice at discussions of this issue at the UN Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) on LAWS. Furthermore, the MOD was preparing to publish a new Defence AI Strategy and "will continue to be proactive in addressing ethical issues surrounding the development and use of AI for military purposes."

Specific applications: Parliamentary reports on live facial recognition technology and algorithmic decision-making

In parallel to these contributions by Parliamentary committees to the overall debate about the future opportunities for and governance of AI in general, other parliamentary bodies have addressed the implications of specific uses of AI systems, particularly the use of AI applications in the public sector.

In May 2018, the House of Commons Science and Technology Committee, under the chairmanship of Sir Norman Lamb in its report "Algorithms in Decision-making" (United Kingdom Parliament, 2018b), reported on the use of algorithms in public and business decision-making and warned of the need to identify and tackle bias and of the need for accountability and transparency on the grey areas in the GDPR as regards automated decision-making.

Subsequently, in its review of the Work of the Biometrics Commissioner and the Forensic Science Regulator in July 2019, the Committee recommended a moratorium on the use of Live Facial Recognition technology (LFRT) (United Kingdom Parliament, 2019).

It said that the Government's biometrics strategy

> was not worth the five-year wait. Arguably it is not a "strategy" at all: it lacks a coherent, forward-looking vision and fails to address the legislative vacuum that the Home Office has allowed to emerge around new biometrics.

It called on the Government:

> to issue a moratorium on the current use of facial recognition technology and no further trials should take place until a legislative framework has been introduced and guidance on trial protocols, and an oversight and evaluation system, has been established.

In each case, the Government was extremely reluctant to promise meaningful action, as their responses to each report show. The response to the latter report only arrived nearly two years after publication.

In March 2019, however, the influential Committee on Standards on Public Life (CSPL), the independent advisory public body that advises the Prime Minister on ethical standards across the whole of public life in the UK, under the Chairmanship of Lord Evans of Weardale, set up an inquiry into AI and Public Standards designed to understand the implications of AI for the Nolan principles. These are the seven principles which are expected to govern conduct in public life in the UK: selflessness, integrity, objectivity, accountability, openness, honesty and leadership. The inquiry also aimed to examine whether the Government's policy was up to the task of upholding those standards as AI is rolled out across our public services, highlighting ethical concerns arising from, for example, data bias and algorithmic "black boxes" and of ensuring that AI is only used for the public good.

The CSPL report, published the following February, "Artificial Intelligence and Public Standards" (United Kingdom Government, 2020c), made a number of key recommendations to strengthen the UK's ethical framework around the deployment of AI in the public sector. Their message to the government was that the UK's regulatory and governance framework for AI in the public sector remains a work in progress and deficiencies are notable. On the issues of transparency and data bias

in particular, there was an urgent need for guidance and regulation. The Government should make clear which principles are to be followed. They also said that upholding public standards would require action from public bodies using AI to deliver frontline services. All public bodies should state how their use of AI complies with the law surrounding data-driven technology and implement clear, risk-based governance for their use of AI and there should be a regulatory assurance body (notably the CDEI), which identifies gaps in the regulatory landscape and provides advice to individual regulators and government on the issues associated with AI. The Government should also consider how an AI impact assessment requirement could be integrated into existing processes to evaluate the potential effects of AI on public standards including on the potential impact of a proposed AI system on public standards at project design stage. Such assessments should be mandatory and should be published.

The Government's again belated response, in May 2021 (United Kingdom Government, 2021d) was, however, broadly positive in contrast to its response to the Science and Technology Committee reports. It agreed that the number and variety of principles on AI may lead to confusion when AI solutions are implemented in the public sector. It asserted that the UK Government had signed up to multilateral principles on AI, including the OECD principles, and was committed to implementing these through its involvement as a founding member of the Global Partnership on AI. The WEF AI Procurement Guidelines had led to a UK-specific AI Procurement Guide. These Guidelines were developed by the Office for Artificial Intelligence, in collaboration with the World Economic Forum Centre for the Fourth Industrial Revolution, the Government Digital Service, Government Commercial Function, and the Crown Commercial Service and seek to enable public bodies to buy AI in a more confident and ethically responsible manner.

In order to ensure more clarity on ethical principles and guidance, the Government had published an online resource, the Data Ethics and AI Guidance Landscape (United Kingdom Government, 2020d), with a list of various data ethics-related resources intended for use by public servants. They would explore the development of an appropriate and effective mechanism to deliver transparency on the use of algorithms facilitating semi-autonomous decision-making within the public sector. The Equality and Human Rights Commission (EHRC) would be developing guidance for public authorities on how to ensure any AI work complies with the public sector equality duty.

The position regarding deployment of specific AI systems by the government, however, is still very unsatisfactory. For example:

- As regards LFRT, the government, in its belated response to the Science and Technology Committee in March 2021 (United Kingdom Parliament, 2021c), promised National College of Policing guidance on the use of live facial recognition (LFR) "consistent with the Bridges' judgment," but the Science and Technology Committee itself took the unusual step of writing to Ministers in the Home Office (United Kingdom Parliament, 2021b) expressing "serious concerns about the lack of progress that has been made by the Government in the areas of forensic market sustainability, laboratory accreditation, biometrics governance, and custody image management" and asking for an update on the national guidance and whether the government intended to introduce a clarified legislative framework for automatic facial recognition technology. Draft guidance is now subject to consultation but has already attracted criticism from present and former Surveillance Camera Commissioners.
- As regards algorithmic decision-making, in the wake of controversy over the use of algorithms in education, housing, and immigration, we have seen the publication of the government's new Ethics, Transparency and Accountability Framework for Automated Decision-Making for use in the public sector (United Kingdom Government, 2021c), but there is no satisfactory compliance and enforcement mechanism via the CDDO or the Cabinet Office to ensure that its principles are adhered to.
- Big Brother Watch's Poverty Panopticon has in the meantime illustrated the widespread issues in algorithmic decision making which have arisen at local government level (Big Brother Watch, 2021).

As a result, in the past year, in respect of both LFRT and algorithmic decision-making, I have put forward private members' bills (the Public Authority Algorithm Bill and the Automated Facial Recognition Technology (Moratorium and Review) Bill (United Kingdom Parliament, 2020b) which are designed to provide a strong legislative and regulatory framework which protects civil liberties. I have also raised the issue of regulation in debates and questions.[50]

The House of Lords' new Justice and Home Affairs Committee is now following up some of these concerns with an inquiry into new technologies in law enforcement (United Kingdom Parliament 2021a).

The scorecard of parliamentary influence

Taking stock of where government action has been taken and policy developed over the more general AI landscape, there is no doubt that progress has been made, although it is difficult to calibrate exactly where Parliamentary influence has been decisive. In most cases, it is more likely to have helped maintain momentum or provide cause for thought rather than fundamentally change the direction of policy.

On the upside, progress in a whole host of areas has been made:

- The ICO, Alan Turing Institute, CDEI, and Office for AI have agreed to work together to develop, roll out, and monitor training for regulators on issues around AI.
- The Office for AI is currently working on a National AI Strategy and has been an active force in the field with its Guide to using AI and Procurement Guidelines.
- The Government has published a Framework for Algorithmic decision-making in the Public Sector.
- The CDEI has proved its worth with numerous reports, such as one on bias in algorithmic decision-making which focused on a number of particular sectors (United Kingdom Government, 2020f) and one on online targeting used to promote and personalize content and target advertising (United Kingdom Government, 2020h).
- The CDEI has also published "snapshot papers" on Deepfakes and AudioVisual Disinformation, AI and Personal Insurance, and Smart Speakers and Voice Assistants (United Kingdom Government, 2019b). In addition, it published the AI Barometer, described as a "major analysis of the most pressing opportunities, risks and governance challenges associated with AI and data use in the UK" (United Kingdom Government, 2020b).
- Valuable work has been commissioned from the Open Data Institute (ODI) on data institutions and trusts.
- The Alan Turing Institute has played a major role in collaboration across the AI landscape nationally and internationally, including bringing together 400 fellows, working on the ExplAIn project with the ICO and developing policy with the OECD and Council of Europe.
- The AI Council, after an uncertain start during the COVID-19 pandemic, has produced its influential AI Roadmap.
- A number of our regulators, such as the FCA and the ICO, have led the way on regulatory sandboxing.

50. See, for example, United Kingdom Parliament (2020c).

In other areas, Parliamentary committees have been less influential. For example:

- We need to explicitly adopt a set of principles nationally, install clear risk evaluation and compliance mechanisms in the public sector, and turn ethics into practical standards for corporate governance to enable the evaluation of use cases and the design of AI systems which can help decide whether and where hard law is needed, as opposed to soft law or guidance. Following from that, we need to develop tools for audit, impact assessment, certification, and continuous monitoring.
- The CDEI is now a key player in the AI landscape and must be put on a statutory basis and its role clearly specified.
- We need to accelerate progress on data trusts and other data-sharing frameworks. The ODI has done good work, but clear legal structures are not in place yet and much more should be done to create trusted vehicles for public data such as for NHS data, drawing on international work where relevant.
- We need to do likewise with progress on online and digital literacy, itself a major route to securing public trust. Simply handing the duty to Ofcom in new online harms legislation as proposed is inadequate.
- There has been little influence on the widespread deployment of LFRT by public bodies.
- A really dominant consideration for us all in this field is the assessment of the impact of AI on jobs and assessment of the skills needed in the future, the diversity in the workforce required and the scale of the reskilling requirement demanded by the move to automation. Government needs to recognize the urgency of the employment implications of AI and the disruption it will cause and the need, by a dimension, to heighten our digital skills and reskilling ambitions. The pace, scale and ambition of UK government action does not match the upskilling challenge facing many people working in the UK. Much more action too needs to be taken to develop greater diversity and inclusion in the tech workforce.

CONCLUSION: GOVERNMENT AT THE REGULATORY CROSSROADS

> As can be seen from the above narrative, in terms of wider AI National Strategy there has been a great deal of agreement between Government and parliamentary bodies about the desired direction of travel, although parliamentary committees have been impatient for greater pace and ambition.
>
> In the creation of a revised national AI strategy, coordination of the work of the key actors—such as the Office for AI, the AI Council, the ICO, the CDEI and the Alan Turing Institute—has been and will continue to be crucial, in delivering plans such as the AI Deal both at the national level and internationally. AI is a complicated and emotive subject. The increased reliance on technology caused by the COVID-19 pandemic has highlighted the opportunities and risks associated with its use, in particular with the use of data. As a result, it has never been clearer that we need to retain public trust in the adoption of AI.
>
> That is clearly accepted by the UK Government but there is some doubt whether it also accepts that making ethical AI a reality involves assessing the risks of AI in context, particularly in terms of impact on civil and social rights, and then, depending on the risk assessed, setting standards, or regulating for the ethical design, development, and deployment of AI systems.

We need much greater definition of when regulation or lesser corporate governance requirements are appropriate. In 2021, the international AI community started to move towards deciding how to do this practically, with the increasing adoption, by international bodies such as the EU and the Council of Europe's Ad Hoc Committee on Artificial Intelligence (CAHAI), of a cross-sector, horizontal risk-based approach to AI governance and regulation.

Key initiatives in that process have been the EU's proposal for an AI Regulation (the "AI Act") (European Commission, 2021), published in April 2021, and the Feasibility Study drawn up and agreed to in December 2020 by CAHAI (Council of Europe, 2020), which explores options for an international legal response based on Council of Europe standards in the field of human rights, democracy, and the rule of law.

We are now coming out of the foothills in determining where we can and should rely on ethical codes and where we should prescribe ethical governance or go the whole hog and regulate. The debate over hard and soft law in this area is by no means concluded but there is no doubt that pooling expertise at international level will bear fruit. The UK is therefore at a crossroads. In April 2021, Britain hosted the G7 meeting of Digital and Technology Ministers and hosted the inaugural Future Tech Forum in November, but we need to go beyond principles in establishing international AI governance standards and solutions. There is a sense that the goals of trustworthy AI and positioning the UK as a leader in the adoption of ethical AI have been diluted.

In my view, to mitigate risks and retain public trust, whether in the public or private sector, the cardinal principle must be that AI needs to be seen to be our servant, not our master. The question is whether that principle is accepted by UK policy makers and regulators together with the duty to ensure that regulatory policies and solutions are classified and calibrated according to ascending degrees of AI risk.

So, does the UK proceed with a similar "horizontal" approach to that adopted by the Council of Europe and the EU or regulate for AI sector by sector as issues present themselves? The way forward could well be an initial overall non-sector-specific requirement for the adoption of AI impact assessments to calculate the risks of the adoption of a particular AI system followed by obligatory regular audit and monitoring of high-risk systems.

In July 2020, the ICO published Guidance on Artificial Intelligence and Data Protection (United Kingdom Government 2020e) to help organizations mitigate the risks of AI arising from a data protection perspective. The guidance set out a framework and a methodology for auditing AI systems. The guidance's proportionate and risk-based approach contains an auditing methodology with tools and procedures for audits and investigations, detailed guidance on AI and data protection, and a toolkit providing practical support to organizations auditing the compliance of their own AI systems.

Following on from this guidance, the ICO has now published a beta version of an AI and Data Protection Risk Toolkit (United Kingdom Government, 2021b) designed to help organizations using AI to understand the risks to individuals' information rights and providing suggestions on best practice measures that can be used to manage and mitigate the risks. As a result, we do have the foundations for an ethical UK AI risk-based regulatory regime, which is also attractive to developers and investors.

The UK Government has now promised to publish an AI Governance White Paper setting out its regulatory proposals early this year. We certainly have many of the right foundations, but it is still unclear how much Parliamentary influence there will be in determining how the Government builds on them.

January 2022

REFERENCES

Axente, M. L. 2021. How do we ensure the responsible use of AI by governments? Digital Tech ITP, April 7. https://digitaltechitp.nz/2021/04/07/how-do-we-ensure-the-responsible-use-of-ai-by-governments/

Big Brother Watch. 2021. *Poverty Panopticon: The hidden algorithms shaping Britain's welfare state*. London. https://bigbrotherwatch.org.uk/wp-content/uploads/2021/07/Poverty-Panopticon.pdf

Council of Europe. 2020. Feasibility Study. Brussels, Ad hoc Committee on Artificial Intelligence (CAHAI). https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

European Commission. 2021. *White Paper on AI: A European approach to excellence and trust. Brussels.* https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Hall, W. and Pesenti, J. 2017. *Growing the artificial intelligence industry in the UK.* London, Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

May, T. 2018. Keynote Speech. 25 January, World Economic Forum, Davos, Switzerland. https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/

OECD. 2019. Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Taylor, C. 2017. Lord Clement-Jones: On regulation and ethical and moral dilemmas in artificial intelligence. LinkedIn. 17 May. https://www.linkedin.com/pulse/lord-clement-jones-regulation-ethical-moral-dilemmas-claire-taylor/?trk=read_related_article-card_title

United Kingdom Government. 2017. *Industrial Strategy: Building a Britain fit for the future*. London, Department for Business, Energy and Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730048/industrial-strategy-white-paper-web-ready-a4-version.pdf

———. 2018. *Government Response to the Lords Select Committee on Artificial Intelligence Report*. London, Department for Business, Energy and Industrial Strategy. https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report

———. 2019a. AI Sector Deal. London, Department for Business, Energy and Industrial Strategy. https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal

———. 2019b. CDEI publishes its first series of three snapshot papers on ethical issues in AI. London, Centre for Data Ethics and Innovation. 12 September. https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai

———. 2019c. *Unlocking Digital Competition: Report of the Digital Competition Expert Panel*. London, Digital Competition Expert Panel. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf

———. 2020a. *A New Pro-competition Regime for Digital Markets: Advice of the Digital Markets Taskforce.* London, Digital Markets Taskforce. https://assets.publishing.service.gov.uk/media/5fce7567e90e07562f98286c/Digital_Taskforce_-_Advice.pdf

———. 2020b. *AI Barometer*. London, Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf

———. 2020c. *Artificial Intelligence and Public Standards*. London, Committee on Standards on Public Life. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/868284/Web_Version_AI_and_Public_Standards.PDF

———. 2020d. Data ethics and AI guidance landscape. Web page. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/guidance/data-ethics-and-ai-guidance-landscape

———. 2020e. *Guidance on AI and Data Protection*. London, Information Commissioner's Office. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/

———. 2020f. *Review into Bias in Algorithmic Decision-making*. London, Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf

———.2020g. *Online Harms White Paper*. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

———.2020h. *CDEI Review of Online Targeting*. London, Centre for Data Ethics and Innovation. https://www.gov.uk/government/publications/cdei-review-of-online-targeting

———. 2021a. *AI Roadmap*. London, AI Council. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949539/AI_Council_AI_Roadmap.pdf

———. 2021b. Blog: New toolkit launched to help organisations using AI to process personal data understand the associated risks and ways of complying with data protection law. London, Information Commissioner's Office. https://ico.org.uk/about-the-ico/media-centre/blog-new-toolkit-launched-to-help-organisations-using-ai/

———. 2021c. *Ethics, Transparency and Accountability Framework for Automated Decision-Making*. London, Office for Artificial Intelligence. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

———. 2021d. *Government Response to the Committee on Standards in Public Life's 2020 Report AI and Public Standards*. London, Department for Digital, Culture, Media and Sport. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/987905/Government_Response_to_the_Committee_on_Standards_in_Public_Life_s_2020_Report_AI_and_Public_Standards__Accessible_version_.pdf

———. 2021e. *Government Response to the House of Lords Select Committee on Artificial Intelligence*. London, Department for Digital, Culture, Media and Sport. https://www.gov.uk/government/publications/government-response-to-the-house-of-lords-select-committee-on-artificial-intelligence/government-response-to-the-house-of-lords-select-committee-on-arti-cial-intelligence

United Kingdom Parliament. 2018a. *AI in the UK: Ready, Willing and Able?* London, House of Lords Artificial Intelligence Committee. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

———. 2018b. *Algorithms in Decision-Making*. London, House of Commons Science and Technology Committee. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/35102.htm

———. 2019. The work of the Biometrics Commissioner and the Forensic Science Regulator. Web page. London, House of Commons Science and Technology Committee. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/1970/197001.htm

———. 2020a. *AI in the UK: No Room for Complacency. Seventh Report of Session*. London, House of Lords Liaison Committee. https://publications.parliament.uk/pa/ld5801/ldselect/ldliaison/196/196.pdf

———. 2020b. *Automated Facial Recognition Technology (Moratorium and Review)* (HL Bill 87). London. https://publications.parliament.uk/pa/bills/lbill/58-01/087/5801087_en_1.html

———. 2020c. Facial recognition surveillance. London, House of Lords Hansard. https://hansard.parliament.uk/Lords/2020-01-27/debates/E1922D0C-2EA8-4ED1-89F3-B7EE2475EDED/FacialRecognitionSurveillance#contribution-A8290171-B9DD-45A6-B2F4-4F87BD55D783

———. 2021a. Call for evidence launched on new technologies in law enforcement. London, Justice and Home Affairs Committee. Announcement. July 22. https://committees.parliament.uk/committee/519/justice-and-home-affairs-committee/news/156778/call-for-evidence-launched-on-new-technologies-in-law-enforcement/

———. 2021b. Forensics and biometrics: follow-up. Letter. July 20. London, House of Commons. https://committees.parliament.uk/publications/6876/documents/72517/default/

———. 2021c. *Government Response to the Work of the Biometrics Commissioner and the Forensic Science Regulator*. London, Department for Digital, Culture, Media and Sport. https://publications.parliament.uk/pa/cm5801/cmselect/cmsctech/1319/131902.htm

ARTIFICIAL INTELLIGENCE AND INDIGENOUS RIGHTS

VALMAINE TOKI

Ngatiwai Nga Puhi, Te Piringa. Professor at the Faculty of Law, University of Waikato, New Zealand

ANDELKA M. PHILLIPS

Senior Lecturer in Law, Science and Technology, School of Law, University of Queensland, Australia; and Research Affiliate, HeLEX Centre, University of Oxford, UK.

SDG3 - Good Health and Well-being
SDG7 - Affordable and Clean Energy
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG12 - Responsible Consumption and Production
SDG13 - Climate Action
SDG15 - Life on Land
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

ARTIFICIAL INTELLIGENCE AND INDIGENOUS RIGHTS

ABSTRACT

Considering that AI will be increasingly used across sectors and become pervasive in society, it is urgent to discuss how it can be leveraged for the benefit of different social groups. In particular, populations that face different realities and live by different worldviews could provide a valuable contribution to the development and application of AI. This chapter explores the intersection between Indigenous rights and artificial intelligence (AI) from a procedural and substantive perspective. In order to do this, it begins with an overview of AI development and how the concept can be understood. It then presents a current view on how AI considers Indigenous rights, which provides context for reviewing differing Indigenous worldviews. As the authors are based in New Zealand, the example of the Māori people, the Indigenous People of New Zealand, is used to highlight procedural and substantive steps that should be taken in the development of technologies. The chapter then explores how AI can expressly recognize and reflect Indigenous rights with the case study of a micro-grid implementation on Aotea/Great Barrier Island. The example demonstrates how allowing a community control over their power supply can in turn allow for enhanced protection of their rights, assist in protecting privacy and facilitate both self-determination and data sovereignty. In an age where data has been dubbed "the new oil," questions about the impact of deployment of a wide range of technologies on Indigenous Peoples are of vital importance. We note that technologies are not neutral and will often pose both risks and benefits for communities, including privacy risks for both individuals and groups. The chapter aims to shed light on some of the current issues in AI development with regards to Indigenous populations and encourage further discussion and research in the AI governance space.

INTRODUCTION

This chapter seeks to explore the intersection between Indigenous rights and artificial intelligence (AI) from a procedural and substantive perspective. To unpack this nexus, the chapter begins with an overview of AI development and how the concept can be understood. It then presents a current view on how AI considers Indigenous rights, which provides context for reviewing differing Indigenous worldviews. As the authors are based in New Zealand, the example of the Māori people, the Indigenous People of New Zealand, is used in relation to procedural and substantive steps that should be taken. The chapter then explores the implementation of a micro-grid on Aotea/Great Barrier Island, a case study that highlights the impact of the deployment of a wide range of technologies on Indigenous Peoples and how AI can expressly recognize and reflect Indigenous rights. We demonstrate how allowing a community control over their power supply can in turn allow for enhanced protection of their rights, assist in protecting privacy and facilitate both self-determination and data sovereignty.

It is particularly important to recognize in light of historical injustices, such as the exploitation of Indigenous Peoples and other marginalized groups in scientific research, that technologies are not neutral and often pose both risks and benefits for communities. Our goal is to shed light on some of these issues and to encourage further discussion and research in this space. We wrote this chapter in 2021, in the time of COVID-19, when parts of our own country were again in lockdown and where the need for contact tracing was increasing the erosion of privacy rights for communities and individuals.

While the General Data Protection Regulation (GDPR) (General Data Protection Regulation, 2016) has exerted a global influence, and while Aotearoa/New Zealand updated our privacy legislation in 2020 and enacted the *Privacy Act 2020* (*Privacy Act*, 2020), COVID-19 has created challenges. Specifically, the need for contact tracing threatens privacy and data protection rights and as a result may be viewed as exacerbating existing inequalities. While we do not question the importance of contact tracing, the advent of smartphones has contributed to increased tracking of the population more generally. With attempts to manage the spread of COVID-19, the need for access to data regarding an individual's movements, vaccination status and COVID-19 testing information is increasing the range of entities that have access to health and other sensitive information. For example, at the time of this writing, New Zealand has recently introduced a vaccine passport system which requires individuals to show their vaccination passport to access services, including hair salons and restaurants.

We also need to be conscious of increased cyber threats to all organizations, including attacks on medical databases, such as the WannaCry attack (Landi, 2019). A more recent example in New Zealand is the ransomware attack on the Waikato District Health Board, which impacted more than 4,000 people (New Zealand Herald, 2021; Keall, 2021). As Phillips has noted previously, there is a real need for us to approach technology from a more holistic and inclusive perspective, and a lack of oversight is unlikely "to lead to a safer, fairer world" (Phillips and Mian, 2019). This requires more public debate and engagement of issues related to technology development, adoption and control. Given the wide-ranging potential of AI-based technology, this need is heightened.

BACKGROUND

The term "AI" covers a wide range of technologies that are currently on the market, in development or speculated to be eventuating in the future. Broadly, there is a distinction between the idea of general intelligence or human-level machine intelligence (HLMI or human-like AI) and narrow AI (Bostrom, 2014, pp. 1-21; Fjelland, 2020; Russell, 2021). Many of the technologies currently utilized in our homes and offices can be viewed as examples of narrow AI. This includes things such as spam filtering on email and voice recognition technology or "AI machines" such as Google's AlphaGo (UK Science and

Technology Committee, 2016, p. 5). AlphaGo was designed to play the game Go. It is a good example to demonstrate what narrow AI entails, i.e., it is capable of performing a specific task (in this case playing a game) or a range of tasks well, but cannot excel in other fields outside of its limited sphere. Other similar examples have been developed to play other games, such as chess.

There is also much speculation around the potential development of HLMI, which would mean that an AI agent could use reason in a variety of situations in the same manner as a human (Bostrom, 2014, pp. 3–5; Russell, 2021, p. 514). There is currently much investment in research towards this goal and much discussion of the possibility of a subsequent "intelligence explosion," also referred to as the Singularity. It is uncertain if or when such an event may occur (Fjelland, 2020; Eliot 2020, chapter 4). Many advances to date have been in quite narrow contexts (Russell, 2021, p. 514). The development of autonomous vehicles is also contributing to this; in this context, AI agents may have to deal with very complex questions (Bradshaw-Martin, 2020). Autonomous vehicles also are a good example of the current limitations of AI (*Technology Quarterly*, 2020).

Definitions of AI vary. While there is no universal definition, John McCarthy, who is a seminal figure in the development of AI and introduced the use of the term, defined AI as "the science and engineering of making intelligent machines, especially intelligent computer programs" and stated that "it is related to the similar task of using computers to understand human intelligence" (McCarthy, 2007). This can be contrasted with the idea of natural intelligence, which is displayed by naturally occurring organisms (Williams and Shipley, 2021, p. 45). This has primarily developed within a spectrum of human autonomy, which does tend to anthropomorphize machines. Machine autonomy has been considered in a wide variety of contexts, ranging from autonomous vehicles to humanoid robots that may in the future exhibit HLMI (Calo, 2017).

Several reports are useful in considering how to define AI more broadly (UK Science and Technology Committee, 2016; Executive Office of the President National Science and Technology Council Committee on Technology, 2016; European Parliament Committee on Legal Affairs, 2016). The UK's House of Commons Science and Technology Committee provides a useful definition, which we rely upon for the purposes of this chapter. According to the Report (House of Commons Science and Technology Committee, 2016, pp. 5–6):

> AI can be loosely thought of as a set of statistical tools and algorithms that combine to form, in part, intelligent software that specializes in a single area or task. This type of software is an evolving assemblage of technologies that enable computers to simulate elements of human behaviour such as learning, reasoning and classification.

We also find Kaplan and Haenlein's definition helpful, where they characterize 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, p. 15). In contrast, machine learning can be viewed as "building algorithms that can learn specific concepts for themselves, without being explicitly programmed" (House of Commons Science and Technology Committee, 2016, p. 6).

Furthermore, in April 2021, the European Commission released a proposal for a Regulation on Artificial Intelligence (Artificial Intelligence Act, COM/2021/206). If the *Artificial Intelligence Act* is adopted in the future, its influence is likely to extend beyond the European Union (EU) in a similar way to the EU's GDPR (General Data Protection Regulation, 2016). Consequently, it is useful to also refer to the proposal's definition of an AI system:

> Artificial intelligence system (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. (Artificial Intelligence Act, art. 3.1.)

This definition does capture a wide range of AI-based technologies, which may have relevance to Indigenous Peoples. Given the influence that the GDPR has exerted on privacy law internationally, it is important that the Proposal be subject to international scrutiny and that Indigenous Peoples be included in this discussion. While the Proposal is subject to further amendment, it has already faced criticism. Most notably, the European Digital Rights (EDRi), in collaboration with 119 civil society groups, has released a statement (EDRi, 2021 and 2021a). This statement includes calls for increased transparency and future-proofing in relation to risk in AI systems, as well as prohibitions of certain types of AI systems that represent unacceptable risks, including a ban on all social scoring systems, as well as "emotion recognition systems; discriminatory biometric categorization; AI physiognomy; systems used to predict future criminal activity; systems to profile and risk-assess in a migration context" (EDRi, 2021). The systems that have been identified as representing unacceptable risks are of particular relevance to Indigenous Peoples, especially predictive analytics, as discussed later in this chapter with regard to predictive policing.

AI has the potential to influence all aspects of society, including criminal justice, cybersecurity and medical diagnosis. While AI is touted as a vehicle to address social issues, it comes with challenges around human rights breaches associated with data protection and non-discrimination. For example, the use of predictive policing software, especially for systems that attempt to predict the likelihood of recidivism, has been shown to have significant problems with bias (D'alessandro et al., 2017; O'Neil, 2016, pp. 85–9; Heaven, 2020). As Heaven (2020) notes:

> A tool called COMPAS, used in many jurisdictions to help make decisions about pretrial release and sentencing, issues a statistical score between 1 and 10 to quantify how likely a person is to be rearrested if released. The problem lies with the data the algorithms feed upon. For one thing, predictive algorithms are easily skewed by arrest rates. According to US Department of Justice figures, you are more than twice as likely to be arrested if you are Black than if you are white. A Black person is five times as likely to be stopped without just cause as a white person.

Another example is PredPol, now known as Geolitica, which relies on processing "historical crime data" to make predictions about where crimes are likely to occur (O'Neil, 2016, p. 85; Geolitica, 2021).

AI can be viewed as a data analysis tool that follows the values of the programmer, but it is the wisdom or mindset behind the programming for AI that makes the difference. The need to develop AI which is representative of a wider range of human values is a challenge. Given that these systems are not usually designed by one individual, but by a design team, there is potential for a more inclusive approach to be taken. However, as Indigenous Peoples are not homogenous and not every community or business at present will include individuals with the right skillset to create systems that are reflective of Indigenous values, a variety of approaches may be needed. It is hoped that models or toolkits could be developed that might assist with this. Communities and businesses contemplating working with Indigenous communities could utilize these models to develop systems that recognize the values of the specific Indigenous community from the outset of a project, similar to the "privacy by design" (Cavoukian, 2011) and "security by design" approaches (Lovejoy, 2020). We hope that this chapter helps to stimulate further discussion of some of these issues.

It is recognized that as control over decision-making is ceded to AI, a power inversion and subsequent erosion of human rights and values becomes a real possibility (Liu and Zawieska, 2017). To ameliorate this erosion of human rights, and correspondingly, to ensure a meaningful recognition of human rights, there is a need to consider the following question. That is, whether the concept of "value alignment"

(Kim and Mejia, 2019), which seeks to ensure that AI incorporates values that are important, is an acceptable solution to ensure that AI is encoded with appropriate human rights values and, if so, whether this could include Indigenous rights (Maitra, 2020, p. 321).

While all Indigenous Peoples are unique, their worldviews do share some similarities. Generally, an Indigenous rights worldview is holistic and relational, underscored by respective cosmologies and a connection with nature. However, AI typically sits outside an Indigenous rights worldview. Consequently, a broader and more pressing issue for Indigenous Peoples is not only the potential erosion of their rights and data within a typical AI paradigm, but also the question: If an Indigenous worldview is to fit within an orthodox AI paradigm, how can this be done? There is a need to consider, firstly, what safeguards are in place, and secondly, how, for instance, are the intellectual property rights attached to an Indigenous worldview protected?

Commentary has noted that AI has "developed in an epistemic echo chamber and the bias in these systems is a feature of white supremacy, a feature that grows out of a whole bunch of interlocked and layered systems" (Lewis, 2020). This heightens the overarching question: how can the right to self-determination for Indigenous Peoples be understood and recognized with an AI realm?

CURRENT VIEW

Indigenous Peoples are affected by the development and the application of AI as community members and subjects of AI endeavors (Walker and Hamilton, 2018). Unsurprisingly, they are asking the question whether AI is the new (r)evolution or the new colonizer for Indigenous Peoples (Whaanga, 2020, p. 35). Some comparisons here can also be made with developments in biotechnology more generally that rely on the resources of Indigenous Peoples and have been dubbed examples of biocolonialism (Whitt, 1998; Indigenous Peoples Council on Biocolonialism, 2006). Beyond value alignment, to date, there has been little meaningful consideration for how Indigenous perspectives and data can be included and protected within an AI realm. Furthermore, Indigenous data is commonly accessed and processed without adequate recognition of Indigenous Peoples' rights (Maitra, 2020, p. 323).

Indigenous rights are whole and indivisible. The UN Declaration on the Rights of Indigenous Peoples provides a framework and understanding for the fundamental rights of Indigenous Peoples (UNDRIP, 2011). This includes, for instance, rights to culture, education, lands, territories, resources and traditional knowledge. The key right is that of self-determination, from which all other rights derive. The operationalization of these rights traditionally sits within an Indigenous worldview.

However, AI does not engage with an Indigenous worldview. Rather, AI reflects the particular values and ideals of the Western scientific worldview (Williams and Shipley, 2021; 2019) and has no normative ability of its own. AI has no conscience; it does not feel joy, guilt or remorse and is wholly unable to care about the overall consequences of its actions or the individual people affected by those actions (Williams and Shipley 2021, p. 44).

If an AI operation is guided by any prescriptive values at all, they are those of its programmers. These programmers are trained and work within the paradigm of the Western scientific worldview, based on a reductionistic ontology of data and a contrived epistemology of algorithms concerned with maximizing the efficiency with which tasks are accomplished and not with the morality of the tasks themselves (Williams and Shipley 2021, p. 44). In this light, AI can only follow prescriptive rules that can be expressed and evaluated in quantitative terms (Williams and Shipley 2021, p. 44).

The technological and philosophical shortcomings of value alignment have triggered concerns on whether this approach can be an appropriate safeguard against human rights breaches and the adequate protection of Indigenous rights, particularly intellectual property and data rights (Maitra, 2020, p. 321).

HOW CAN AI RECOGNIZE INDIGENOUS RIGHTS?

This section considers what is needed for the recognition of Indigenous Peoples' rights in the context of AI systems in terms of procedural and substantive law. It is important to recognize at the outset that it is predominantly men operating within the Western scientific worldview who have been responsible for AI programming. Unsurprisingly, this programming reflects their biases and notions of ethics and wisdom (Weidenbener, 2019). Subsequently, there is no guarantee that engraining values into higher levels of automated AI will be adequate protection for fundamental human or Indigenous rights (Bostrom, 2014, pp. 185-207).

To adequately recognize Indigenous rights within an AI framework, procedural and substantive measures are required. These can then ensure that, if AI develops as its own autonomous entity that will influence our social structures and identities as human beings, Indigenous rights are recognized. Such measures should recognize the importance of personal and community data and data rights. For instance, when data is gathered from Indigenous communities, it should be acknowledged or recognized as Indigenous data that is collected within an Indigenous context.

Procedural rights measures

Procedural rights are clearly provided for in article 18 of the United Nations Declaration on the Rights of Indigenous Peoples (*Declaration on the Rights of Indigenous Peoples*, GA Res 61/295, UN GAOR, 61st sess, UN Doc A/RES/47/1 (2007) 'UNDRIP').

> Indigenous Peoples have the right to **participate in decision-making** in matters which would **affect their rights**, through representatives chosen by themselves in accordance with their own procedures, as well as to maintain and develop their own **indigenous decision-making institutions** (emphasis added).

The key right is that of self-determination articulated in article 3 of the UNDRIP: "Indigenous Peoples have the right of self-determination. By virtue of that right they freely determine their political status and freely pursue their economic, social and cultural development."

Māori data sovereignty espouses the inherent rights and interests that Māori have in relation to the collection, ownership and application of Māori data, including within any AI framework (Kukutai and Taylor, 2016; Te Mana Raraunga Māori Data Sovereignty Network). Any program that seeks to capture Indigenous data using a particular algorithm should recognize this right in not only program design but also in ensuring Indigenous participation in the control and management of any AI program that seeks to include Indigenous data. The Indigenous Navigator is an example of how this can be achieved. The Navigator provides tools for tracking how Indigenous Peoples' rights are recognized. The data collected by the Indigenous Navigator is not the official statistical data, but captures Indigenous Peoples' perceptions and experiences in relation to the framework (IWGIA and ILO, 2021, p 20). This process involves recognizing the right of free, prior and informed consent (IWGIA and ILO, 2021, p. 17). However, it would also be helpful if technical standards and models were developed in collaboration with the Navigator, as this could help developers and programmers to understand what is needed. For example, we could look to the work of the PCI Security Standards Council, which has produced guidelines and

implemented a certification system to improve data security in payment systems globally (PCI Security Standards Council, 2021). It has also developed the Payment Card Industry Data Security Standard, which is a technical standard. We suggest that similar technical standards and certification systems could be set at an international level, which could then be used to develop AI systems and other technological systems used by Indigenous communities.

Substantive rights measures

The UN Declaration on the Rights of Indigenous Peoples articulates the following fundamental rights (emphasis added):

> Indigenous Peoples have the right to **maintain and strengthen** their **distinct** political, legal, economic, social and **cultural institutions**, while retaining their rights to participate fully, if they so choose, in the political, economic, social and cultural life of the State. (article 5)
>
> Indigenous Peoples have the right to revitalize, **use, develop and transmit to future generations their histories, languages, oral traditions, philosophies, writing systems and literatures,** and to designate and retain their own names for communities, places, and persons. (article 13)
>
> States shall consult and cooperate in good faith with the Indigenous Peoples concerned through their own representative institutions in order to obtain their **free, prior and informed consent** before adopting and implementing legislative or administrative measures that may affect them. (article 19)
>
> Indigenous Peoples have the right to maintain, control, protect and develop their cultural **heritage, traditional knowledge and traditional cultural expressions**, as well as the manifestations of their **sciences, technologies** and cultures, including human and genetic resources, seeds, medicines, knowledge of the properties of fauna and flora, oral traditions, literatures, designs, sports and traditional games and visual and performing arts. They also have the right to maintain, control, protect and develop their **intellectual property over such cultural heritage, traditional knowledge, and traditional cultural expressions.** (article 31).

Any AI program that seeks to use or extract Indigenous data through an AI algorithm (or similar) should recognize not only the right of free prior and informed consent, but also rights such as those associated with traditional knowledge, traditional cultural expressions and the manifestations of sciences and technologies. When Indigenous data associated with traditional knowledge is obtained and used without free, prior and informed consent, this is a clear breach of this right. One example of this is when the traditional knowledge that undergirds specific Indigenous medicinal plant remedies is taken and used by commercial companies without the free, prior and informed consent of the Indigenous community. Other examples deal with the direct involvement of Indigenous Peoples in medical research with attempts to patent the cell-line developed from the blood sample of a Guayami woman from Panama and a cell-line developed from a Hagahai donor (WIPO, 2006; IPCB and Harry, 1995).

Undergirding these fundamental rights is the Indigenous worldview. Read together, these rights, driven by the overarching right of self-determination, provide a compelling narrative for meaningful recognition prior to any AI program that seeks to access Indigenous peoples' data or similar, both procedurally and substantively.

AN INDIGENOUS WORLDVIEW

Indigenous Peoples, although from different global regions, share a similar worldview that is derived from nature and cosmology. To capture Indigenous knowledge within an AI framework is challenging, but integrating Indigenous perspectives would allow the building of "a different kind of AI" (Kesserwan, 2018)—one that would reflect and maintain a relational ethic that is reciprocal, as identified in an Indigenous worldview. We provide three brief hypotheses of how indigenous worldviews can add value to AI development drawing from the examples of Navajo, Lakota and Hawaii peoples. We then explore in more detail the Māori people's case.

Navajo

For the Dine (Navajo) peoples, their worldview recognizes and honors their reciprocal responsibilities to the universe that sustains them (Haskie, 2002, citing Griffin-Pierce, 1992). This worldview is captured in the concept of Hózhó, a complex wellness philosophy and belief system comprised of principles to guide thoughts, actions, behaviors and speech (Kahn-John and Koithan, 2015).

The Navajo, like many Indigenous Peoples, ascribe to tenets such as harmony and balance. Their belief systems center on the interrelatedness and connectedness among animate and inanimate beings. However, they also recognize the need for individual wellness and the interdependence of physical, emotional, psychological and spiritual well-being (Haskie, 2002, p. 25, citing Cleary and Peacock, 1998, p. 25). This existing ethical framework of harmony and moral behavior could potentially be placed into an AI framework that seeks, for instance, a more just outcome within our criminal justice system.

Lakota

Similarly, for the Lakota peoples, their worldview assumes that everything in the universe possesses an interior dimension (the soul) and a physical dimension (the body) (Posthumous, 2018). Intrinsic to this is a sense of responsibility to both the animate and the inanimate, the essence of Lakota life (Deloria,1998).

From this starting point, the hypothesis becomes of whether a Lakota AI framework could fill radically different roles, ranging from autonomous weaponry to mass surveillance, while maintaining the relational ontology (Lewis et al., 2019). It is suggested that to capture this, AI's development could be halted intermittently to establish a relational approach (Lewis et al., 2019).

In addition, to overcome the differences in physicality – the opposition between the soul and the body –, it is suggested that the distinguishing features of each AI system, such as its mission, code or creators, be placed at its center, to ensure it is correctly considered in a collective (Lewis et al., 2019). However, it is acknowledged that the programmer would need to be tasked with this centering which presents, in and of itself, a new range of challenges.

Hawaii

In the Hawaiian (kānaka maoli) worldview, the foundational concept of *pono* is an "ethical approach (...) which privileges multiplicities over singularities" to achieve balance and harmony. Within this worldview, *pono* is never reduced to prioritize the individual over a relationship. The well-being of everyone involved within the relationship must be taken into consideration, and self-interest is always secondary (Lewis et al., 2019).

AI is a tool created by humans for human progress. If the Hawaiian (kānaka maoli) worldview is applied to an AI realm, then similarly to both the Lakota and Navajo worldviews, to capture this ethical framework and sharing between AI and humans, the notion of autonomy must be redefined or applied intermittently. Either way, a compromise is required, which is not ideal.

If the previous hypotheses are all considered and mechanisms to align human values in machines are extended—based upon the central tenet of treating all relationships as paramount—we can ameliorate some of the problems with value alignment in AI development. However, without the overarching acknowledgment of self-determination of peoples, it is unsure how effective such values alignment may be.

Rather, a more appropriate approach to move beyond value alignment consists of taking Indigenous epistemologies as a pre-existing value system that requires mutual respect amongst humans and machines.

Māori

The Māori worldview is centered on *tikanga Māori*. *Tikanga* Māori is a complex three-dimensional philosophy that communicates concepts from the inside. The accepted meaning of *tikanga* is "straight and direct," coupled with moral notions of justice and fairness (Benton et al., 2013, 429). However, this can vary according to the people involved and the particular circumstances (Toki, 2018). *Tikanga* Māori is a contextual concept (New Zealand Law Commission, 2001). *Tikanga* Māori is consistently recognized by the courts in New Zealand; it informs New Zealand common law and is acknowledged as an integral strand of the legal system and as an "applicable law" (*Trans-Tasman Resources Ltd v Taranaki-Whanganui Conservation Board* [2021] NZSC 127 at [169]).

This concept is sourced from *Te Ao* Māori, or the Māori World, the world in which Māori lived (Marsden, 1992, p. 117). Māori cosmology and creation stories are intrinsic to *Te Ao* Māori, which establishes the relationships, or *whakapapa,* between the animate and the inanimate, meaning between people, the environment and the spiritual world (Waitangi Tribunal, 2014, p. 20). The interplay between these elements underpins a mechanism similar to that of a social constitution (Toki, 2018). The principle of *whakapapa* is fundamental to *Te Ao* Māori (Toki, 2018). It is a complex network of reality linking all objects (Waitangi Tribunal, 2014, pp. 22–25). As a relational construct, it provides an explanation of how the universe emerged and how the convergence of complementary or balancing pairs created new forms of life (Marsden, 2003). *Whakapapa* has always been central to the identity of an individual. The individual forms part of the collective and, in turn, is linked to others by *whakapapa*. *Whanaungatanga,* in turn, is the "glue" that holds the "parts" together; it is often defined as "the state or circumstances of being a relative, that is, kinship and the rights, responsibilities, and expected modes of behavior that accompany the relationship" (Benton et al., 2013, p. 524). *Whanaungatanga,* as a component of *tikanga,* "embraces *whakapapa* and focuses on relationships" (Mead, 2003, p. 28). It is indispensable to Māori as *whanau* (family) provide for the physical, emotional and spiritual well-being of individuals. Just as "individuals expect to be supported" by the collective, so too does "the collective expect to be supported by the individuals, this is an obligation and a fundamental principle" (Mead, 2003, p. 28).

So, *tikanga* is the structure that gives effect to basic principles or ground rules (Toki, 2018). Concepts such as *mana* and *tapu* assist in the regulation of the relationships or *whakapapa* between people, the environment and the spiritual world (Toki, 2018). The aim of *tikanga* Māori is to achieve balance and harmony, balance within the individual and balance within the community or wider collective. The regulators—*tapu* and *mana*—assist to restore any imbalance, a process that is underpinned by reciprocity, *aroha* (love) and *manaaki* (care). *Aroha* is an emotional concept that is an almost instinctual way of reacting in relationships. It is a central component of Māori ethos and is known to take on the meaning of a healing process (Benton et al., 2013, p. 47). For *kaumatua* and *kuia* the principle of *aroha* is the basis for the giving, sharing and support amongst *whanau.*

The process to restore balance is called *utu. Utu* is the "exercise for the right of compensation" to "return for anything; satisfaction, ransom, reward or response, to make response by way of payment or answer," and it is linked to the concept of *mana* (Benton et al., 2013, p. 46). Often referred to as the principle of reciprocity or equivalence, an important purpose of the process for *utu* is to restore balance and harmony and to maintain relationships or *whanaungatanga* (Mead, 2003, p. 31).

It is difficult to isolate one concept such as *mana* and fold it into an AI framework without related concepts such as *tapu*, *whakapapa* and *utu*. *Mana* on its own loses its essence and in isolation runs the risk of being redefined.

Applying this relational and interconnectedness principle from within an Indigenous worldview has the potential to contribute features to AI development that Western scientific approaches do not. As some Indigenous worldviews do not distinguish between the animate and the inanimate, it could be that this pre-existing relational value system could be adopted within an AI realm as an ethical framework. Equipped with this knowledge, we can begin to construct relational frameworks to protect and empower.

MĀORI CASE STUDY

The following case study considers the practical application of AI to a situation that seeks to achieve well-being for an Indigenous (Māori) community.

Aotea/Great Barrier Island is a remote island located approximately 100 km northeast of central Auckland. It is 285 square kilometers in size and has several small Māori communities. There is no reticulated power system on Aotea. People live off the grid, running their own solar and battery power systems. These systems are supplemented by petrol or diesel generators, natural gas and wood fires, and in virtually all cases, the solar and battery systems do not provide anywhere near the total energy needs for households. There is a high reliance on back-up generators (Aotea Great Barrier Island Local Board Plan, 2020).

The lack of infrastructure on Aotea provides a key opportunity to improve the lives of disadvantaged Māori, and to contribute to New Zealand's efforts to reduce carbon emissions and expand clean energy use.

A proposed solution to this issue is a fully renewable energy-based smart electricity micro-grid system that takes a *tikanga* approach, as well as a phased fractal-structured approach to interconnect micro-grids from the various small Māori communities (Apperley, 2019). The eventual goal of establishing this micro-grid is that energy can then be shared among the Māori households and Māori community on Aotea/Great Barrier Island. The system will also operate solely on Aotea, which should help to optimize the protection of the Aotea community's privacy rights. The solution envisions that through the fractal-structured micro-grid, energy will be both generated and shared within the Māori community at a relatively early stage. The *marae* (customary meeting house) may be the center point within the system where storage and allocation could take place.

This example demonstrates how a holistic approach drawing on existing electricity data, community data, and a *tikanga* or relational approach can be used to solve the practical problem of energy consumption, use and availability. Allowing the community to be in charge of the micro-grid can be viewed as one example of community empowerment.

An important aspect of AI-powered solutions to community challenges is the use of data. In today's world there are tremendous challenges for the privacy rights of all citizens, but especially Indigenous Peoples, together with others who have been marginalized.

The framework to gather data in this case study is conventional – as opposed to indigenous – and employs technology that can be considered as part of the "internet of things (IoT)." IoT "generally refers to scenarios where network connectivity and computing capability extends to objects, sensors and everyday items not normally considered computers, allowing these devices to generate, exchange and consume data with minimal human intervention. There is, however, no single, universal definition" (Internet Society, 2015, p. 5).

It's important to note that data from smart electricity meters can be used to make inferences about a wide variety of things, which could include potentially sensitive data. This includes, for instance,

> behaviors of residents including bathroom activities, cooking, housework, sleep cycles, and meal times can be inferred from seemingly non-sensitive smart meter readings. It has been shown that even the current TV channel and specific audiovisual content displayed on a television can be identified based on the corresponding household's electricity usage profile. (Kröger, 2019)

It is also possible to infer other sensitive data. This includes religious affiliation, which is possible to infer from energy usage patterns, in particular by comparing the data with that of other households on religious festival days (Karwe and Müller, 2015, p. 228; Cleemput, 2018, p. 3; Reimann, 2019). It is also possible to make inferences about health based on use of medical devices (Pham and Månsson, 2019) and to deduce other matters, such as employment status, as smart meters can allow for identification of when individual appliances are used in the home (Anderson, 2016; Murrill, 2012; Greveler et al., 2012).

Providing local communities with the technology to help themselves and keeping it under their control is one way of addressing this privacy challenge. However, the implementation of a community-controlled scheme will still need to take account of privacy and data security issues. If the Indigenous community is involved in developing the system from its earliest stage, then they could also use both approaches discussed earlier, privacy by design together with data sovereignty by design (Data Sovereignty Now, 2020; Nagel and Lycklama, 2021). Following Nagel and Lycklama's approach would mean that an Indigenous community should have complete self-determination over their data. Furthermore, given the amount of information that can be gleaned from smart meter data, it would be wise to have some functionality turned off by default so that the system can comply with New Zealand Information Privacy Principles, in that data is only collected when it is necessary (Principle 1). This is also in accordance with the principle of data minimization set out in article 5 of the GDPR (General Data Protection Regulation, 2016, art. 5).

Given potential security and privacy risks, it would also be advisable to implement a security-by-design approach, which is in line with principle 5 of New Zealand's Information Privacy Principles together with the requirements of article 32 of the GDPR and the "integrity and confidentiality" principle set out in article 5 (General Data Protection Regulation, 2016, art. 32 and 5). To develop further this consideration regarding data protection and its implications for indigenous communities, the following sub-sections will analyze the differences between a conventional approach to data gathering in the smart-grid project and an indigenous one.

Conventional Approach

The smart micro-grid collects and shares data, akin to a digital nervous system. In this context, AI can be viewed as the brain of the system. Applying these two technologies to a grid system powered by solar energy, AI will be used to operate the electrical utility not only in one household but among a cluster of households that have opted into the grid. The key aspects of the proposed smart grid are as follows:

1. It will aim to provide the most efficient use of electricity possible among a cluster of homes, powered by the demand of individual households as well as appliances of the household itself.
2. It will aim to optimize appliances so that high energy use does not occur at the same time among energy-demanding appliances, such as fridges, ovens, washing machines and so on.
3. Surplus solar energy will be directed towards a reservoir-like battery system that is also optimized to allow for efficient energy use once the sun has gone down.
4. Display, notification and remote control will be available through a dashboard on devices such as smartphones, computers, tablets or smart appliances themselves.

AI will analyze the data and complete required and set tasks. This data, once compiled, will provide predictions of future energy use and will construct a form of identity associated with a particular household or with groups of houses. However, it is important to stress that implementing this system should incorporate the approaches of privacy by design (Cavoukian, 2011) and data sovereignty by design (Data Sovereignty Now, 2020).

The conventional approach and framework provide definite advantages and structure to such a proposal. However, given that the case study is within a Māori community, a *tikanga* Māori or Indigenous approach is not only unique but pivotal.

Indigenous Approach

This project will widen the scope of the data compared to what is ordinarily collected, such as metrics associated with economy, to data that better reflects *Te Ao* Māori, or a Māori worldview. For instance, values not normally considered important in a conventional AI realm, such as how *kuia* and *kaumatua* (elders) use energy, will be considered to better allow for their needs. *Tikanga* principles such as *whanaungatanga* that embrace *whakapapa* and focus on relationships strengthen this approach. Associated concepts of *manaaki* and *aroha* are further aligning obligations. This approach is not novel but employed by the Indigenous Navigator program mentioned earlier. The program allows tools for Indigenous communities to monitor how their rights are recognized.

In addition to adopting the aforementioned approach, it would be beneficial to develop other complementary tools that Indigenous communities could use in establishing their own technological systems, which could help enhance self-determination and data sovereignty at the local level. This could also help developers and programmers in understanding what they need to do to recognize Indigenous communities' values.

The overall aim of *tikanga*, a Māori worldview, is balance, which the smart-grid project seeks to achieve—environmental balance and Indigenous well-being. A meaningful application of the ethical *tikanga* framework within an AI realm will contribute to achieving this. If the scope of the AI is informed by Indigenous value sets, with an Indigenous programmer or programmed by someone given Indigenous guidelines to adhere to, the AI will be more consistent with these principles.

Ultimately, we need to reimagine AI as a tool. To better process data associated with Indigenous Peoples' rights and principles, the AI tool must be designed and guided by these principles. This involves discussion and collaboration between diverse stakeholders and developing models that developers, programmers and Indigenous Peoples can draw upon.

CONCLUSION

This chapter engages in a critical dialogue surrounding the value of Indigenous perspectives to AI, emphasizing the acknowledgment of an Indigenous worldview, the rights that underpin it, and an introduction to particular relational frameworks.

Fitting an Indigenous worldview within a non-Indigenous framework such as AI is akin to fitting a round peg into a square hole. To make this fit requires one to adapt to the other. Including an Indigenous worldview comes with the benefit of working within a pre-existing ethical and relational framework. However, such inclusion is not recommended without guarantees of data privacy, security and intellectual property protection, and the appropriate procedural and substantive recognition of the right of self-determination.

If AI were viewed as a tool that Indigenous Peoples could program, and if their fundamental rights were adequately protected, an Indigenous worldview could be achieved within an AI realm. We need to keep in mind, though, that any automated system that has the potential to collect sensitive data does pose privacy risks as well. At present, even automated systems require some human input. There is a need for public discussion and engagement in establishing systems of this kind. The implementation of the system mentioned in the Māori case study should also take account of such risks and implement privacy by design and security by design approaches that also incorporate respect for data sovereignty from the outset.

REFERENCES

Anderson, B., Lin, S., Newing, A., Bahaj, A. and J. P. 2017. Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. Computers, Environment and Urban Systems, 63, pp. 58-67. https://doi.org/10.1016/j.compenvurbsys.2016.06.003

Aotea Great Barrier Island Local Board. 2020. Aotea Great Barrier Island Local Board Plan 2020. https://www.aucklandcouncil.govt.nz/about-auckland-council/how-auckland-council-works/local-boards/all-local-boards/great-barrier-local-board/Documents/aotea-great-barrier-local-board-plan-2020-english.pdf

Apperley, M. 2019. Modelling fractal-structured smart microgrids: Exploring signals and protocols. In M. Negnevitsky and V. Sultan (eds.), *Proceedings of ENERGY 2019, The Ninth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*. Athens, Greece, IARIA, pp. 13–17.

Benton, R., Frame, A., and Meredith, P. (eds.). 2013. *Te Mātāpunenga: A Compendium of References to the Concepts and Institutions of Māori Customary Law, compiled for Te Matahauariki Institute*. Wellington, Victoria University Press.

Bostrom, N. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in *Advanced Artificial Agents. Minds and Machines,* Vol. 22, No. 2, pp. 71-85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. 2014. *Superintelligence: Paths, Dangers and Strategies.* London, UK, Oxford University Press, pp. 1-21, 185-207.

Bradshaw-Martin, H. 2020. Could your self-driving car choose to kill you? BBC Science Focus. November 17, 2020. https://www.sciencefocus.com/future-technology/could-your-self-driving-car-choose-to-kill-you/

Calo, R. 2017. *Artificial Intelligence Policy: A Roadmap.* UC Davis Law Review, Vol. 51, pp. 399-435. DOI: https://doi.org/10.2139/ssrn.3015350

Cath, C. J. N., et al. 2016. Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach. *Oxford Internet Institute.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2906249

Cavoukian, A. 2011. Privacy by Design: The 7 Foundational Principles. IAPP. https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

Cleemput, S. 2018. Secure and privacy-friendly smart electricity metering. Dissertation, KU Leuven, Belgium. https://lirias.kuleuven.be/retrieve/509996

Data Sovereignty Now. 2020. Data Sovereignty Now stimulating the data economy executive summary. https://datasovereigntynow.org (Accessed 12 September 2021); see also https://datasovereigntynow.org/wp-content/uploads/2020/09/stimulating-the-data-economy_executive-summary-1.pdf

d'Alessandro, B., O'Neil, C. and LaGatta, T. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data,* Vol. *5,* No. 2, pp. 120-134.

Deloria, E. C. 1998. *Speaking of Indians.* Lincoln, University of Nebraska Press.

EDRi. 2021. *An EU Artificial Intelligence Act for Fundamental Rights. Civil Society calls on the EU to put fundamental rights first in the AI Act*. November 30, 2021. https://edri.org/our-work/civil-society-calls-on-the-eu-to-put-fundamental-rights-first-in-the-ai-act/

EDRi. 2021a. *An EU Artificial Intelligence Act for Fundamental Rights: a civil society statement*. https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf

Eliot, L. 2020. Decisive Essays on AI and Law. LBE Press Publishing, chapter 4.

European Parliament. (2015/2103(INL)) *EU Parliament Resolution with recommendations to the Commission on Civil Law Rules on Robotics*. https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect

European Parliament Committee on Legal Affairs. 2016. *Civil Law Rules on Robotics* (2015/2103 (INL)). Brussels, Belgium. https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ 2 119/1.

Executive Office of the President National Science and Technology Council Committee on Technology. 2016. *Preparing for the Future of Artificial Intelligence*. Washington D.C. USA. http://www.eenews.net/assets/2016/10/12/document_gw_03.pdf

Fjelland, R. 2020. Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun.* Vol. 7, No. 10, pp. 1-9. https://doi.org/10.1057/s41599-020-0494-4

Franklin, A. 2020. The Fourth Amendment in Your Shower: Naperville, Reasonable Expectations of Privacy, and the Intimate Nature of Electric Smart Meter Data. *NCL Rev.,* Vol. 99, No. 4 pp. 1141-1166. https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=6848&context=nclr

Geolitica. 2021. *Homepage*. https://geolitica.com

Greveler, U., et al. 2012. Multimedia content identification through smart meter power usage profiles. In: *Proceedings of the International Conference on Information and Knowledge Engineering (IKE),* p. 1. WorldComp.

Haskie, M. J. 2002. Preserving Culture: Practicing the Navajo Principles of Hozho Doo K'e. Dissertation, UMI No. 3077247, Ann Arbor, MI, Proquest. https://www.proquest.com/openview/34b9993f90d41bd6a482011691cb023a/1?pq_origsite=gscholar&cbl=18750&diss=y

Heaven, W. D. 2020. *Predictive policing algorithms are racist. They need to be dismantled.* MIT Technology Review. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

House of Commons Science and Technology Committee. 2016. *Robotics and artificial intelligence.* Fifth Report of Session. London, United Kingdom, pp. 5-6, 16-24, 36-8. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

Indigenous Peoples Council on Biocolonialism. http://www.ipcb.org

Indigenous Peoples Council on Biocolonialism and Harry, D. 1995. Patenting of Life and Its Implications For Indigenous Peoples. *Information About Intellectual Property Rights*, No. 7. January 1995. http://www.ipcb.org/publications/briefing_papers/files/patents.html

Indigenous Peoples Council on Biocolonialism. 2006. *The Convention On Biological Diversity's International Regime On Access & Benefit Sharing: Background & Considerations For Indigenous Peoples*. Indigenous Peoples Council on Biocolonialism, Briefing Paper. http://www.ipcb.org/pdf_files/absbriefcop8.pdf

Internet Society. 2015. *The Internet of Things (IoT): An Overview.* (White Paper Oct 2015) p. 5 https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-IoT-Overview-20151221-en.pdf

IWGIA and ILO. 2021. *Indigenous Peoples in a changing world of work – Exploring Indigenous Peoples economic and social rights through the Indigenous Navigator,* (p. 20). https://indigenousnavigator.org/sites/indigenousnavigator.org/files/media/document/Indigenous%20peoples%20in%20a%20changing%20world%20of%20work%20-%20wcms_792208.pdf

Kahn-John, M., and Koithan, M. 2015. Living in Health, Harmony, and Beauty: The Diné (Navajo) Hózhó Wellness Philosophy. *Global Advances in Health and Medicine Journal*, Vol. 4, No. 3, pp. 24-30. https://doi.org/10.7453/gahmj.2015.044

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,* Vol. 62, No., 1, pp. 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Karwe, M., and Müller, G. 2015. *DPIP: A Demand Response Privacy Preserving Interaction Protocol*. In International Conference on Business Information Systems, pp. 224-234. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-26762-3_20

Keall, C. 2021. Why are our defences so shaky? The Waikato DHB ransomware attack in 20 questions. *New Zealand Herald.* 29 May 2021. https://www.nzherald.co.nz/business/why-are-our-defences-so-shaky-the-waikato-dhb-ransomware-attack-in-20-questions/4NDSFQD6FST4LHH3UEIIRLABBY/

Kesserwan, K. 2018. *How Can Indigenous Knowledge Shape Our View Of AI Policy Options.* February 16. https://policyoptions.irpp.org/magazines/february-2018/how-can-indigenous-knowledge-shape-our-view-of-ai/

Kim, T. W. and Mejia, S. 2019. From Artificial Intelligence to Artificial Wisdom: What Socrates Teaches Us. *Computer,* Vol. 52, Issue 10, pp. 70-74. https://doi.org/10.1109/MC.2019.2929723

Kröger, J. 2019. Unexpected Inferences from Sensor Data: A Hidden Privacy Threat in the Internet of Things. In: Strous L., Cerf V. (eds) *Internet of Things. Information Processing in an Increasingly Connected World. IFIPIoT 2018. IFIP Advances in Information and Communication Technology,* Vol. 548. Springer, Cham. https://doi.org/10.1007/978-3-030-15651-0_13

Kukutai, T., Taylor, J. 2016. Indigenous Data Sovereignty Towards an Agenda. *Centre for Aboriginal Economic Policy Research (CAEPR),* Vol. 38. ANU, Australia 10.22459/CAEPR38.11.2016.

Landi, H. 2019. *Lingering Impacts from Wannacry: 40% of healthcare organizations hit by WannaCry in past 6 months*. Fierce Healthcare. 29 May 2019. https://www.fiercehealthcare.com/tech/lingering-impacts-from-wannacry-40-healthcare-organizations-suffered-from-attack-past-6-months

Lewis, J. 2020. *Creating ethical AI from Indigenous perspectives.* University of Alberta. Folio. https://www.ualberta.ca/folio/2020/10/creating-ethical-ai-from-indigenous-perspectives.html

Lewis, J. E. Arista, N., Pechawis, A., and Kite, S. 2019. Making Kin with the Machines. *Journal of Design and Science*. https://doi.org/10.21428/bfafd97b

Liu, H. Y. and Zawieska, K. 2017. From Responsible Robotics Towards a Human Rights Regime Oriented to the Challenges of Robotics and Artificial Intelligence. *Ethics and Information Technology,* 22, pp. 1-13. DOI: https://doi.org/10.1007/s10676-017-9443-3

Lovejoy, K. 2020. *How to manage cyber risk with a Security by Design approach*. EY. 7 February 2020. https://www.ey.com/en_nz/consulting/how-to-manage-cyber-risk-with-a-security-by-design-approach

Maitra, S. 2020. Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings. *AIES '20,* New York, NY, USA. https://dl.acm.org/doi/10.1145/3375627.3375845

Marsden, M. 1992. God, *Man and Universe: A Māori View* in Michael King (ed) *Te Ao Hurihuri: Aspects of Māoritanga.* Auckland, Reed Books, p. 117.

Marsden, M. 2003. *The Natural World and Natural Resources* in Charles Royal (ed.), *The Woven Universe Selected Writings of Rev Maori Marsden*. Masterton, Estate of Rev. Maori Marsden.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. 31 August 1955. *AI Magazine*, Vol. 27, No. 4, p. 12. doi: 10.1609/aimag.v27i4.1904.

McCarthy, J. 2007. *What Is Artificial Intelligence?*. Computer Science Department, Stanford University. http://www-formal.stanford.edu/jmc/whatisai/node1.html

Mead, H. M. 2003. *Tikanga Māori: Living by Māori Values.* Wellington, Huia Publishers, p. 28.

Murrill, B. J., Liu, E. C. and Thompson, R. M. 2012. *Smart Meter Data: Privacy and Cybersecurity, report*. Washington D.C. https://digital.library.unt.edu/ark:/67531/metadc87204/ (Accessed September 10, 2021.) University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu Crediting UNT Libraries Government Documents Department.

Nagel, L. and Lycklama, D. 2021. Design Principles for Data Spaces. Position Paper. Version 1.0. Berlin. DOI: http://doi.org/10.5281/zenodo.5105744

New Zealand Government 2020, Privacy Act, Wellington NZ.

New Zealand Herald. 2021. *Waikato DHB cyber attack: 4200 people's personal details disclosed on dark web*. New Zealand Herald. https://www.nzherald.co.nz/nz/waikato-dhb-cyber-attack-4200-peoples-personal-details-disclosed-on-dark-web/LCSXDX4W3HTZ4FCISHAL4T32IM/

New Zealand Law Commission. 2001. *Māori Custom and Values in New Zealand Law.* Wellington, NZ, NZLC SP9. https://www.lawcom.govt.nz/sites/default/files/projectAvailableFormats/NZLC%20SP9.pdf

O'Neil, C. 2016. *Weapons of Math Destruction*. St Ives, Penguin, pp. 84-100.

PCI Security Standards Council. https://www.pcisecuritystandards.org

Pham, C. T. and Månsson, D. 2019. A study on realistic energy storage systems for the privacy of smart meter readings of residential users. *IEEE Access*, 7, pp. 150262-150270.

Phillips, A. M. and Mian, I. S. 2019. Governance and Assessment of Future Spaces: A Discussion of Some Issues Raised by the Possibilities of Human–Machine Mergers. *Development,* Vol. 62, pp. 66–80 https://doi.org/10.1057/s41301-019-00208-1

Posthumous, D. 2018. *All My Relatives: Exploring Lakota Ontology, Belief, and Ritual.* Lincoln, University of Nebraska Press.

Proposal for a Regulation of The European Parliament and of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts Com/2021/206.

Reimann, R. 2019. *TechDispatch #2: Smart Meters in Smart Homes*. European Data Protection Supervisor. 16 October 2019. https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-2-smart-meters-smart-homes_fr

Russell, S. 2021. The history and future of AI. *Oxford Review of Economic Policy*, Vol. 37, No. 3, pp. 509–520.

Supreme Court of New Zealand. 2021. *Trans-Tasman Resources Ltd v Taranaki-Whanganui Conservation Board*. NZSC 127.

Technology Quarterly. 2020. *Driverless cars show the limits of today's AI*. The Economist. 13 June 2020. https://www.economist.com/technology-quarterly/2020/06/11/driverless-cars-show-the-limits-of-todays-ai

Te Mana Raraunga Māori Data Sovereignty Network, *Te Mana Raraunga – Māori Data Sovereignty Network Charter*. https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5913020d15cf7dde1df34482/1494417935052/Te+Mana+Raraunga+Charter+%28Final+%26+Approved%29.pdf

Toki, V. 2018. *Indigenous Courts, Self Determination and Criminal Justice*. Oxford, Routledge.

United Nations. 2011. United Nations Declaration on the Rights of Indigenous Peoples. https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html

United States Court of Appeals. 2018. *Naperville Smart Meter Awareness v. City of Naperville*, No. 16-3766. 7th Cir.

Waitangi Tribunal. 2014. *He Whakaputanga me te Tiriti The Declaration and the Treaty The Report on Stage 1 of the Te Paparahi o Te Raki Inquiry. Wai 1040.* https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_85648980/Te%20Raki%20W.pdf

Walker, R. S., and Hamilton, M. J. 2018. Machine Learning with Remote Sensing Data to Locate Uncontacted Indigenous Villages in Amazonia. *PeerJ Preprints* 6:e27307v1 https://doi.org/10.7287/peerj.preprints.27307v1

Weidenbener, L. 2019. Many Questions, Fewer Answers at Intersection of AI, Ethics. *Indianapolis Business Journal*, Vol. 40, pp. 20-21. https://www.ibj.com/articles/73969-innovation-issue-many-questions-fewer-answers-at-intersection-of-ai-ethics

Whaanga, H. 2020. AI: A New (R)evolution or the New Coloniser for Indigenous Peoples. In Lewis, J. E., (ed.), *Indigenous Protocol and Artificial Intelligence Position Paper.* Honolulu, Hawai'i: The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR), p. 35. https://spectrum.library.concordia.ca/986506/7/Indigenous_Protocol_and_AI_2020.pdf

Whitt, L. A., 1998. Biocolonialism and the commodification of knowledge. *Science as Culture,* Vol. 7, No. 1, https://doi.org/10.1080/09505439809526490

Williams, D. H., and Shipley, G. P. 2019. Limitations of the Western Scientific Worldview for the Study of Metaphysically Inclusive Peoples. *Open Journal of Philosophy*, Vol. 9, pp. 295-317. doi:10.4236/ojpp.2019.93020

Williams, D. H., and Shipley, G. P. 2021. Enhancing Artificial Intelligence with Indigenous Wisdom. *Open Journal of Philosophy*, Vol. 11, pp. 43-58. doi:10.4236/ojpp.2021.111005

WIPO. 2006. Bioethics and Patent Law: the Cases of Moore and the Hagahai People, (September 2006). http://www.wipo.int/wipo_magazine/en/2006/05/article_0008.html

Yang, L., et al. 2014. Inferring occupancy from opportunistically available sensor data. In *Pervasive Computing and Communications (PerCom),* IEEE, pp. 60–68.

HEADLIGHTS, NOT REAR-VIEW MIRRORS: SEEING, RECOGNIZING, CONSIDERING AND WRITING LGBTI PEOPLE INTO ARTIFICIAL INTELLIGENCE'S LIFECYCLE

JED HORNER

Tech Trust and Safety Practitioner. Formerly, Project Director, Australian Human Rights Centre (AHRC), Faculty of Law, UNSW Sydney.

SDG3 - Good Health and Wellbeing
SDG5 - Gender Equality
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

HEADLIGHTS, NOT REAR-VIEW MIRRORS: SEEING, RECOGNIZING, CONSIDERING AND WRITING LGBTI PEOPLE INTO ARTIFICIAL INTELLIGENCE'S LIFECYCLE

Our future survival is predicated upon our ability to relate within equality.

—Audre Lorde (1980, p. 358)

ABSTRACT

Bias and discrimination, and the harms they can cause, are not new concerns. The potential of artificial intelligence (AI) to exhibit bias and exacerbate discrimination is at the forefront of global policy discussions thanks to the current renewed interest in AI. For historically marginalized population groups, such as lesbian, gay, bisexual, transgender, queer and intersex (LGBTI) people, these concerns are pronounced, and are grounded in experiences of legal and social exclusion. Our response to these concerns, as policymakers, members of affected communities, developers and corporate representatives, should be to craft responsible AI. This should be practical, multi-layered and attuned to the needs of LGBTI people and others. A lifecycle approach, which remains open to adaptation as practices change, is a necessary prerequisite. As a contribution to this approach, I outline some specific practices that could play a more central role in maximizing the benefits and reducing the harms associated with AI for LGBTI people, if adopted at scale. Expanded social audits, the adoption of recognized international standards (and improved engagement within the processes leading to their development) and regular reviews that examine the structural adequacy of existing laws and regulations to manage the impacts of AI, including as they relate to LGBTI people, are but three underpinnings of such an approach, as we seek to write a new chapter for tech policy that is more responsive and inclusive.

INTRODUCTION

The triumvirate of fairness, accountability and transparency (FAT) features prominently in international discussions on responsible artificial intelligence (AI) (Raji et al., 2020; Selbst et al., 2019). As the focus on axes of difference in these discussions has shifted from one largely preoccupied with gender and "race," other axes, including sexual orientation, gender identity and intersex status, are increasingly coming into view. This brings into focus people who are lesbian, gay, bisexual, non-binary, transgender, intersex and queer (LGBTI). These discussions are essential, as are the practical approaches they are engendering through principles, toolkits and frameworks, sometimes at supranational level (European Commission, 2021) but often within technical teams. But these discussions alone are not sufficient (Bowles, 2018; Nolan and Frishling, 2020). Part of the challenge to the adequacy of existing approaches, in isolation, is the enduring pernicious legacy of *de jure* and *de facto* discrimination, which has left an "indelible mark on the lives of LGBTI people, as a diverse population group" (Horner, 2017, p. 99). This requires a broader, more encompassing response, which takes seriously the structural factors so enmeshed in this history of discrimination.

In this chapter, I argue that to craft "responsible AI" in a manner attuned to the needs of LGBTI people, we need to embrace an expanded lifecycle approach which is multi-layered and will continue to morph as good practice evolves and as gaps and omissions become apparent. As a contribution to articulating what this approach might look like, I outline some specific practices for consideration. I acknowledge and pay an intellectual debt to those who have labored to shape, create and elevate the practical antecedents to these approaches. Historically, this includes American corporations, the past Rev. Leon Sullivan (Stewart, 2011), think tanks (such as Carnegie and the Ford Foundation), and foot-soldiers, including my own uncle, who, in working to try and transform the labor conditions of multi-national companies operating during the Apartheid regime in South Africa, had the temerity to call for "evolutionary change" (Horner, 1971). None of these contributions should simply be dismissed as we rush to create new frameworks during times of intensified political struggle, not always cognizant of the old. Indeed, we should not jettison insights, goodwill, disciplinary knowledge, regulatory precedents and aspects of good practice, a point already well made by responsible AI practitioners (Raji et al., 2020; Marchant, 2011; Mittelstadt et al., 2016).

In channeling this expanded lifecycle approach that engages the political and historical, I argue that we should simultaneously focus on a multi-pronged approach. The first involves conducting social audits, so that we understand the broader socio-political context in which AI is developed, used, scaled and evaluated before we conduct impact assessments in a product sense. The second involves adopting (and indeed developing) recognized international standards, and the third concerns evaluating the structural adequacy of legislative and regulatory frameworks. The latter relates to specific axes of difference such as sexual orientation, gender identity and intersex status in areas of public life and in relation to AI. Evaluating these frameworks is not just a responsibility of governments, but of corporations and civil society too, something which has historically been understood (Gray and Karp, 1994; First, 1973). Given recent developments in the European Union (European Commission, 2021), the United States (National Security Commission on Artificial Intelligence, 2021) and other countries, a multi-layered approach that engages industry, government, civil society and members of affected communities is time-critical. In the absence of such an approach, and when deferring instead to narrow technical solutions, it is possible that the adverse experiences of LGBTI people, when it comes to the operation of AI, will pose real and practical barriers to the full realization of their human rights. Given the history of the differential and inequitable production and diffusion of technology, including its impacts, this might also exacerbate divisions between people in the Global North and South, as well as population groups within countries (Benjamin, 2019; Eubanks, 2018). This includes LGBTI people, who are more vulnerable and susceptible to violence and discrimination. This is a place we can intervene in order to prevent and change it for the benefit of humankind more generally, and LGBTI people more specifically. The enduring question is, how?

UNPACKING AI'S CLOSET: SEXUAL ORIENTATION, SOCIAL SORTING AND INTELLIGENT ALGORITHMS

During the first wave of COVID-19, a same-sex couple based in the United States, who were on work visas and impacted by the effects of the virus on the local job market, pivoted to generate an income for themselves by selling their album online (Golding-Young, 2020). They had worked together to make music for eight years. They posted a video on Facebook as a paid advertisement in an attempt to reach their fans. Their post was rejected, with Facebook reportedly labeling it as "adult sexually explicit content." The assumed cause: a picture of their foreheads touching—a picture they had used for years. They tested the system, assuming that if it was a heterosexual couple, similar generic rules (presumably "community standards") would apply to seemingly romantic or intimate images. The experience, they report, was different (Golding-Young, 2020). Not quite two sets of rules, but arguably two interpretations of them. The couple argues:

> We have been heartened recently by the improved representation of LGBTQ people on television, and we are grateful that most people we meet are accepting of our relationship. It's enough to make you think that maybe society has fully accepted that "love is love." Unfortunately, our recent experience with Facebook suggests otherwise. When Facebook's platform refused to allow us to fully express ourselves as both artists and a same-sex couple, it brought back painful memories of discrimination against the LGBTQ community. (Golding-Young, 2020)

There are many ways to think through this specific case—the ambiguities, the inconsistencies and the ethical challenges. One way is to invoke the notion that AI has a "black-box" challenge, where we cannot see the decisions being made behind the scenes, and they are not adequately explained to us as users, consumers or citizens (Pasquale, 2015; Rudin and Radin, 2019). The other might be to refer to "AI's closet." The closet is a widely used metaphor with multiple meanings. It is omnipresent, like AI has become. One dominant interpretation is "a room for privacy and retirement" and another refers to the notion of secrets, or "skeletons in the closet" (Sedgwick, 2008, p. 65). People talk about the closet, are aware of it, yet can never entirely agree on its meaning. This has historically rendered lesbian, gay, bisexual, transgender and sometimes intersex people paradoxically both invisible and visible depending on how and where they are. The closet's contours are shaped by laws, popular media, social attitudes and individual dispositions (Horner, 2017). It is a space of liminality—something in-between full social citizenship and marginalization, imposed by a confluence of factors; something the couple above arguably experienced.

Some might consider the relevance of the closet to be eroding—in certain places, under some conditions, for specific people. But, in a world marked by increasing digital connectivity, perhaps it is simply taking on a new form, remaining an archive of aspiration and trepidation, desire and despair, pleasure and pain, danger and emancipation. To think of the closet in this way, in relation to AI, does something useful. It both accommodates different meanings that attach to AI and invokes the notion of what Bucher (2016, p. 31) terms AI's "algorithmic imaginary." This imaginary encapsulates what LGBTI people and the broader public consider to be possible thanks to AI and engendered by AI—from the real and the now to the imagined.

As the vignette concerning the same-sex couple above highlights, we are already living in the age of intelligent algorithms. These algorithms consume data at an unprecedented scale to sort our social world, shape and cater to our preferences and fears, and enable the monetization of aspects of ourselves that were previously off-limits. As I have outlined, they also police and adjudicate, in increasingly automated ways. In this world—where information has become a currency of its own, a form of capital that challenges how we configure and organize wealth and power—a political economy of information has emerged that "instrumentalizes difference, rather than sameness" (Wark, 2019, p. 31). This information-based political economy, spurred by the product development it is dependent on, has

promised much for LGBTI people, who are historically marginalized, both in a *de jure* and a *de-facto* sense (Horner, 2017, p. 101). The rise of AI, comprehended as a core part of this emergent information economy, has promised free expression, connectivity and, ultimately, a form of emancipation, once constrained by older forms of technology and earlier configurations of political power.

Think of the attempt to make an income, in the vignette above. Alternatively, think of sex. Physical social encounters are increasingly replaced by altogether different sexual encounters enabled by platforms such as Grindr and Scruff for generations both young and old (Albury et al., 2017). These platforms, emblematic of AI for many LGBTI people, organize visual content (such as selfies), leverage locational data (who is within a few hundred feet, or a particular pin on a map) and provide for preferences to be expressed (body type, age, and so on). In doing so, they enable the kinds of digital and physical connections that users might desire, producing pleasure or indeed something else (Albury et al., 2017; Race, 2009). There is a growing body of literature exploring the effects that interactions on these platforms produce for LGBTI people. This includes the way in which platform design and configuration—including the previous ability to classify users according to race, in the case of Grindr—entrench existing social antagonisms and experiences of racial discrimination (Maslen, 2019).

Through these examples, bundled into what might popularly be considered AI, it is easy to glimpse how intelligent algorithms now play a pivotal role in how LGBTI people find sexual partners, form relationships, engage in commercial activities, voice political opinions and much more. This complex relationship to the promise of AI, for LGBTI people, whether experienced through the (more than) social network Facebook or dating apps, might be "cruelly optimistic" (Berlant, 2011). By this, I refer to Berlant's (2011, p. 1) definition of cruel optimism as "a state of attachment to an object (a feeling, relationship, aspiration) in advance of its loss; a strong compelling vision of the 'good life' that ultimately transpires to be an impediment to the realization of such aspirations in the first place." This might be the future conjured by AI (i.e., sex, another income, improved consumer experience, or a relationship) or an escape from current physical homophobia, only to see it replaced by a more nefarious online form you cannot turn off when you close your front door.

These are not theoretical considerations. Asking these questions helps reorient our focus from tech as merely instrumental, a *fait accompli*, towards a more critical engagement where we can identify and examine our relationship with AI as LGBTI people. From this baseline, we can ask: What are we trying to achieve with a particular form of technology? What assumptions about its own capacities are we making, relative to our desires, hopes and fears? How can we manage some of the downsides we experience, as a diverse population group, that it might amplify? This enables us, with others, to intervene in a considered way to re-shape the trajectory of tech. But to do so effectively requires an articulation of what AI is—one that traverses the scientific, the popular and the political.

DEFINING AI: LIFECYCLES, NOT JUST ALGORITHMS

While there is a fixation with algorithms as a proxy for AI and its social, political, environmental and economic impacts, AI is ultimately about more than algorithms (Metcalf and Crawford, 2016; Mittelstadt et al., 2016; Raji et al., 2020; Selbst et al., 2019). Mittelstadt et al. (2016, p. 2) have argued that "it makes little sense to consider the ethics of algorithms independent of how they are implemented and executed in computer programs, software and information systems." There is, instead, an AI life cycle involving raw materials, data inputs, algorithms, optimization, auditing processes and business planning, all of which are shaped by human decision-making. Together, these form the object we call AI. Some have adopted the term "cyber-physical" or "cybernetic" to describe AI, which captures aspects of this way of seeing AI (Bell et al., 2021). To explain and help people to visualize such a value-chain driven approach, Bratton (2015) has introduced the powerful idea of the stack, denoting the

components of what we understand as modern computing, extending to AI. Bratton's stack encompasses the earth layer (for example, the components of technical devices) through to the interface and the user. This works as a heuristic to comprehend what AI is, quite literally and materially, including from a political economy perspective. You might be able to imagine a smart home device or a dating or hook-up app through this lens, for instance, thinking through not just AI but the adjacent power needed to enable it (for example, the operating system, chips and battery to power your smartphone). This understanding of AI as a lifecycle has implications for how we define it, shifting our focus from lines of code in isolation to business processes, stakeholder engagement and the decisions we take concerning how technology is developed, adopted and scaled. This includes the way in which its development and impacts are managed, including for LGBTI people. Such decisions always involve moments of exclusion where we choose to privilege a factor, an attribute or a market segment, or to trade off forms of social harm and commercial imperative (Mouffe, 2005). A lifecycle approach, by definition, then, is about responsibility, corporate or otherwise, as we make these decisions.

This approach to defining AI encourages and can enable accountability precisely because it shifts the discussion from inevitability or the notion of "unruly technology" to the idea that we, as humans, through "the political", might exert influence over the direction of technology, shaping and molding how it is developed, adopted and scaled (Mouffe, 2005). This includes its potential impacts on specific affected communities. A decision, therefore, to collect data on sexual orientation on a large scale in a way that enables political advertising, or to design a product such as a hook-up or dating app based on selective, and perhaps responsible, use of similar such data, is not simply a knee-jerk reaction. Rather, it is a considered market response to an identified need, segment and opportunity informed (or not) by legal, ethical and political considerations. The AI that we see, use and relate to, therefore, is not accidental or a product of some technological dream time; it is constructed by humans with their bias, power and privilege (Benjamin, 2019). Indeed, even the ability to use specific subject positions such as "gay," or "bisexual," or the way a product team might define "race" (i.e., black, white, Asian) or gender (i.e., male, female, non-binary, x, etc.) is contingent and shifts over time (Laclau, 2005).

In short, there is no product—no AI artefact—without a lifecycle. Yet I maintain that the cycle itself is longer and more complex than designers or engineers might have imagined, encompassing wider socio-political factors. In such a differentiated lifecycle approach, decisions about design or subsequent adoption are not neutral. Neither are potentially adverse decisions to be seen as belonging exclusively to technical teams (for example, engineering), according to this viewpoint. This approach recognizes the importance of product, brand and, more broadly, entity-wide risk appetite and corresponding accountability as part of a true lifecycle.

One definition which approximates the approach I am referring to was contained in early iterations of ISO/IEC Draft International Standard 22989 (within Joint Technical Committee 1, Sub-Committee 42 on AI). This articulated AI as "a set of methods or automated entities that together build, optimize and apply a model so that the system can, for a given set of predefined tasks, compute predictions, recommendations, or decisions. AI systems are designed to operate with varying levels of automation". (ISO/IEC, 2020). What this definition encouragingly encompassed was a focus on a lifecycle—a range of interconnected parts or functions (such as algorithmic optimization) involving automation to varying degrees. What it does not do is reduce AI to all instances of automation, nor does it imply that activities such as data collection, even in a more automated sense, are necessarily instances of AI in isolation.

Whilst the definition that I am advocating for might not be contentious in business settings, given the widely understood practices of corporate strategy, risk management (PRISMA, 2020; Bemthuis et al., 2020) and even, more recently, political risk management (Rice and Zegart, 2018), it is important to foreground. This is precisely because of the divergent ways in which scholars, practitioners, public institutions and businesses view AI, understand its potential impact and assign responsibility—and more often than not, blame—for its shortcomings within the AI lifecycle (Australian Human

Rights Commission, 2021; Eubanks, 2018). The approach to AI that I have outlined, that of the lifecycle, embraces a multi-stakeholder view in terms of who develops AI, uses AI, assesses the impacts of AI, regulates AI and accounts for the growth of AI, both within countries and internationally.

ENACTING RESPONSIBLE AI: STEPPING INTO THE LIFECYCLE

Marchant (2011, p. 200) argues that, in response to technological developments that might pose harms, some of which I have described, we can either "(1) slow the pace of development or (2) improve our capacity to adapt." The first option would seem problematic, because it impacts on the diffusion of technology and the right to access the benefits of science that comes with the development and adoption of AI, from more complex neural networks to adaptations of machine learning. It might also inadvertently entrench the dominance of AI in one geographic area, posing socio-economic and security challenges (National Security Commission on Artificial Intelligence, 2021). The second response seems intuitive and involves developing, scaling and embedding approaches that are technically sound, attuned to socio-political contexts and able to meet fundamental human rights norms. Here, we can think of laws, principles, company codes, audits and a wide array of practices in existence (Gray and Karp, 1994; Raji et al., 2020; Whittlestone et al., 2019). One positive and encouraging recent example is the development of comprehensive resources through ABOUT ML (Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles) under the auspices of the Partnership on AI (Partnership on AI, 2021). Expanding this focus and building on this momentum, next I outline specific steps that can be taken together or as discrete contributions to enable a more expansive lifecycle approach to managing the impacts of AI for LGBTI people. The iterative and interconnected nature of this approach is outlined in Figure 1.

| FIGURE 1 |

An Integrated, iterative model of an expanded lifecycle approach to AI.

Social audits

Inputs or baseline reference points

- International human rights instruments
- Relevant "ground" or "attribute" (e.g. sexual orientation, gender identity, etc.)
- Universal Periodic Review reports by the UN Human Rights Council
- UN Sustainable Development Goals (SDGs)
- Relevant national laws and policies
- Data on social disparities (routinely reported and incidentally reported)
- Voices of affected communities (e.g. qualitative literature and inputs from participatory processes)

Entity-specific practices

Adoption of international standards from recognized Standards Development Organizations (SDOs)

- Audits and subsequent certification
- Principles
- Design rules
- Data sheets and other artefacts
- Recourse, redress, and complaints-handling procedures
- System documentation management and review

Ongoing monitoring, review, and revision

Stocktakes of legislation, regulation, and binding policy documents

- Amendments to legislation or regulation
- Changes in strategic and operational policies
- Report submission to multi-stakeholder processes (e.g. national human rights institutions and the UN Human Rights Council)
- Influence on practices and processes (e.g. drive change in evidence base and literature)

Define the issues first: Conduct more expansive social audits in key areas to identify the issues and understand their magnitude

> Despite being the past participle of the Latin verb "dare," i.e., "to give," data instead are always produced by people, out of what they observe, fail to see, or suppress in the world in which they live. A corollary, in the case of people, is that a hallmark of privilege is who and what one can afford to ignore. (Krieger, 2021, p. 2)

It is often hard to act on what we cannot and will not see, and on what we are unable to agree on or determine how to measure. As Krieger (2021, p. 2) notes, this is often the consequence of conscious decisions—acts of commission and omission. Understanding the focus of responsible AI for LGBTI people is difficult in the absence of clear agreement over what the fundamental areas of obligation, need or concern are at a national level, let alone a regional or international one. We might, for example, enumerate specific human rights but not others, or rely on the limited experiences of members of a team to identify known risks, but not those that have a high likelihood of materializing, due to our limited frames of reference. We might pay attention to specific disparities for LGBTI people in health or employment but not in other areas. So too, we might neglect persistent structural barriers (for example, laws that disadvantage LGBTI people), privileging an analysis of social practices over other levels of analysis. Part of this might be attributable to the composition of teams, where disciplines such as anthropology or sociology might be privileged, much like computer science has been. Another part of this might be attributable to the artifacts we might develop in this area, always reflecting an accepted way of doing things.

One model to enable more comprehensive analysis and awareness of the baseline for LGBTI people, and inform the practice of responsible AI, is a social audit. Although not new or necessarily unique (Nolan and Frishling, 2020), audits provide a comprehensive process to analyze available data, identify gaps in data availability and quality, and develop new explanatory models. They are already used within AI product development and can be sensitized to broader stakeholder perspectives, Google's SMACTR (Scoping, Mapping, Artifact Collection, Testing, and Reflection) model being one example (Raji et al., 2020). However, the ideal social audit I describe is structural in nature and aims to provide the baseline material required for impact assessments themselves. Its gaze is broader; it focuses on levels and appreciates intensity of exposure and pathways that maintain disadvantage for LGBTI people. The goal of such social audits is to inform the assumptions about risk, vulnerability and susceptibility that policymakers, product managers, designers and engineers might have. It does so based on bringing together data, insights, human rights norms and published literature, and it is shaped by the voices of affected communities.

Certain instructive international models might provide some illustrations of what an ideal social audit approach looks like. The United Kingdom, for example, has pioneered a Race Disparities Audit with a publicly available dataset in key areas of analysis. The Race Disparities Audit process in the UK has also given rise to a realization of the shortfalls of data collection standards in many critical areas, which is undoubtedly positive in ensuring consistent measurement to inform specific interventions and, in some cases, mitigation measures. This process has been reviewed by the UK House of Commons Women and Equalities Committee (2018) which endorsed it as well as calling for a clear action plan. The UN Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance has similarly remarked of the Race Disparity Audit, "The RDA and its database are worthy of emulation by governments all over the world. I strongly commend this initiative" (United Nations High Commissioner for Human Rights, 2019).

For LGBTI people, such audits can be undertaken by companies, governments, or civil society organizations, with specific reference to the nature of the known harms LGBTI people face. Methodologically, they should go further, though, and purposely engage wider views and different

disciplinary perspectives as well as taking proactive measures to ensure that affected communities are represented not just in the design of such processes but in analysis too. Ideally, these audits should also focus on three key areas of concern as a baseline:

1. **The adequacy of existing data,** including data on social disparities. Questions should examine how reliable the data is, its face validity, how it is collected and whether it actually captures the magnitude of harms—for example, intensity and duration of exposure.

2. **The explanatory models used to construct, analyze and describe data:** Does the framework of analysis assign blame for outcomes based on assumptions about people of specific gender identities or sexual orientations or their physical sex characteristics?

3. **Structural factors (such as laws) that impact key domains:** How does this map to exposure data? Are the pathways from these legislative and regulatory impacts clear, specified and explained? Were these laws recently repealed, with a lag in social attitudes, or entrenched for decades prior?

Finally, social audits should be undertaken with the help of teams with expertise in risk management, social science research, engineering, computer sciences and other core disciplines such as law. Given the depth of analysis within fields such as public health and disciplines such as epidemiology, which measure literal social harms, the framework of analysis for such social audits could benefit from using these disciplinary concepts, tools and methodologies (Krieger, 2021). As an example, in assessing the impacts of specific AI-related harms, or indeed those incurred in its development (Benjamin, 2019), measures might adopt a consideration not just of susceptibility and likely exposure; instead, they might also explore the extent to which exposure to given harms might contribute to maintaining or entrenching disparities, "requiring careful development of a priori hypotheses about timing and intensity of exposure in relation to the outcome(s) under study" (Krieger, 2000, p. 57). Whilst outputs from such a social audit should be detailed and exhibit methodological rigor, they should also exhibit clear analysis and illustrate data for a non-specialist audience too. After all, the intent of such a social audit is to change practices, not merely modes of analysis. If we cannot define the persistent and enduring social disparities affecting LGBTI people and understand the pathways that maintain these forms of disadvantage, we cannot have an informed discussion to address these issues, including as they are challenged, ossified or amplified through AI. Social audits provide a necessary first step to widen and deepen our understanding, engagement and ultimately analyses for responsible AI.

Develop and adopt international standards to guide AI development, deployment and evaluation at scale

Given our different lived experiences, national legal contexts, and the geographic and socio-economically shaped diffusion of technology, LGBTI people need to think about the most efficient and impactful opportunities to influence responsible AI globally. Advocates for LGBTI rights have done this successfully in the past by leveraging or piggybacking on civil rights and nascent social justice movements to achieve a fuller realization of our fundamental human rights (Lixinski, 2020). This has included through domestic law changes and diversity and inclusion initiatives, for example.

In an increasingly multi-polar world, a driver to do this might come from regulatory moves in Europe (European Commission, 2021), the United States or anywhere else there is concentrated market and regulatory power, where institutions and nation-states might shift their gaze to the risks and specific social harms for LGBTI people and others arising from AI's use. Very often, these moves require interpretation and more practical mechanisms to enable not just compliance but ongoing monitoring and review. This presents opportunities, many of which might involve practical knowledge exchange and collaboration between civil society, researchers, government agencies and tech companies themselves.

But this energy and intellect, and these efforts, need to be better channeled and directed so that artifacts (i.e., standards, technical specifications) are produced that are capable of widespread use across entities of all sizes. This will prevent good practice from remaining diffuse or indeed even proprietary.

One such route—with precedent and an existing infrastructure that accommodates a multi-stakeholder approach and embraces multilateralism—is international standard setting. Through recognized standards development organizations (SDOs), such as the International Organization for Standardization (ISO) or International Electrotechnical Commission (IEC), many responsible tech companies, such as Microsoft, Google and IBM, already participate in shaping information security standards, information technology governance standards and, more recently, standards for artificial intelligence. What these processes can deliver through their governance model is a thorough and rigorous process that maintains the language, structure and methodological approach that are vital for artifacts in the context of commercial contracts, regulatory call-up (when appropriate) and broader voluntary use within industry (Cihon, 2019).

The development of international standards is not necessarily at variance with international regulatory moves in AI that matter for LGBTI people. Instead, standards development here is often complementary, even necessary. For example, privacy will remain a critical right to be protected for LGBTI people and every human being. International standards such as ISO/IEC 27701, mapped to the requirements of the European General Data Protection Regulation (GDPR) and other select national privacy laws, are already providing a framework for enterprises of all sizes that want to implement a more comprehensive approach to privacy information management in practice (Standards Australia, 2020, p. 28).

As the pace of AI standards development increases, there are clear opportunities for LGBTI civil society organizations and individual contributors to ensure that emerging standards in bias, risk management and other related areas have due regard to specific social harms that affect LGBTI people. In some instances, this can be achieved through mirroring and referencing effective practices and methods in these standards. In other instances, where practices might be more nascent and further study is required before standardization takes place, it might be through developing specific technical reports. One intermediary point might be what is termed pre-standardization work, where consortia work to identify, define and outline a standards-based response to an issue before commencing work through a national standards body or international standards development organization to develop a standard based on a comprehensive draft. This not only increases the chance that a new work item proposal (NWIP) leading to a standard will be approved, but arguably improves the drafting rigor and opportunities for multi-stakeholder engagement.

The specific challenge for LGBTI people is to ensure that as other civil society organizations partake in these standards setting processes (for example those focused on consumers broadly), from the national level to the international level, we create the same level of momentum and engagement. We might ask the following constructive questions at the national level: Which organization, from a civil society perspective, is ensuring the voices of LGBTI are reflected in the work of national standards bodies concerning AI? And in the shaping of the standards-setting process, is material being used that adequately reflects existing research on the benefits and harms associated with AI for LGBTI people as international standards are being developed?

We need more LGBTI perspectives and voices in the standards development process. The challenge, therefore, is threefold. First, we must shape the development of standards using our lived experience. Second, we must codify that in a way that is sensitized but generalizable—we all have human rights (see Figure 2). Finally, we need to encourage the subsequent uptake of recognized international standards to ensure there is a marked shift in practices, norms and market behavior as a result.

| FIGURE 2 |

Shaping international standards: a model of organization and participation for LGBTI stakeholders.

Assess the structural adequacy of laws and regulations

Finally, governments should proactively, and in a consolidated way, conduct stock-takes of laws, regulations and even binding policy directives that impact LGBTI people's enjoyment of their human rights in relation to AI. The express purposes of this should be to identify empirical developments in AI as they impact LGBTI people and to assess the related adequacy of existing laws and their consistency with human rights norms. These reviews should not be limited to AI-specific or technology-centric legislation and regulation, but instead focus on the domains or areas of public life in which impacts are being felt and where there are known or likely deltas or gaps. For example, they might focus on gaps between stated aims, such as full legal equality, and the state of existing law in relation to anti-discrimination. Here, religious exemptions might impact the way in which companies or organizations can lawfully hire and fire (Horner, 2017). These same laws might subsequently shape the way that algorithms for employment screening are lawfully allowed to operate in spite of disquiet and a rising consumer base that disavows such specific uses of AI. Again, the focus is not solely on the downsides of AI alone, but on material risks and consequent harms for LGBTI people, and indeed members of other affected communities, across domains of public life. This would necessarily drill down to explore questions such as whether the digital amplification of discrimination and similar social harms might exacerbate existing structural inequities and vulnerabilities at a population level (Horner, 2017; Bourgois et al., 2017; Metzl and Hansel, 2014).

The benefits of this approach are that, if conducted with integrity, it centers those who are affected by AI and engages the disproportionate and discriminatory impact on members of these population groups (Raji et al., 2020). It also provides for a full, considered social and legal analysis of what the precise examples, challenges and potential solutions are, including through law reform. In the age of "regulatory capitalism" where "interventions can begin from anywhere from within its networks and then, through diffusion mechanisms, can quite rapidly globalise" (Drahos, 2017, p. 776), this approach provides for dialogue, analysis and transparency. It equips governments, civil society and tech companies themselves with finer-grained understandings that are contextual to inform their policy responses globally.

In Australia, the Human Rights Commission undertook a similar approach between 2018 and 2021, providing a final report in early 2021 that outlined a series of measures—some voluntary, some structural—to address the challenges they identified in relation to AI (Australian Human Rights Commission, 2021). This broader stocktake approach can accommodate LGBTI people on the basis of protected attributes and as an affected community of concern, but more firmly grounding this approach by focusing on diverse population groups such as LGBTI people can provide instructive insights for future stocktake-related initiatives.

These stocktakes need not be solely undertaken by national, state or provincial governments or by their departments or agencies. Professional associations, quasi-regulatory bodies and others can play a pivotal role, and one that exhibits leadership. The New Zealand Law Foundation, in collaboration with the University of Otago, for example, has undertaken a broad consultive process resulting in a detailed analysis of the human rights impacts of AI in New Zealand and with reference to the country's developed human rights architecture (Gavaghan et al., 2019). This has implications for how New Zealand manages the impacts of AI in the future, including for LGBTI people, and with reference to domains such as criminal justice (predictive policing) and employment law.

CONCLUSION

Tackling AI's gaps for LGBTI people, which reflect the prejudices, aspirations, fears and oversights of the humans who build AI systems and train them by design, necessitates the realization that AI is a lifecycle. Effectively identifying, challenging and addressing these gaps and acts of commission and omission, which can manifest in bias and material discrimination, requires an abundance of creativity and a willingness to leverage material insights. This includes the diverse lived experiences of members of affected LGBTI communities, as well as concepts and methodologies from disciplines, including epidemiology, that are proximate to the study of human harms such as discrimination—experiences which remain all too real for many LGBTI people (Horner, 2017). To date, these have arguably not been as widely leveraged as they should be within the nascent movement for responsible AI.

Through more structured, considered and concrete activity that builds on the work already underway (Raji et al., 2020; Australian Human Rights Commission, 2021; National Security Commission on Artificial Intelligence, 2021), AI's impacts and trajectory in relation to LGBTI people can be influenced and (re)shaped so that it is more reflective of fundamental human rights. As Bowles (2018, p. 197) argues, "rather than tackle the culture problem head on, it's better to focus on concrete change. Pushing beyond user and business needs to consider society as a stakeholder encourages technologists to appreciate their place in a wider community, governed by a social contract".

This requires what I term an "expanded lifecycle" approach. As part of a broader suite of measures, this approach entails: social audits, the adoption of recognized international standards, and regular consolidated stocktakes of legislation and regulation. Each of these tactics should expressly incorporate the perspectives of LGBTI people, as members of an affected community that has historically been marginalized. This will require ongoing, coordinated and collective efforts that materialize in specific measures and practices—efforts that members of LGBTI communities are experienced in creating, navigating and sustaining. Drawing on Treichler 's (1999, p. 1) inspirational words, who was writing in the midst of another public health crisis that challenged LGBTI people, and specifically gay men:

> it is the careful examination of language and culture that enables us, as members of intersecting social constellations, to think carefully about ideas in the midst of a crisis: to use our intelligence and critical faculties to consider theoretical problems, develop policy, and articulate long term social needs, even as we acknowledge the urgency of the [....] crisis and try to satisfy its relentless demands for immediate action.

These are capacities we collectively have and must harness to ensure that responsible AI, as an evolving set of concrete and real practices, becomes the answer to AI's existing gaps and acts of commission and omission as they manifest for LGBTI people. If we are to truly relate within equality, surely this is the baseline that future generations, and even the current ones, expect.

REFERENCES

Albury, K., Burgess, J., Light, B., Race, K. and Wilken, R. 2017. Data cultures of mobile dating and hook-up apps: Emerging issues for critical social science research. *Big Data & Society,* Vol. 4, No. 2, pp. 1–11.

Australian Human Rights Commission. 2021. *Human Rights and Technology: Final Report*. Sydney, Australian Human Rights Commission.

Bell, G., Gould, M., Martin, B., McLennan, A. and O'Brien, E. 2021. Do more data equal more truth? Toward a cybernetic approach to data. *Australian Journal of Social Issues,* Vol. 56, No. 2, pp. 213–222.

Bemthuis, R., Iacob, M.-E. and Havinga, P. 2020. A design of the resilient enterprise: A reference architecture for emergent behaviors control. *Sensors*, Vol. 20, No. 22, p. 667.

Benjamin, R. 2019. *Race After Technology*. London, Polity Press.

Berlant, L. 2011. *Cruel Optimism.* Durham, N.C., Duke University Press.

Bourgois, P., Holmes, S., Sue, K. and Quesada, J. 2017. Structural vulnerability: Operationalizing the concept to address health disparities in clinical care. *Academic Medicine,* Vol. 92, No. 3, pp. 299–307.

Bowles, C. 2018. *Future Ethics.* East Sussex: New Next Press.

Bratton, B. 2015. *The Stack: On Software and Sovereignty.* Cambridge, M.A.: MIT Press.

Bucher, T. 2016. The algorithmic imaginary: Exploring the ordinary effects of Facebook algorithms. *Information, Communication & Society,* Vol. 20, No.1, pp. 30–44.

Cihon, P. 2019. *Standards for AI Governance: International Standards to Enable Global Co-ordination in AI Research & Development*. Oxford, Future of Humanity Institute (University of Oxford).

Drahos, P. 2017. Regulating capitalism's processes of destruction. Drahos, P. (ed.) *Regulatory Theory: Foundations and Applications.* Canberra, The Australian National University, pp. 761–783.

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

European Commission. 2021. *Impact Assessment of the Regulation on Artificial Intelligence*. Brussels: European Commission. https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence

First, R. 1973. The South African Connection: From Polaroid to Oppenheimer. *Issues: A Journal of Opinion,* Vol. 3, No. 2, pp. 2–6.

Gavaghan, C., Knott, A., Maclaurin, J., Zerilli, J. and Liddicoat, J. 2019. *Government Use of Artificial Intelligence in New Zealand*. Wellington, New Zealand Law Foundation and University of Otago.

Golding-Young, S. 2020. *Facebook's Discrimination Against the LGBT Community*. ACLU. https://www.aclu.org/news/lgbtq-rights/facebooks-discrimination-against-the-lgbt-community/

Gray, K. R. and Karp, R., E. 1994. Corporate social responsibility: The Sullivan Principles and South Africa. *Visions in Leisure and Business,* Vol. 12, No. 4, Article 2.

Horner, D. B. 1971. *United States Corporate Investment and Social Change in South Africa*. Johannesburg, South African Institute of Race Relations.

Horner, J. 2017. Expanding the gaze: LGBTI people, discrimination and disadvantage in Australia. Durbach, A., Edgeworth, B. and Sentas, V. (eds.) *Law and Poverty in Australia: 40 Years After the Poverty Commission.* Sydney, Federation Press, pp. 92–102.

House of Commons Women and Equalities Committee. 2018. *Race Disparity Audit: Third Report of Session 2017–19*. London, House of Commons.

ISO/IEC. 2020. Draft International Standard 22989. Geneva, ISO/IEC.

Kreiger, N. 2000. Discrimination and health. Berkman, L. F. and Kawachi, I. (eds.) *Social Epidemiology.* Oxford, Oxford University Press, pp. 36–75.

Kreiger, N. 2021. Structural racism, health inequities, and the two-edged sword of data: Structural problems require structural solutions. *Frontiers in Public Health,* Vol. 9.

Laclau, E. 2005. Populism: What's in a name? Panizza, F. (ed.) *Populism and the Mirror of Democracy.* London, Verso, pp. 32–49.

Lixinski, L. 2020. Rights litigation piggybacking: Legal mobilization strategies in LGBTIQ international human rights jurisprudence. *Florida Journal of International Law,* Vol. 31**,** No. 3, pp. 273–314.

Lorde, A. 1980. Age, race, class, and sex: Women redefining difference. Rothenberg, P. S. (ed.) *Racism and Sexism: An Integrated Study.* New York, St. Martin's Press, pp. 352–359.

Marchant, G. E. 2011. Addressing the pacing problem. Marchant, G. E. A., Braden, R. Herkert, J. R. (ed.) *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight.* Dordrecht, Springer, pp. 199–205.

Maslen, A. T. 2019. *White for White: An Exploration of Gay Racism on the World's Most Popular Platform for Gay and Bisexual Men.* London, London School of Economics and Political Science.

Metcalf, J. and Crawford, K. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society,* Vol. 3., No. 1.

Metzl, J. M. and Hansel, H. 2014. Structural competency: Theorizing a new medical engagement with stigma and inequality. *Social Science & Medicine,* Vol. 103, pp. 126–133.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, Vol 3, No. 2, pp. 1–26.

Mouffe, C. 2005. *On the Political.* London, Routledge.

National Security Commission on Artificial Intelligence (NSCAI). 2021. *Final Report*. Washington, D.C., NSCAI.

Nolan, J., and Frishling, N. 2020. Human rights due diligence and the (over) reliance on social auditing in supply chains. Deva, S. and Birchall, D. (eds.), *Research Handbook on Human Rights and Business.* Cheltenham, Edward Elgar Publishing, pp. 108–129.

Partnership on AI. 2021. *ABOUT ML Resources Library*. https://partnershiponai.org/about-ml-resources-library/

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, M.A., Harvard University Press.

PRISMA. 2020. *Guidelines to Innovate Responsibly: The PRISMA Roadmap to Integrate Responsible Research and Innovation (RRI) in Industrial Strategies*. Rome: Italian Association for Industrial Research.

Race, K. 2009. *Pleasure Consuming Medicine: The Queer Politics of Drugs.* Durham, N.C., Duke University Press.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson B., Smith-Loud, J., Theron, D. and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, Barcelona. New York, ACM.

Rice, C. and Zegart, A. 2018. *Political Risk: How Businesses and Organisations Can Anticipate Global Insecurity.* London, Weidenfeld & Nicolson.

Rudin, C. and Radin, J. 2019. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition, *Harvard Data Science Review,* Vol. 1, No. 2.

Sedgwick, E. K. 2008. *Epistemology of the Closet.* Berkeley, University of California Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. and Vertesi, J. Fairness and abstraction in sociotechnical systems. FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, Atlanta. New York, ACM.

Standards Australia. 2020. *An Artificial Intelligence Standards Roadmap: Making Australia's Voice Heard*. Sydney, Standards Australia.

Stewart, J. B. 2011. Amandla! The Sullivan Principles and the battle to end apartheid in South Africa, 1975–87. *Journal of African American History,* Vol. 96, No. 1, pp. 62–89.

Treichler, P. A. 1999. *How to Have Theory in an Epidemic: Cultural Chronicles of AIDS.* Durham, N.C., Duke University Press.

United Nations High Commissioner for Human Rights. 2019. *End of Mission Statement of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance at the Conclusion of Her Mission to the United Kingdom of Great Britain and Northern Ireland.* https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=23073&LangID=E

Wark, M. 2019. *Capital is Dead: Is This Something Worse?* London, Verso.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. and Cave, S. 2019. *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research*. London, Nuffield Foundation.

INCLUSIVE INNOVATION IN ARTIFICIAL INTELLIGENCE: FROM FRAGMENTATION TO WHOLENESS

ÉLIANE UBALIJORO

Executive Director for Sustainability in the Digital Age and Canada Hub director for Future Earth, Montreal, Canada.

GUYLAINE POISSON

Associate Professor, Information and Computer Sciences Department, University of Hawaii at Manoa, Honolulu, Hawaii, USA.

NAHLA CURRAN

Undergraduate student, Department of Economics, Philosophy and Political Science, University of British Columbia, Okanagan, Kelowna, Canada.

KYUNGIM BAEK

Associate Professor, Information and Computer Sciences Department, University of Hawaii at Manoa, Honolulu, Hawaii, USA.

NILUFAR SABET-KASSOUF

Strategic Programs Manager for Sustainability in the Digital Age and Future Earth, Montreal, Canada.

MÉLISANDE TENG

PhD student, Mila, University of Montreal, and LEADS Intern with Future Earth, Montreal, Canada.

SDG5 - Gender Equality
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

INCLUSIVE INNOVATION IN ARTIFICIAL INTELLIGENCE: FROM FRAGMENTATION TO WHOLENESS

ABSTRACT

Artificial intelligence is shaping the future of humanity. But what happens when only a fraction of society is at the table that is defining that future? Trends in innovation are centered around profit and growth, with inclusivity loosely at their edges. The result is a fragmented digital age that increases biases, disparities and inequalities and that demands societal trade-offs of global well-being and technological trade-offs of AI performance and reliability. In this chapter, we explore the role of the digital divide, the lack of diversity and representation in AI and STEM, and the influence of innovation on funding and research incentives within academia, government and industry when it comes to defining AI policy and stakeholder engagement. We consider the impacts of siloed efforts for increasing diversity and inclusion, and we examine how they fall short in moving the needle towards systemic change. We argue for shifting the understanding of innovation to one of inclusive innovation and offer examples of how we might begin to drive this shift. Placing inclusivity at the heart of the future we are shaping through digital technology will allow us to move from a fragmented digital age to one of wholeness that benefits all.

INTRODUCTION

Artificial intelligence (AI) is a driver of innovations in a wide variety of sectors and industries that have different needs and different problems to solve. As such, it should not be developed in the silos of the tech world or any single discipline. The design, development, deployment and assessment of technology such as AI is complex and requires interdisciplinarity. But what is innovation? The answer to this question is important for understanding what is currently driving or steering the development of new technologies.

Innovations through AI-based technologies and applications are rapidly changing many aspects of our lives. Unfortunately, it is not always for the good of humanity. In this chapter, we argue that the unintended consequences of unfit AI (these include increasing biases, disparities and inequalities) can be addressed by shifting our understanding of innovation to one of inclusive innovation. "Inclusive innovation" refers to "the means by which new goods and services are developed for and/or by those who have been excluded from the development mainstream; particularly the billions living on lowest incomes," ultimately broadening and diversifying the scope of stakeholders (Heeks et al., 2013, p. 1). This shift in thinking should be applied at all stages of AI-based technology development and deployment, as well as in policymaking and system improvement. Innovations in AI should be intrinsically inclusive and interdisciplinary. According to Dr. Katia Walsh, "artificial intelligence is the result of human intelligence, enabled by its vast talents and also susceptible to its limitations. Therefore, it is imperative that all teams that work in technology and AI are as diverse as possible" (as cited in Larsen, 2021). The degree to which AI can truly benefit the entire planet and beyond is tied to how much it accounts for this diversity.

In this chapter, we explore some of the major barriers to truly inclusive innovation and the measures needed for ensuring no one is left behind. These barriers include the divides and inherent biases already present in AI, how and why AI projects are currently funded, and the underlying investment incentives that are driving the direction of AI. The impact of technology that is not inclusive enough, as well as the lack of effort in addressing the root causes, is largely underestimated and is a barrier in the development of AI to benefit all. AI technologies are shaping the future of humanity and there needs to be a thoughtful reflection around this issue if we are to progress properly.

THE QUEST FOR INNOVATION

Since Alan Turing (1950) discussed how to build and test intelligent machines and the term "artificial intelligence" was coined in 1956 (McCarthy et al., 1955), there have been successes and setbacks throughout the seventy years of modern AI history. The surge in excitement for AI in the past decade has been driven by access to large amounts of data, cheaper and faster computers and the development of machine learning techniques, in particular deep learning. Nowadays, AI has permeated many aspects of our lives, from social media newsfeeds and online shopping to drug discovery (Fleming, 2018; Jiménez-Luna et al., 2021; Lada et al., 2021) and the fight against epidemics (Cho, 2020; Zeng et al., 2021). AI is one of the major forces revolutionizing human society and it is bringing in its trace a new era, the digital age.

Sadly, these advances have created a new type of global divide between the tech-rich and the tech-poor. The fast-paced advances of AI have deepened and widened the digital divide and amplified existing biases in, for example, academia, gender, race and rich-poor countries or populations (Carter et al., 2020). The current biases and divides in AI mirror some of the biases and divides that have plagued our societies for centuries. Whereas advances in technologies throughout history have often amplified colonialism, there is a real danger of moving towards new forms of colonialism reinforced by digital technologies and the current drivers of innovation in this field. Digital colonialism occurs when "large scale tech companies extract, analyze, and own user data for profit and market influence with nominal benefit to the data source" (Coleman, 2019, p. 417). It is the "exercise of imperial control at the architecture level of the digital ecosystem: software, hardware and network connectivity, which then gives rise to related forms of domination" (Kwet, 2019, p. 1). Take, for example, the division of work, with the invisible workers of AI such as data annotators, often from less privileged communities, who endure isolation and often difficult working conditions (Gray and Suri, 2019) and the biases in the data being used to train AI systems. As we design, develop and gather the data that are fed to machines so they can

learn, we are inevitably transferring our biases to AI. Addressing the neglected issues in AI development and policies will not only serve to improve the technology itself but could be instrumental in addressing both current and future systemic biases and divides (May, 2020).

It is clear to us that AI-based technologies give us two possible avenues: we will inadvertently perpetuate new forms of colonialism in the digital age (Voskoboynik, 2018) or humanity can move forward, in an inclusive manner, to pursue the common goal of resolving present global challenges and to drive impactful and beneficial innovations together. How can we make sure that the correct path is followed?

REIMAGINING THE KEY STAKEHOLDERS OF AI

A key to success for researchers in the academic world is to get funding for their research and publish in prominent journals and conference proceedings. To achieve this, many early-career researchers are advised to prioritize innovative research in their work. But what is innovation? When evaluating research proposals, funding agencies describe innovation as "creative, original, and transformative concepts and activities" or "unique and innovative methods, approaches, concepts, or advanced technologies" (National Science Foundation (NSF) and National Aeronautics and Space Administration (NASA) as cited in Falk-Krzesinski and Tobin, 2015, p. 15). With limited funding opportunities, a project often needs to show significant advances in the field to be rated as innovative. This often means striving for the newest and fastest method that requires abundant resources such as technologies, algorithms and systems developed with the use of high computing power, large storage, fast and reliable internet or cellular access (Thompson et al., 2020, p. 2).

This, however, narrows the scope of AI and its capacity to be used to meet global needs; currently its use is limited to what mostly drives profit in the Global North. In addition to the inherent issues this poses for a just and equitable society, it also presents technical challenges in developing "trustworthy and verifiable AI" (Dengel et al., 2021, p. 91) that is adaptable to limited resource settings. As stated by Dengel et al. (2021, p. 93),

> current research evaluation methods and academic criteria tend to favor vertical, short-term, narrow, highly focused, community- and discipline-dependent research. It is the responsibility of all scientists in the academic world to foster a methodological shift that facilitates (or at least does not penalize) long-term, horizontal, interdisciplinary, and very ambitious research.

This is also true for industry. As stated in the United Nations Conference on Trade and Development Technology and Innovation, "as with any new technology, many companies, when they innovate and develop new goods and services, they tend to focus on higher-income consumers that can bear the higher initial prices of these products" (UNCTAD, 2021, p. 125). Unfortunately, the suitability of those new technologies for developing countries is often overlooked ('Utoikamanu, 2018).

To achieve the needed shifts in research and development, we need to include all stakeholders that are impacted by innovations in AI, not just those currently benefiting from it. In this way, we can broaden the perception of innovation to include adaptability and new applications of existing techniques.

In order to widen the capacities of AI and shift to more inclusive innovations, we need to first grasp the inherent biases that are both driving and perpetuated by the current standards for innovation.

The digital divide and inclusive innovation

The digital divide is defined by the Organization for Economic Co-operation and Development (2001, p. 5) as "the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard to both their opportunities to access information and communication technologies (ICTs) and to their use of the Internet for a wide variety of activities."

This divide is most pronounced between the Global North and Global South. For example, in 2018 in Europe, 80 percent of the population was using the internet compared to only 25 percent of the population in Sub-Saharan Africa (UNCTAD, 2021, p. 78). Resources, both financial and technological, are predominantly concentrated in and directed to the Global North, often excluding stakeholders in the Global South from the global scientific research and innovation scene (Chan et al., 2021; Garcia, 2021; Mishra, 2021; Reidpath and Allotey, 2019; Skupien and Rüffin, 2020). However, significant divides within countries are also a major factor. The impacts of these divides were brought to light by the COVID-19 pandemic as the world shifted to life online, with work, shopping, healthcare services and education, requiring a computer and internet connection (United Nations, 2020). While this is a general issue and the pandemic example is recent, within the field of AI research and education, the lack of resources due to the tech divide is a pressing issue.

The technological divide contributes significantly to the lack of diversity in AI innovations. Or we could say that it overlooks or undervalues some great innovations. As this field and its funding depend on updated technology, the divide favors students and researchers from more privileged socioeconomic backgrounds (American University, 2020). Those that cannot afford computers or have no or slow internet access are excluded, and that impacts the field as a whole. When the vast majority of researchers in AI-based technologies are from a similar background, the rest of the world is squeezed out. Technology is often designed and developed by scientists for one sub-section of their country in one part of the world. The result is inequality fueled by the next updated technology.

To shrink this tech divide we need to reconsider the criteria for evaluating the quality of innovation in AI such that inclusive innovation is considered key. This would foster greater AI-based technologies that are adapted for different communities and would widen the pool of stakeholders. As we consider how the grand challenges of this century disproportionately affect marginalized people, having them at the forefront of innovation is critical to scaling tech innovation for good in the service of humanity and the planet.

Diversity and representation in AI and STEM

When inclusion is perceived as an act of charity or seen as accepting to lower our ambitions in order to advance the technology, we will always end up with biased technologies. Inclusive innovation needs to be seen for what it is: striving for a more inclusive, balanced technology that will benefit all. If we want to have machines capable of solving complex problems, we need to expose them to a wide variety of data. This means that people with diverse background expertise and experiences should be involved in all aspects of the development process of AI and AI-based technologies—data acquisition to train the AI systems, design, development, deployment, operation, monitoring and maintenance. This level of diversity should also be represented at all levels of AI-related policy- and decision-making. A lack of diversity and a failure to represent all stakeholders allow for omission by ignorance and not necessarily by intent, making it more difficult to address it explicitly (*Coded Bias*, 2020).

Diversity and representation issues are, of course, not inherent to AI. Historically, the fields of science, technology, engineering, and mathematics (STEM) have had a predominantly white, male base (Dancy et al., 2020, p. 1). Marginalization in STEM fields undeniably affects many communities, including Indigenous people, people with disabilities and the LGBTQ+ community (Miller and Downey, 2020; Schneiderwind and Johnson, 2020). Our focus in this chapter is on race, gender and socioeconomic status.

| TABLE 1 |

Percentage of people employed in the US in computer and mathematical occupations (Bureau of Labor Statistics, 2010; 2020).

	2010	2020
Women	25.8	25.2
Men	74.2*	74.8*
White	77.2*	65.4
Black or African-American	6.7	9.1
Asian	16.1	23.0

*Estimated percentage

Gender bias

Table 1 shows an uneven distribution of labor in the computer and mathematics field. It comes as no surprise that women in computer and math occupations represent only one-fourth of the field, and that alarmingly, over the last ten years, representation has slightly decreased. Although we can attribute the poor results to science-driven organizations not hiring women as much as men (Picture a Scientist, 2020), the gender gap in this field begins far earlier. As children, girls navigate the stereotypes that come with STEM from their parents, social norms and teachers, which demotivates many from pursuing interests in the field (Hill, C., 2020). Girls who do have an interest or perform well in math or science may not actually go into any STEM fields because they believe that these occupations are "inappropriate for their gender" (Hill et al., 2010, p. 22). The factors that influence women and girls from a young age demotivate many of them early on. Also, severe gender bias present in the workplace may cause women to leave STEM-related careers. This includes workplace environment, family responsibilities and implicit bias (Hill et al., 2010, pp. 24–25).

Implicit bias against women can be a significant impediment to success and advancing in a career; it can even be a factor in women's choice to leave. One example of the consequences of this bias is the tone of recommendation letters for women, where personality traits are often highlighted over technical expertise (Trix and Psenka, 2003, p. 215). This and other consequences of implicit bias reduce the involvement of women in AI-based technology design and their presence as policymakers engaged in science. Furthermore, women considered successful in their field are more derogated and less well-liked than successful men, which contributes to a negative workplace environment and can make it almost impossible for women to move forward. In the private sector, women in STEM leave due to unclear advancement opportunities, feeling isolated, an unsupportive environment and an extreme schedule (Hill et al., 2010, p. 24). When a workplace tries to push you out, with a lack of opportunities for advancement compounded by constant microaggressions, there is no real reason to stay.

In terms of marital status and family responsibilities, there are also clear differences between men and women. In STEM academia, single women are more likely to have a tenure-track position than their married counterparts. As well, due to the demands of the field and the tradition of women as primary caregivers, women abstain from having children or delay maternity (Hill et al., 2010, p. 26). Furthermore, a study into retention in engineering found that women were more likely to leave due to "time and family-related issues" (Frehill et al., 2008). These gender-based factors all contribute to minimal women applicants and low retention of women in STEM.

Racial bias

From 2010 to 2020, there was an upward trend of non-white people employed in computer and math occupations. The factors that contribute to a small percentage of people of color in this field are similar to the reasons outlined in the gender gap. In this section, we focus on the underrepresentation of Black, Asian and Hispanic or Latinx people in STEM. From a young age, implicit bias can make the difference between a student continuing their education or dropping out. One study found that low-income Black students that have at least one Black teacher in third, fourth, or fifth grade are 29 percent less likely to drop out of high school (Dodge, 2018). At the high school level, when STEM is usually introduced to students, they can begin pursuing their interests before going to college; but those opportunities are not equal for all. A study by Teach for America found that "one in four schools [in the US] offers computer science courses" (Dodge, 2018). Typically, schools in upper-class neighborhoods with a predominately white student body have this exposure to computer science, leaving minority and low-income students behind. Without this previous exposure in school, it is difficult for students to cultivate their interest in this subject and to believe that they can pursue a college education in a field perceived as requiring high innate ability (Leslie et al., 2015; Miller, 2017; Riegle-Crumb et al., 2019). They also often have a weaker sense of identity and of belonging to the "typical computer scientist" culture (Metcalf et al., 2018, p. 613). This perpetuates the belief (similarly as with women in STEM) that STEM careers are not appropriate for minorities, despite their interest (Dodge, 2018). We can see the consequences of such biases in the education system (in the United States, for example) by looking at the low percentage of Black, Asian and Hispanic people entering the STEM workforce (Barber et al., 2020; Clark and Hurd, 2020).

The workplace itself can be another battleground if the challenges of the education system are conquered. The racism and racial bias found in STEM significantly affect diversity in the field (McGee and Bentley, 2017; McGee, 2020). In San Francisco, for example, 60 percent of Black people and 42 percent of Asians and Hispanics in STEM have experienced some sort of racially motivated discrimination (Dodge, 2018). This discrimination is not always in the form of hate speech. Like with women, it comes in the form of wage gaps, microaggressions, not nominating minorities for advancement, not giving minorities important projects, or placing less value in their work (Dodge, 2018). These factors all contribute to a negative workplace that not only harms minorities but decreases their interest in the field and leads to them often choosing to leave the field altogether (Dodge, 2018). Therefore, attracting and retaining more minority populations into STEM education is a necessary first step to ameliorate the biases in AI.

We highlighted two major gaps that occur early in the education system: the implicit biases among educators and the lack of access to computer science courses for minority children. Educators are susceptible to unintentional, implicit bias and the role they play in whether or not children pursue their interests in STEM cannot be underestimated (Bushweller, 2021). It is therefore important that proper bias training be prioritized early in the education system, as gaps in this space amplify those we see later in STEM (Warikoo et al., 2016).

In terms of the lack of computer science courses, a possible solution to address this is to encourage non-profit, ideally minority-led, organizations that offer computer science programs. Hiring teachers, donating updated technology and finding suitable spaces are all essential parts for this to succeed. The advantage of having this be minority-led is that minority children perform better when they are taught by someone with a similar background (Rosen, 2018). Supporting community-led extracurricular programs is also a way to incentivize the uptake of such courses. As a result, a higher demand from communities will increase the likelihood that computer science courses be offered in the academic curriculum. This is part of the solution for inclusive innovation in AI and should not be neglected.

Socioeconomic bias

Finally, another significant bias in the field is socioeconomic status. A Yale study found that the way an individual pronounces certain words is telling of their social status (Cummings, 2019). While this is not a major issue in and of itself, a person's socioeconomic status can influence an employer's decision to hire them. The same study with 274 "individuals with hiring experience" found that, when lacking any information about qualifications, employers identified candidates from high socioeconomic status as better for the job than those from lower status (Cummings, 2019). Additionally, those who were from a higher social class were given better pay and more opportunities for bonuses.

This issue of bias is more general to the entire workforce. However, if we go back to the question of racial bias in STEM, we see that there is an intersection between race and income, although gender also intersects with these two. In the United States, many low-income neighborhoods tend to be dominated by minorities, more specifically Black and Latinx people. This is due to a long history of discrimination that segregated and ghettoized minorities (Firebaugh and Acciai, 2016, p. 13372). The result is poorly funded schools and limited access to jobs. Coupled with employers that are biased towards high-income applicants, a young person's socioeconomic status can influence their long-term future. There is no requirement for employers to hire a certain percentage from low-income neighborhoods. But without socioeconomic diversity in STEM, there is little representation from another sizable portion of the population. In addition, STEM-related technology that is developed to help these low-income neighborhoods will only be provided through a high-income lens.

Accounting for diversity in STEM is especially important in resolving some of the current major challenges with AI. One of these challenges is around "the lack of highly skilled experts in building AI systems" (Dengel et al., 2021, p. 93). As we have discussed, a significant portion of the population is currently excluded from developing and contributing talent and expertise to the AI field as a result of systemic biases (gender, race, socioeconomic status) even within countries that are on the tech-rich side of the digital divide. Another major challenge is with the efficiency of AI systems, given the insufficient representativeness in data fed into these systems (Kuhlman et al., 2020).

Humans feed their limited experiences and prejudices to a blank-slate algorithm and little by little it learns to reproduce this behavior. In the end, we have technology that is unreliable through no fault of its own. It simply did what it was supposed to: learn and replicate.

Algorithmic bias applied

The lack of resources and technological divide driving the lack of diversity in research comes to a head with the issue of learned bias in AI-based technologies. These algorithmic biases manifest themselves in applications as diverse as facial recognition technologies and hiring tools. In the documentary *Coded Bias*, Joy Buolamwini discovers that the AI in her Aspire Mirror project – a "device that enables you to look at yourself and see a reflection on your face based on what inspires you or what you hope

to empathize with" based on a face detection software[51] – does not recognize her face as a Black woman (*Coded Bias*, 2020). She resorts to wearing a plain white mask for her face to be seen. While this may simply seem to be an issue of a software error or bug, this technology has already begun to be applied to real-life uses, and Buolamwini's experience is replicated a thousandfold.

One of the most common uses of AI-based technologies is in surveillance and security, typically for facial recognition. *Coded Bias* explores this issue in detail. As Buolamwini explains, because the facial recognition algorithm is programmed by white men, they feed white, male faces to the algorithm. After bringing this issue up to companies like Microsoft and IBM, Buolamwini saw that IBM improved the accuracy of their algorithm to recognize not only skin color but gender, seen in Table 2.

| TABLE 2 |

IBM algorithm accuracy from 2017 and 2018 (Buolamwini, 2019).

	2017	2018
Skin color and gender		
Darker male	88.0%	99.4%
Lighter male	99.7%	99.7%
Darker female	65.3%	83.5%
Lighter female	92.9%	97.6%

In June 2020, Robert Williams, a Black man was arrested in Michigan, United States, for larceny following the use of facial recognition on the robber (Hill, 2020). Due to the police's confidence in the algorithm, they arrested him without doing a due diligence (i.e., checking his alibi, questioning witnesses, and so on). He was subsequently released, and the charges dropped, but the mistake made by the algorithm and poor police work could have cost Robert Williams his life.[52] With an overabundance of cameras installed, the use of facial recognition as a surveillance tool is slowly becoming a reality. And with that, the misidentification and prosecution of innocent people may skyrocket (Raji et al., 2020). Following the thread of bias in policing and security, AI-based technology is also found to unequally allocate police officers to certain communities (Heaven, 2020). There has historically been an over-policing of non-white communities, so-called "ghettoized" locations. An algorithm used in such contexts will learn where police and resources need to be allocated based on this historical data. It will learn to "increase vigilance in areas with a higher perceived propensity for crime, and will lead to an inequitable distribution of police and, in turn, inequitable criminalization" (Osoba and Welser IV, 2017, pp. 14–15).

51. For futher information about the Aspire Mirror Project, see: http://www.aspiremirror.com/

52. Another such incident happened in 2017. A Palestinian worker was wrongfully arrested because Facebook's automated translation system mistranslated a "good morning" post written in Arabic as "attack them" in Hebrew and "hurt them" in English. See Y. Berger (2017).

Consequently, the algorithm can and will lead to an increase of minorities in prison for petty crimes, such as marijuana possession, speeding or being homeless, amplifying the inherent biases in the system (Heaven, 2020; O'Donnell, 2019). Failing to correct for these biases will reinforce them within AI systems and perpetuate their unevenly assignment of police to marginalized communities.

There is also a substantial amount of algorithmic bias in AI-enabled hiring processes. Employers are overconfident in algorithms that will in fact expand the gaps created by previous biases in hiring—and this often occurs without the awareness of employers (Hickok, 2020). As Miranda Bogen (2019) explains, AI in hiring works at multiple levels before an applicant even applies. Targeted ads for jobs through Facebook, LinkedIn and Indeed contribute to reinforcing racial and gender stereotypes by predicting "who is most likely to click on the ad" (Bogen, 2019). A joint study by Northeastern University and the University of Southern California looked into the skewed delivery for job ads on Facebook. For example, in the most extreme cases, jobs as a cashier "reach an 85 percent female audience" and positions in taxi companies "reach a 75 percent Black audience" despite the employer's openness to all demographics (Ali et al., 2019, p. 4). However, the algorithm learned from the recruiters' applicant preference and targets people that align with this preference. Once again, its job is to adapt, learn, and replicate the data it receives.

Along the hiring process, the algorithm can eliminate a significant number of candidates who may have experience but do not use keywords or phrases the algorithm was trained on (Bogen, 2019). Some algorithms may also use past hiring decisions as guidance on who to reject, which can perpetuate discrimination (Dastin, 2018). Other hiring tools will determine who will be successful in a position, using past experience, performance reviews, tenure, and sometimes a lack of negative signals such as disciplinary action (Bogen, 2019). These hiring algorithms of course include the field of AI itself. The human biases we discuss in this chapter (gender, race and socioeconomic status) are compounded and replicated by hiring algorithms, perpetuating the vicious cycle that fuels the lack of representation in computer science and AI programming.

The challenges with AI that we have discussed so far—lack of diversity, applied algorithmic bias, siloed and discipline-dependent research and non-inclusive innovation—are often undervalued in terms of their impact on the quality of the technology and workforce, as well as on the future of humanity. By forgoing truly inclusive innovation, we are essentially trading off global well-being and prosperity alongside higher standards for AI performance and reliability (Dengel et al., 2021) in exchange for short-term profit. A major driver of this trade-off lies within the current funding structure for AI research, an issue we address in the next section.

AI FUNDING STRUCTURES AND INCENTIVES

At the root of the barriers to truly inclusive AI are questions of how and why AI research is funded. As it is now, the selected projects receiving funding from industries or government agencies sadly do not prioritize inclusion and diversity. They are often not interdisciplinary or collaborative and fail to account for growing human, social and natural capital to the same degree as they look for returns on investment. Those innovations influence policymakers in terms of what will drive the direction of new technologies, which then feeds back to how funding is directed. And thus, a complex vicious cycle is perpetuated.

Though not built intentionally, a vicious cycle is created from the interconnectedness of AI research projects, funding sources and policies and is reinforced by the limited diversity of stakeholders benefiting from and influencing AI innovations.

AI and academia

Most research in AI and new technologies is directly or indirectly supported or conducted by corporations. According to the Congressional Research Service's 2021 report on federal research and development funding (2020), 54 percent of applied research and 85 percent of development in the U.S. were funded by Business (see Figure 1). A recent assessment on AI policy and funding in Canada shows that even government funding of AI is primarily directed to "industry and academia with links to industry. Academia often acts as an intermediary between industry and government. Indirectly, these funds can still benefit for-profit organizations" (Brandusescu, 2021, p. 37). This can have a very strong impact on academic research, policymaking and the extent to which corporations influence innovation in AI.

The presence of the private sector in the academic field of AI is inextricable. According to the AI index report produced by the Stanford Institute for Human-Centered Artificial Intelligence (Zhang et al., 2021, p. 21), more than 15 percent of the peer-reviewed AI publications in 2019 are from corporations in every major country and region of the world. Industry is also absorbing the majority of AI expertise coming out of academia; in 2019, 65 percent of Ph.D. candidates in AI in North America went into industry after graduating (Zhang et al., 2021, p. 4). Corporations also sponsor or are highly present at many conferences and workshops in the field (Alford, 2021). For example, at the International Conference on Learning Representation (ICLR) in 2021, nearly 30 percent of the accepted papers were from corporations such as Google, Amazon, IBM and Facebook. Also, Google had four papers among the eight outstanding paper award winners and Facebook had one (ICLR, 2021).

Most corporations working in AI are driven by R&D agendas that are highly influenced by market demands and return on investment. Innovations that revolutionize the field, bring in new assets and broaden horizons are central to the R&D agendas. Also, expertise and skillsets developed in academia are a key resource. This is one of the reasons for the industry to fund academic research. It follows, then, that the industry's needs ultimately influence many funded academic research projects. This dynamic between industry and academia creates a twofold tension. First, given that most major players in the AI and technology industry are concentrated in the Global North (Chan et al., 2021), the gaps that are limiting innovation that truly benefits all are further amplified. Second, the industry-academia dynamic disproportionately drives the direction of the field towards private sector interests rather than the public good.

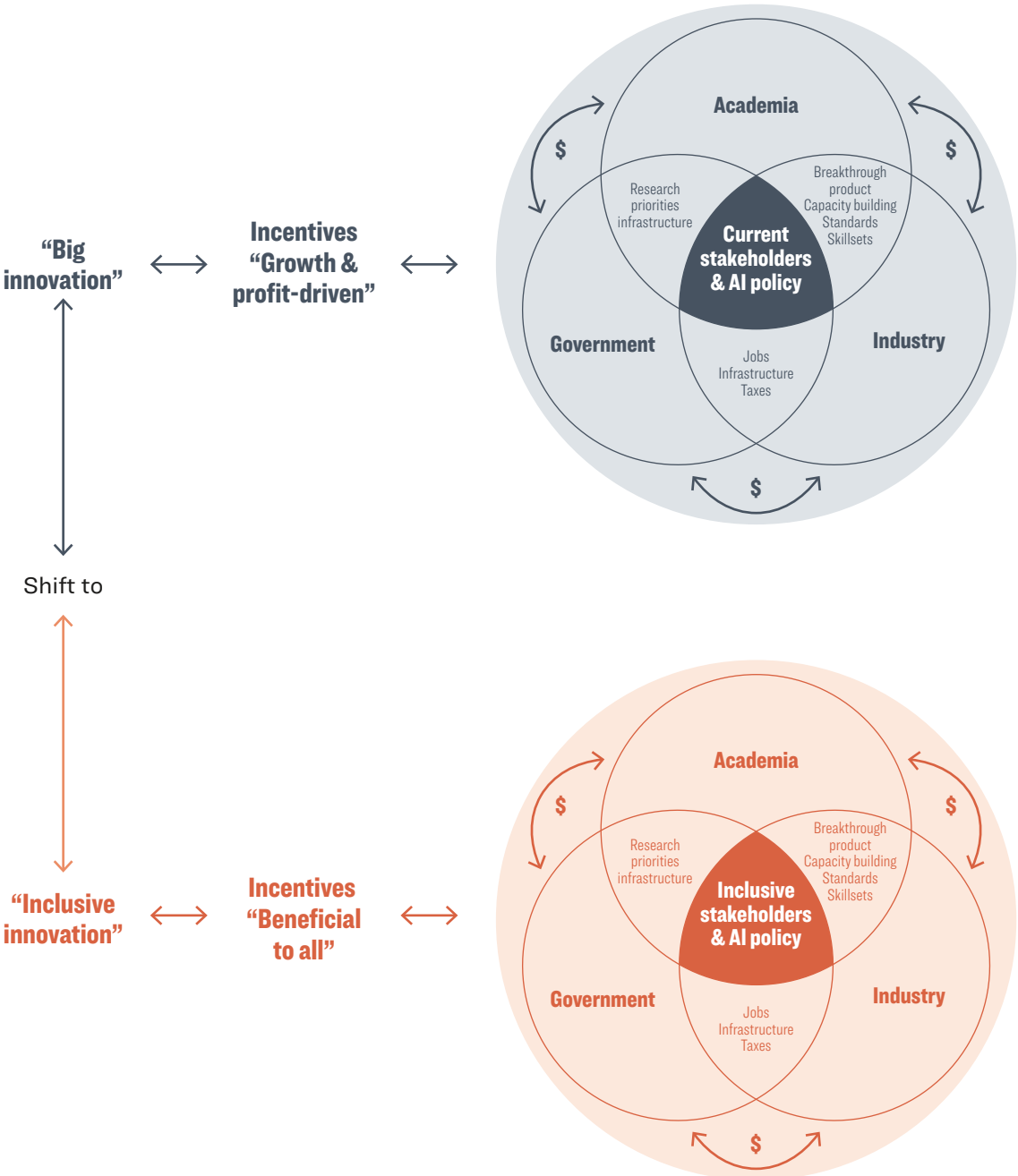
Close collaborations between industry and academia are not inherently problematic and can benefit research and education in academic institutions (Etzioni, 2019). Stakeholder capitalism—"a form of capitalism in which companies seek long-term value creation by taking into account the needs of all their stakeholders, and society at large"—can be seen as a solution that works for people and the planet (Schwab and Vanham, 2021). However, this requires industries to place interdisciplinarity and inclusive innovation at the heart of their AI strategy, ultimately shifting market demand towards the public good and including marginalized people and communities among key stakeholders.

Feedback loops: Government funding, private sector incentives and policy

Excitement about AI is shifting funding away from more basic research towards applied research and "big innovation" that can be commercialized in the short and medium terms. Thus, AI is rapidly shifting the funding playing field for both the public and private sectors. Applied R&D is often incentivized by the potential for return on investment in terms of both profit and growth. Most applied research is currently funded by the private industry (Congressional Research Service 2020, Figure 1). Since the industry also indirectly funds academic research (for example, by supporting government funding programs) (Brandusescu, 2021), it is difficult to dismiss the strong influence of the private sector on the direction of AI. Further, this influence is feeding back into policy strategies for economic growth, which also influence government funding programs (see Figure 1).

| FIGURE 1 |

AI funding structures and incentives Adapted from models by Kimatu, J. N. (2016) and Ondimu, S. (2012).



As such, "it is worth questioning how the innovation economy is influenced by private interests and private power—and by extension, how AI public policy gets written" (Brandusescu, 2021, p. 38). Considering the feedback mechanism between public or government funding and innovations within the private AI sector, the need to place inclusivity at the heart of these innovations has never been more pressing if we want to have AI-based technologies that benefit all and are trusted by all. Innovation is a critical driver of research and funding incentives with direct impacts on academia, government and industry. Shifting from "big innovation" to inclusive innovation can shift funding and research dynamics, AI policy and stakeholder engagement to ensure that no one is left behind.

There are advantages to extending the focus of AI beyond its current siloed emphasis on science and technology and into fields such as neuroscience, computational linguistics, ethics, sociology and anthropology (Rahwan et al., 2019, p. 477). These advantages include increased interdisciplinarity and integration of the skillsets beyond the technical that are sorely missing from AI in general and that, as such, are hindering the field's progress (Dengel et al., 2021).

The *U.S. National Artificial Intelligence Act of 2020* (United States Congress, 2020) seeks to diversify funding to AI research and its applications to a wider scope of government agencies beyond national defense, which was previously the main driver of AI policies in the U.S. (Delgado and Levy, 2021). This act and other policies and initiatives are starting to change how funding agencies are operating, recognizing the needed change in funding priorities: "Artificial intelligence is increasingly becoming a highly interdisciplinary field with expertise required from a diverse range of scientific and other scholarly disciplines that traditionally work independently and continue to face cultural and institutional barriers to large scale collaboration" (United States Congress, 2020).

However, as stated in one of the U.S. Congress findings, "current federal investments and funding mechanisms are largely insufficient to incentivize and support the large-scale interdisciplinary and public-private collaborations that will be required to advance trustworthy artificial intelligence systems in the United States" (United States Congress, 2020, pp. 3–4). This is not surprising when we consider the heavy influence of private-sector interests on funding criteria for research and innovation outlined above. Also, given these criteria, it is not uncommon for researchers to adapt their work to fit the available funding opportunities. Therefore, in addition to AI public policy being caught in this vicious cycle, the quality of the AI itself only needs to satisfy the demands of the market. And unfortunately, presently the AI agenda is mainly in the hands of a limited number and diversity of stakeholders (Delgado and Levy, 2021). In order to shift the direction of incentives and break the cycle, criteria for funding should prioritize inclusive innovation as a focal point. Profit and growth need to be weighed against opportunities for scaling thriving human, natural and social capital.

One way for inclusive innovation to be prioritized is for government and industry to support more projects that are community-based, collaborative and interdisciplinary. Unfortunately, presently reaching for highly rated innovative projects too often means not selecting to work on inclusive AI in a limited-resource setting designed for local impact since this kind of project does not explicitly revolutionize the field in the short term and does not attract funding. For example, the new National Science Foundation (NSF, 2021) commitment to increase funding for applied AI may seek to diversify research. However, without anchoring these types of initiatives in inclusive innovation, this may actually move money towards tech innovation and away from fundamental research that does not present short-term commercial viability, which will disadvantage students who are interested in research focused on the public good (Viglione, 2020). Changing the incentives currently driving the vicious cycle in funding can allow for more early- to mid-career researchers to take on projects that prioritize inclusive innovation and nurture expertise for inclusive AI. We argue that inclusive innovation is the real and only innovation that should be considered in AI if we want it to benefit all. When AI innovation occurs in silos and is mostly incentivized by the industry, pressured by shareholders and profit, this outcome is not possible.

THE ISSUE OF TRICKLE-DOWN SCIENCE

To justify the current lack of inclusion in innovation, the concept of trickle-down economics has at times been extended to "trickle-down science." Having a high concentration of resources and scholars in the Global North is expected to "produce the best science" whose "methods, theories, and insights" will trickle down into the Global South (Reidpath and Allotey, 2019, p. 1). Just as with trickle-down economics, this is not viable, and in fact the opposite is happening (Reidpath and Allotey, 2019, p. 1), partly because of the funding incentives and the current drivers of market demands discussed earlier in this chapter. For example, the high demand for resources coupled with fewer regulations and privacy protections in the Global South is driving the increased exploitation of both human resources (for instance, to perform activities such as data mining) and natural resources, such as the extraction of minerals (Arezki, 2021; Arun, 2020, p. 594; Mishra, 2021).

It is also important to consider the political environment of regions where AI-based applications are deployed. Oftentimes technology developed for the privileged few can be harmful in less-resourced regions. For example, the UN investigators' report on the genocide of the Rohingya population in Myanmar in 2017 noted that "Facebook [had] been a useful instrument for those seeking to spread hate" (Human Rights Council, 2018, p. 34). This demonstrates the powerful effect that social media technologies can have on human rights when used in a place where the political and media environments are not healthy.

Rapid innovation often occurs at the expense of those who would supposedly benefit from the trickle-down philosophy, for whom the harmful repercussions disproportionately outweigh any potential benefit (Schia, 2018, p. 827). This is only reinforced by the continued exclusion of marginalized people and communities as key stakeholders. As Shirley Chisholm stated, "If they don't give you a seat at the table, bring a folding chair." The importance of ensuring representation, however challenging it may be, cannot be understated. But even then, the work is far from done.

WHAT IT REALLY MEANS TO HAVE A SEAT AT THE TABLE

The challenges described in this chapter are of course not just related to AI or how we perceive innovation. They are representative of broader systemic issues that are evolving daily. A key barrier to addressing them is that current efforts are siloed rather than approached from a systems perspective.

Many AI initiatives are already making a lot of progress toward having AI benefit all and including more voices. These include AI4ALL,[53] the African Master's in Machine Intelligence,[54] Quantum Leap Africa,[55] the Centro de Excelencia en Inteligencia Artificial en Medellín[56] and the African Supercomputing Center at Morocco's UM6P university[57].

53. See the AI4ALL website (2021) for more information.

54. For more information on AIMS and the African Master's in Machine Intelligence, see AIMS (2021).

55. See Quantum Leap Africa (2021) for additional details.

56. This Center was set up in partnership between Ruta N in Colombia and the Institute for Robotic Process Automation & Artificial Intelligence (IRPAAI). See Ruta N (2018) for more information.

57. The ASCC is in the University Mohamed VI Polytechnic, more information is found on their website (see ASCC 2020).

Corporations such as Google and Microsoft, as well as foundations, policymakers and government funding agencies, do invest in some ways in AI projects for the social good,[58] and as such, they are funding projects that would probably not otherwise be selected within current funding guidelines. However, when these initiatives are not anchored in inclusive innovation, they can sometimes backfire and intensify the marginalization of minorities (Latonero, 2019). For example, by exclusively placing these funding opportunities outside of mainstream funding cycles, the notion that these projects are marginal is reinforced rather than addressed.

This occurs on an individual level as well. When the only entry points for marginalized people into innovative research is through specialized programs, the feeling of the imposter syndrome (Tulshyan and Burey, 2021), so common to minorities in science and high-level positions, is accentuated. These programs intend to reduce the gap and include more minorities behind the scenes and as decision-makers. But they rarely address the broader systemic issues that result in prejudices, toxic environments and toxic colleagues that perpetuate the idea that minorities need to be invited into the circle of scientists and leaders. As discussed in previous sections, the personal self-doubt that starts in childhood coupled with a panoply of social barriers and expectations contribute to the silencing of minority voices that are at the table.

Unfortunately, these inclusion programs are often perceived as enough to bridge the gaps (Puritty et al., 2017). Of course, in practice, they are not. We see this with the percentage of women in math and computer science jobs in the U.S. dropping from 25.8 percent in 2010 to 25.2 percent in 2020 (See Table 1). They are good initiatives, so why are they not working as intended? As mentioned by Dengel et al. (2021, p. 90), "we still need a lot of work in research and a paradigm shift in AI to develop a real AI for humanity—a human-centric AI." We argue that an essential requirement for this paradigm shift is to place inclusivity at the center of innovation rather than on the peripheries or as an afterthought.

We are not the first to point out all the biases and problems in AI technologies. We are also not the first to mention how much progress has been made. But it is important to continue elevating the standard for inclusivity and innovation. This can only serve to improve the systems in which we operate and the innovations we strive for (as we saw with facial recognition systems, for example). According to Giridharadas (2021), an author known for his critique of elites' exclusionary take on world issues that should be subjected to collective action:

> all grand challenges [...] require public, institutional, democratic and universal solutions. They need to solve the problem at the root and for everyone. What we do together is more interesting, compelling, more powerful, more valuable than what we do alone. Current neo-liberal myth is that what we do alone is better and more beautiful than what we do together. We need to bring back the notion that we live in society within which we have interdependence. Valuing what we do together needs to be reclaimed. Only this collective intelligence will allow us to solve the grand challenges we face.

58. Examples include the Google Impact Challenge for Women and Girls (2021), the Microsoft AI for Good Research Lab (Microsoft Research, 2021), and the Creating Sustained Social Impact, by their Corporate Citizenship branch (Microsoft Corporate Citizenship, 2021).

CONCLUSION

The development of new algorithms, the advancement in computational resources and the availability of abundant data have driven the recent surge of innovation in AI. This is driving transformations in a wide range of industries and sectors that will likely revolutionize society as did past industrial revolutions. As such, humanity is once again confronted with the danger of perpetuating the repercussions of inequitable systems change driven by colonial mentalities and socio-economic divides. In particular, the digital divide is amplifying inequalities in terms of access to AI and the harmful consequences of human bias in AI-based technologies.

As discussed throughout the chapter, addressing the neglected issues in AI development starts by addressing our own human biases. The quality and accuracy of AI-based systems are compromised by the lack of diversity in data and human resources at all stages of AI development.

This is amplified by the marginalization based on gender, race and socioeconomic status of entire groups within STEM, which also worsens the talent shortage currently challenging the AI field.

As a result of a sky-rocketing demand for AI, a vicious cycle for funding, both private and public, is reinforced by the underlying incentives for rapid, short-term profit and economic growth, steering the direction of AI and the digital age. This vicious cycle is common to the rapid growth mentality focused on "big innovation." As we argue in this chapter, the focal point needs to shift to one of inclusive innovation, thereby increasing the diversity of voices and enabling greater capacity-building, especially within marginalized, resource-poor communities. For AI to truly reflect the power of human consciousness it should be a representation of the beauty and the power of diversity.

The increasing interconnectedness of global systems and challenges is shifting the emphasis from pure profit to valuing natural, human and social capital. There is no way for people and the planet to thrive without this shift. Prioritizing local solutions that embody universal ethical principles of trust, responsibility and empathy is key. Though it may be tempting to prioritize rapid growth and short-term profit for the sake of innovation, this approach will inevitably limit our AI systems to benefit a privileged few rather than to humanity as a whole. Once AI research and development is driven by inclusive innovation, we will be able to shift from a fragmented AI to one of wholeness that benefits all, including future generations.

REFERENCES

AI4ALL. 2021. Home page. https://ai-4-all.org/

AIMS (African Masters in Machine Intelligence). 2021. Home page. https://aimsammi.org

Alford, A. 2021. AI conference recap: Google, Microsoft, Facebook, and others at ICLR 2021. *InfoQ*. June 8. https://www.infoq.com/news/2021/06/conference-recap-iclr-2021/

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A. and Rieke, A. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *Proceedings of the ACM on Human-Computer Interaction,* Vol. 3, No. 199, pp. 1–30. https://dl.acm.org/doi/pdf/10.1145/3359301

American University. 2020. Understanding the digital divide in education. School of Education Online blog, December 15. https://soeonline.american.edu/blog/digital-divide-in-education

Arezki, R. 2021. Transnational governance of natural resources for the 21st century. Brookings Institution blog, July 7. https://www.brookings.edu/blog/future-development/2021/07/07/transnational-governance-of-natural-resources-for-the-21st-century/

Arun, C. 2020. AI and the Global South: Designing for other worlds. M. D. Dubber, F. Pasquale and S. Das (eds), *The Oxford Handbook of Ethics of AI*. Oxford, Oxford University Press, pp. 589–606.

ASCC (African SuperComputing Center). 2020. Home page.

Barber, P. H., Hayes, T. B., Johnson, T. L. and Márquez-Magaña, L. 2020. Systemic racism in higher education. *Science*, Vol. 369, No. 6510, pp. 1440–1441. https://www.science.org/doi/pdf/10.1126/science.abd7140

Berger, Y. 2017. Israel arrests Palestinian because Facebook translated "good morning" to "attack them." *Haaretz*, October 22. https://www.haaretz.com/israel-news/palestinian-arrested-over-mistranslated-good-morning-facebook-post-1.5459427

Bogen, M. 2019. All the ways hiring algorithms can introduce bias. *Harvard Business Review*, May 6. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

Brandusescu, A. 2021. *Artificial intelligence policy and funding in Canada: Public investments, private interests*. Centre for Interdisciplinary Research on Montreal, pp. 11–51. https://www.mcgill.ca/centre-montreal/files/centre-montreal/aipolicyandfunding_report_updated_mar5.pdf

Buolamwini, J. 2019. Compassion through computation: Fighting algorithmic bias. Video, World Economic Forum. https://youtu.be/5PGYOYZKsdY

Bureau of Labor Statistics (United States). 2010. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity, labor force statistics from the current Population Survey. https://www.bls.gov/cps/aa2010/cpsaat11.pdf

––––. 2020. Employed persons by detailed occupation, sex, race, and Hispanic or Latino Ethnicity, labor force statistics from the current Population Survey. https://www.bls.gov/cps/cpsaat11.pdf

Bushweller, K. 2021. How to get more students of color into STEM: Tackle bias, expand resources. Education Week web article, March 2. https://www.edweek.org/technology/how-to-get-more-students-of-color-into-stem-tackle-bias-expand-resources/2021/03

Carter, L., Liu, D. and Cantrell, C. 2020. Exploring the intersection of the digital divide and artificial intelligence: A hermeneutic literature review. *AIS Transactions on Human-Computer Interaction,* Vol. 12, No. 4, pp. 253–275. https://aisel.aisnet.org/thci/vol12/iss4/5/

Chan, A., Okolo, C. T., Terner, Z. and Wang, A. 2021. *The limits of global inclusion in AI development*. Association for the Advancement of Artificial Intelligence. https://arxiv.org/pdf/2102.01265.pdf

Cho, A. 2020. Artificial intelligence systems aim to sniff out signs of COVID-19 outbreaks. *Science*, May 12. https://www.sciencemag.org/news/2020/05/artificial-intelligence-systems-aim-sniff-out-signs-covid-19-outbreaks

Clark, U. S. and Hurd, Y. L. 2020. Addressing racism and disparities in the biomedical sciences. *Nature Human Behaviour*, Vol. 4, No. 8, pp. 774–777. https://www.nature.com/articles/s41562-020-0917-7

Coded Bias. 2020. Motion picture, 7th Empire Media, Brooklyn, directed by Shalini Kantayya.

Coleman, D. 2019. Digital colonialism: The 21st century scramble for Africa through the extraction and control of user data and the limitations of data protection laws. *Michigan Journal of Race and Law,* Vol. 24, No. 2, pp. 417–439. https://repository.law.umich.edu/mjrl/vol24/iss2/6

Congressional Research Service (United States). 2020. *Federal Research and Development (R&D) Funding: FY2021*. https://fas.org/sgp/crs/misc/R46341.pdf

Cummings, M. 2019. Yale study shows class bias in hiring based on few seconds of speech. *YaleNews*, October 21. https://news.yale.edu/2019/10/21/yale-study-shows-class-bias-hiring-based-few-seconds-speech

Dancy, M., Rainey, K., Stearns, E., Mickelson, R. and Moller, S. 2020. Undergraduates' awareness of white and male privilege in STEM. *International Journal of STEM Education,* Vol. 7, No. 52, pp. 1–17. https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-020-00250-3

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 10. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Delgado, F. A. and Levy, K. 2021. A community-centered research agenda for AI innovation policy. *Cornell Policy Review*, May 4. https://www.cornellpolicyreview.com/a-community-centered-research-agenda-for-ai-innovation-policy/

Dengel, A., Etzioni, O., DeCario, N., Hoos, H., Li, F., Tsujii, J. and Traverso, P. 2021. Next big challenges in core AI technology. B. Braunschweig and M. Ghallab (eds.), *Reflections on Artificial Intelligence for Humanity*. Lecture Notes in Computer Science, Vol. 12600, Springer, Cham, pp. 90–115. https://doi.org/10.1007/978-3-030-69128-8_7

Dodge, A. 2018. What you need to know about the stem race gap. Ozobot blog, February 20. https://ozobot.com/blog/need-know-stem-race-gap

Etzioni, O. 2019. AI academy under siege. *Inside Higher Ed*, November 20. https://www.insidehighered.com/views/2019/11/20/how-stop-brain-drain-artificial-intelligence-experts-out-academia-opinion

Falk-Krzesinski, H. J. and Tobin, S. C. 2015. How do I review thee? Let me count the ways: A comparison of research grant proposal review criteria across US federal funding agencies. *The Journal of Research Administration*, Vol. 46, No. 2, pp. 79–94. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4892374/

Firebaugh, G. and Acciai, F. 2016. For blacks in America, the gap in neighborhood poverty has declined faster than segregation. *Proceedings of the National Academy of Sciences*, Vol. 113, No. 47, pp. 13372–13377. https://www.pnas.org/content/pnas/113/47/13372.full.pdf

Fleming, N. 2018. How artificial intelligence is changing drug discovery. *Nature*, Vol. 557, No. 7706, pp. 55–57. link.gale.com/apps/doc/A572639347/AONE

Frehill, L. M., Di Fabio, N., Hill, S., Traeger, K., and Buono, J. 2008. Women in engineering: A review of the 2007 literature. *SWE Magazine*, Vol. 54, pp. 6–30.

Garcia, E. 2021. The international governance of AI: Where is the Global South? The Good AI blog, January 28. https://thegoodai.co/2021/01/28/the-international-governance-of-ai-where-is-the-global-south/

Giridharadas, A. 2021. Philanthropy and the state: Who is funding what and why? Video, UCL Institute for Innovation and Public Purpose. https://www.youtube.com/watch?v=fOAkNu7Y6f4

Google. 2021. Google Impact Challenge for Women and Girls: Introduction. https://impactchallenge.withgoogle.com/womenandgirls2021

Gray, M. L. and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston, Houghton Mifflin Harcourt.

Heaven, W.D. 2020. Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*, July 17. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice

Heeks, R., Amalia, M., Kintu, R., and Shah, N. 2013. Inclusive innovation: Definition, conceptualisation and future research priorities. *Manchester Center for Development Informatics*, No. 53, pp. 1–28. https://www.researchgate.net/publication/334613068_Inclusive_Innovation_Definition_Conceptualisation_and_Future_Research_Priorities

Hickok, M. 2020. Why was your job application rejected? Bias in recruitment algorithms, part 1. Montreal Ethics AI Institute blog, July 12. https://montrealethics.ai/why-was-your-job-application-rejected-bias-in-recruitment-algorithms-part-1/

Hill, C. 2020. The STEM gap: Women and girls in science, technology, engineering and math. AAUW resources section. https://www.aauw.org/resources/research/the-stem-gap/

Hill, C., Corbett, C. and St. Rose, A. 2010. *Why so Few? Women in Science, Technology, Engineering, and Mathematics*. Washington DC, AAUW. https://www.aauw.org/app/uploads/2020/03/why-so-few-research.pdf

Hill, K. 2020. Wrongfully accused by an algorithm. *The New York Times*, June 24. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Human Rights Council. 2018. *Report of the Independent International Fact-Finding Mission on Myanmar*. Geneva. https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf

ICLR (International Conference on Learning Representations). 2021. Announcing ICLR 2021 Outstanding Paper Awards. https://iclr-conf.medium.com/announcing-iclr-2021-outstanding-paper-awards-9ae0514734ab

Jiménez-Luna, J., Grisoni, F., Weskamp, N. and Schneider, G. 2021. Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert Opinion on Drug Discovery*, Vol. 16, No. 9, pp. 1–11. https://www.tandfonline.com/doi/pdf/10.1080/17460441.2021.1909567

Kimatu, J. N. 2016. Evolution of strategic interactions from the triple to quad helix innovation models for sustainable development in the era of globalization. *Journal of Innovation and Entrepreneurship,* Vol. 5, No. 16, pp. 1–7. https://innovation-entrepreneurship.springeropen.com/articles/10.1186/s13731-016-0044-x

Kuhlman, C., Jackson, L. and Chunara, R. 2020. No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv Preprint.* https://arxiv.org/pdf/2002.11836.pdf

Kwet, M. 2019. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class,* Vol. 60, No. 4, pp. 3–26. https://journals.sagepub.com/doi/pdf/10.1177/0306396818823172

Lada, A., Wang, M. and Yan, T. 2021. How machine learning powers Facebook's news feed ranking algorithm. Engineering at Meta blog, January 26. https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/

Larsen, J. 2021. Levi-Strauss' Dr. Katia Walsh on why diversity in AI and ML is non-negotiable. *VentureBeat*, August 2. https://venturebeat.com/2021/08/02/levi-strauss-dr-katia-walsh-on-why-diversity-is-non-negotiable-in-ai-and-machine-learning/

Latonero, M. 2019. Opinion: AI for good is often bad. *Wired*, November 18. https://www.wired.com/story/opinion-ai-for-good-is-often-bad/

Leslie, S., Cimpian, A., Meyer, M. and Freeland, E. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, Vol. 347, No. 6219, pp. 262–265. https://www.science.org/doi/full/10.1126/science.1261375

May, A. 2020. Dr. Fei-Fei Li: "We can make humanity better in so many ways." *Artificial Intelligence in Medicine*, December 12. https://ai-med.io/ai-champions/dr-fei-fei-li-we-can-make-humanity-better-in-so-many-ways/

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. 1955. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McGee, E. O. 2020. Interrogating structural racism in STEM higher education. *Educational Researcher*, Vol. 49, No. 9, pp. 633–644. https://journals.sagepub.com/doi/full/10.3102/0013189X20972718

McGee, E. and Bentley, L. 2017. The troubled success of Black women in STEM. *Cognition and Instruction*, Vol. 35, No. 4, pp. 265–289. https://www.tandfonline.com/doi/pdf/10.1080/07370008.2017.1355211

Metcalf, H.E., Crenshaw, T.L., Chambers, E.W. and Heeren, C. 2018. Diversity across a decade: A case study on undergraduate computing culture at the University of Illinois. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education. Association of Computing Machinery*, pp. 610–615. https://dl.acm.org/doi/abs/10.1145/3159450.3159497

Microsoft Corporate Citizenship. 2021. Creating sustained societal impact. https://www.microsoft.com/en-hk/sparkhk/creating-sustained-societal-impact

Microsoft Research. 2021. AI for Good Research Lab: Overview. https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/

Miller, O. 2017. The myth of innate ability in tech. Personal blog, January 9. http://omojumiller.com/articles/The-Myth-Of-Innate-Ability-In-Tech 4

Miller, R. A. and Downey, M. 2020. Examining the STEM climate for queer students with disabilities. *Journal of Postsecondary Education and Disability*, Vol. 33, No. 2, pp. 169–181. https://www.researchgate.net/publication/334654579_Examining_the_STEM_Climate_for_Queer_Students_with_Disabilities

Mishra, S. 2021. Opinion: Is AI deepening the divide between the Global North and South? *Newsweek*, March 9. https://www.newsweek.com/ai-deepening-divide-between-global-north-south-opinion-1574141

NSF. 2021. National Science Foundation Graduate Research Fellowship Program. https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=6201

O'Donnell, R. M. 2019. Challenging racist predictive policing algorithms under the equal protection clause. *New York University Law Review*, Vol. 94, No. 544, pp. 544–580. https://www.nyulawreview.org/wp-content/uploads/2019/06/NYULawReview-94-3-ODonnell.pdf

Ondimu, S. 2012. Possible approaches to commercialisable university research in Kenya. *The 7th KUAT Scientific, Technological and Industrialization Conference*, pp. 1–16. https://www.researchgate.net/publication/328095915_Possible_Approaches_to_Commercialisable_University_Research_in_Kenya

Organization for Economic Co-operation and Development. 2001. *Understanding the Digital Divide*. https://www.oecd.org/digital/ieconomy/1888451.pdf

Osoba, O. and Welser IV, W. 2017. *An Intelligence in our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, RAND Corporation. https://www.rand.org/pubs/research_reports/RR1744.html

Picture a Scientist. 2020. Motion picture, Uprising Production, Antarctica, directed by Ian Cheney and Sharon Shattuck.

Puritty, C., Strickland, L. R., Alia, E., Blonder, B., Klein, E., Kohl, M. T., McGee, E., Quintana, M., Ridley, R. E., Tellman, B. and Gerber, L. R. 2017. Without inclusion, diversity initiatives may not be enough. *Science*, Vol. 357, No. 6356, pp. 1101–1102. https://www.science.org/doi/full/10.1126/science.aai9054

Quantum Leap Africa. 2021. Preparing Africa for the Coming Quantum Revolution. https://quantumleapafrica.org

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., Mcelreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B. and Wellman, M. 2019. Machine behaviour. *Nature*, Vol. 568, No. 7753, pp. 477–486. doi: 10.1038/s41586-019-1138-y.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J. and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* pp. 145–151. https://dl.acm.org/doi/pdf/10.1145/3375627.3375820

Reidpath, D. and Allotey, P. 2019. The problem of "trickle-down science" from the Global North to the Global South. *BMJ Global Health*, Vol. 4, No. 4, pp. 1–3. https://gh.bmj.com/content/bmjgh/4/4/e001719.full.pdf

Riegle-Crumb, C., King, B. and Irizarry, Y. 2019. Does STEM stand out? Examining racial/ethnic gaps in persistence across postsecondary fields. *Educational Researcher*, Vol. 48, No. 3, pp. 133–144. https://journals.sagepub.com/doi/pdf/10.3102/0013189X19831006

Rosen, J. 2018. Black students who have one Black teacher are more likely to go to college. Johns Hopkins University Hub, November 12. https://hub.jhu.edu/2018/11/12/black-students-black-teachers-college-gap/

Ruta N. 2018. *Ruta N Medellín: Centro de Innovación y Negocios Inicio.* https://www.rutanmedellin.org/es/

Schia, N. N. 2018. The cyber frontier and digital pitfalls in the Global South. *Third World Quarterly*, Vol. 39, No. 5, pp. 821–837. https://www.tandfonline.com/doi/pdf/10.1080/01436597.2017.1408403

Schwab, K. and Vanham, P. 2021. What is stakeholder capitalism? *European Business Review*, January 22. https://www.europeanbusinessreview.eu/page.asp?pid=4603

Schneiderwind, J. and Johnson, J. M. 2020. Why are students with disabilities so invisible in STEM education? *Education Week*, July 27. https://www.edweek.org/education/opinion-why-are-students-with-disabilities-so-invisible-in-stem-education/2020/07

Skupien, S. and Rüffin, N. 2019. The geography of research funding: Semantics and beyond. *Journal of Studies in International Education*, Vol. 24, No. 1, pp. 24–38. https://journals.sagepub.com/doi/pdf/10.1177/1028315319889896

Thompson, N. C., Greenewald, K., Lee, K. and Manso, G. F. 2020. The computational limits of deep learning. *MIT Initiative on the Digital Economy Research Brief,* Vol. 4, pp. 1–16. https://arxiv.org/pdf/2007.05558.pdf

Trix, F. and Psenka, C. 2003. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, Vol. 14, No. 2, pp. 191–220. https://journals.sagepub.com/doi/pdf/10.1177/0957926503014002277

Tulshyan, R. and Burey, J. A. 2021. Stop telling women they have imposter syndrome. *Harvard Business Review*, February 11. https://hbr.org/2021/02/stop-telling-women-they-have-imposter-syndrome

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, Vol. 54, No. 236, pp. 433–460.

United Nations. 2020. Digital divide "a matter of life and death" amid COVID-19 crisis, Secretary-General warns virtual meeting, stressing universal connectivity key for health, development. Press release, June 11. https://www.un.org/press/en/2020/sgsm20118.doc.htm

UNCTAD. 2021. *Technology and Innovation Report: Catching Technological Waves—Innovation with Equity*. New York, United Nations Publications. https://unctad.org/webflyer/technology-and-innovation-report-2021

United States Congress. 2020. H.R.6216 – National Artificial Intelligence Initiative Act of 2020, pp. 1–56. https://www.congress.gov/bill/116th-congress/house-bill/6216/text#toc-H7A238FDF26594A338CB94267854F51D4

'Utoikamanu, F. 2018. Closing the technology gap in least developed countries. *UN Chronicle*, December. https://www.un.org/en/chronicle/article/closing-technology-gap-least-developed-countries

Viglione, G. 2020. NSF grant changes raise alarm about commitment to basic research. *Nature*, Vol. 584, No. 7820, pp. 177–178. https://www.nature.com/articles/d41586-020-02272-x

Voskoboynik, D. M. 2018. To fix the climate crisis, we must face up to our imperial past. *OpenDemocracy*, October 8. https://www.opendemocracy.net/en/opendemocracyuk/to-fix-climate-crisis-we-must-acknowledge-our-imperial-past/

Warikoo, N., Sinclair, S., Fei, J. and Jacoby-Senghor, D. 2016. Examining racial bias in education: A new approach. *Educational Researcher,* Vol. 45, No. 9, pp. 508–514. https://journals.sagepub.com/doi/full/10.3102/0013189X16683408

Zeng, D., Cao, Z. and Neill, D. B. 2021. Artificial intelligence-enabled public health surveillance – from local detection to global epidemic monitoring and control. L. Xing, M. L. Giger and J. K. Min (eds), *Artificial Intelligence in Medicine*, pp. 437–453. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7484813/

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J. and Perrault, R. 2021. *Artificial intelligence index report*. Stanford Institute for Human-Centered Artificial Intelligence. https://aiindex.stanford.edu/report/

PARADOXES OF PARTICIPATION IN INCLUSIVE AI GOVERNANCE: FOUR KEY APPROACHES FROM GLOBAL SOUTH AND CIVIL SOCIETY DISCOURSE

MARIE-THERESE PNG

PhD candidate at the Oxford Internet Institute and Google DeepMind Scholar. She was previously Technology Advisor to the UN Secretary General's High-Level Panel on Digital Cooperation and currently advises organizations on the ethics of large language models and the environmental impacts of information infrastructure.

SGD8 - Decent Work and Economic Growth
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG15 - Life on Land
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

PARADOXES OF PARTICIPATION IN INCLUSIVE AI GOVERNANCE: FOUR KEY APPROACHES FROM GLOBAL SOUTH AND CIVIL SOCIETY DISCOURSE

ABSTRACT

It is estimated that AI could fuel additional economic output of around US$13 trillion by 2030. However, the countries and communities best positioned to profit from this output are those with the most economic power in the Global North, while the costs are carried by those already disadvantaged, disproportionately in the Global South. Inclusive AI governance initiatives aim to address such distributional inequalities but have yet to address the structural issues that underpin these inequalities. In addition, inclusive AI governance initiatives do not prioritize issues particularly relevant to the Global South, such as Western infrastructure and regulatory dominance, exclusionary ownership, cultural or contextual incompatibilities, digital and material extraction, beta-testing, and workers' rights.

This chapter proposes a methodical approach to address these foundational issues that can be adopted by those working towards meaningful inclusion in AI governance as well as in relevant areas of international trade law, intellectual property, technical standards and certification, and human rights. Four key prerequisites to effective inclusive AI governance are suggested: understanding the Global South AI discourse; co-constructing formal roles of Global South civil society, industry and state actors in global AI governance processes; identifying and resolving barriers to Global South participation; and historically contextualizing geopolitical inequalities in AI governance.

INTRODUCTION

In these critically formative stages of AI governance and given the triopoly of AI governance activity by North America, China and Europe, leaders of AI governance initiatives are recognizing their responsibility to ensure that AI deployment and regulation do not lock in intranational and international inequalities. Largely, under the thematics of "AI for good," we see greater efforts towards integrating civil society and Global South stakeholders. This is guided by the logic that metrics, guardrails and protective mechanisms must be defined by those who know and experience the costs, and cannot be adequately defined by those who are distanced from AI risks by dimensions of power and institutional safety (Ulnicane et al., 2020; Milan and Gutiérrez, 2015; Schiff et al., 2021).

Nonetheless, "global governance practices often generate competing social effects, by which inclusionary trends combine with more exclusionary tendencies" (Pouliot and Thérien, 2017), generating the "paradox of participation"- wherein inclusion can exist while structural harms persist, and by which methods that aim to increase citizen participation nonetheless result in yet further establishment domination (Cleaver, 1999; Bliss and Neumann, 2008; Williams, 2004; Ahmed, 2012). This chapter interrogates whether inclusive AI governance initiatives materially benefit those who disproportionately bear the risks of AI systems. It proposes that the purpose of inclusion in AI governance is structural reform—redistributing resources, agenda-setting and decision-making power (Fraser, 2005)—and that beyond inclusion, Global South and civil society actors are sources for alternative governance mechanisms.

This chapter thus proposes a methodical approach summarized in four recommendations which can be adopted as baseline requirements by those working towards meaningful inclusion in AI governance within technical standards organizations, governments, international organizations and industry. Sub-sections examine concrete concerns from the Global South, including natural resource mining, cheap digital labor, funding regimes and Western regulatory dominance.

Recommendation 1: Understand the AI discourse from the Global South (global civil society, state actors, industry actors, public discourse) to ensure a meaningful scoping and integration of Global South demands and goals, and understand the alignments and non-alignments with the standing governance process.

Recommendation 2: Co-construct formal roles for Global South civil society, industry and state actors in global AI governance processes. This is necessary to ensure that the integration of Global South actors is productive rather than performative and achieves the goal of restructuring more robust and comprehensive governance processes.

Recommendation 3: Identify and resolve barriers that prevent Global South actors from accessing structural and infrastructural decision-making power and avoid "paradoxes of participation." This must examine the potential and limitations of Global South actors as well as historic processes of inequality.

Recommendation 4: Contextualize geopolitical inequalities in AI governance within an analysis of power and historic-political dynamics, e.g., precedents of power asymmetries and transnational exclusion in the global governance of other emerging technologies.

DEFINITION OF KEY CONCEPTS

Artificial intelligence

Artificial intelligence (AI) is a broad branch of computer science which aims to, through an array of techniques, build machines capable of performing tasks such as visual perception, speech recognition, decision-making and language translation, that typically require human cognition.

In this chapter, we will comprehensively understand AI as an area of research, an underlying technology for digital products, and an industry. We also understand AI for its political utility and its materiality, meaning its hardware and infrastructure (data centers, graphic processing units (GPUs), etc.), their supply chains and their dependencies on human labor (e.g., data laborers and annotators) (Crawford, 2021).

Global governance

Global governance describes the collaborative development of ethics, policy and regulation for "issues that have become too complex for a single state to address alone," and is "a product of neo-liberal paradigm shifts in international political and economic relations" (Jang et al., 2016). The global governance of AI, as with the governance of other emerging technologies (Ulnicane et al., 2021), involves multi-stakeholder structures to fill in governance gaps, wherein "actors from private and civil society sectors [...] assume authoritative roles previously considered the purview of the State" (Jang et al., 2016).

Global South and Global North

There are distinct differences between how AI is conceptualized in the dominant policy discourse in the Global North and the discourse emergent from the Global South.

Since the end of the Cold War, the Global North has been associated with stable states and economies, whereas the Global South has referred to economically disadvantaged nation-states. A more developed understanding of the Global South casts it as the "deterritorialized geography of capitalism's externalities", where the Global South is not only defined as countries located in the geographic Southern hemisphere, but accounts also "for subjugated peoples within the borders of wealthier countries, such that there are Souths in the geographic North and Norths in the geographic South" (Mahler, 2017). It is also essential to consider how these people disproportionately carry the costs of extraction and exploitation by capitalist economies. Global South perspectives center the displaced priorities, concerns and voices of the global majority, especially in the context of colonial legacies. As Singh and Guzmán (2021) articulate, "we treat 'Global South' as an imperative to focus on cognate lived experiences of the excluded, silenced, and marginalized populations as they contend with data and AI on an everyday basis."

The dominant AI discourse is spearheaded by the "West," or the Global North, i.e., Western Europe and North America, referring to traditionally powerful, industrialized and wealthy state actors, and—within the growing multi-stakeholder governance model—industry, standards-setting organizations and military research and funding bodies. It is also a reproduction of political, epistemic, economic and moral hierarchy developed during European colonization. As Glissant and Dash (1999) put it, "The West is not in the West. It is a project, not a place."

Both the Global South and Global North are heterogeneous. Given the complex plurality of Global South stakeholders, it is crucial to examine the limitations and utility of "Global South" as an analytical framework for present-day power asymmetries and the unequal distribution of AI risks. On one hand, it is a useful unifier for solidarity-building, but on the other it obscures the heterogeneity and internal incongruence of the Global South AI discourse. The Global "Souths" (Connell, 2007; Comaroff and

Comaroff, 2016) represent divergent "political regimes, levels of development, ideologies, and geopolitical interests" (Weiss, 2016) which engender regional contestation and set real limitations on coordination and collective mobilization. The AI discourse from the Souths "operate on a wide spectrum between [the] optimism of leapfrogging and digital transformation of societies on one end and the pessimism of human suffering caused by new forms of data capitalism and colonialism on the other" (Singh, 2021).

Agendas between the Global South and North are not to be seen as inherently dichotomous or antagonistic. There are many countries that "occupy an interstitial position between North and South"—for example within the "Global East" (Müller, 2018). Further, the North/South binary does not account for subjugated peoples within the borders of wealthier countries, and vice versa—"economic Souths in the geographic North and Norths in the geographic South" (Mahler, 2017).

North/South binaries are also blurred by the Chinese government and industry leadership in AI governance and applied research and development (R&D). Powerful tech industry actors include GAFAM (Google, Amazon, Facebook, Apple, Microsoft) in the United States and BATX (Baidu, Alibaba, Tencent, Xiaomi) in China. China's geopolitical, research, production and standards-setting power inevitably shapes dominant discourse and is of great significance to the Global South. As Lee (2019) puts it: "Unless they [developing countries] wish to plunge their people into poverty, they will be forced to negotiate with whichever country supplies most of their AI software—China or the United States—to essentially become that country's economic dependent."

RECOMMENDATION 1: UNDERSTAND THE GLOBAL SOUTH AI DISCOURSE

The meaningful integration of Global South actors—as well as their demands and goals—into AI governance processes requires an adequate understanding of the AI discourse from different countries within the Global South, including activity from civil society, state actors, industry, research institutions and the broader public. This section summarizes a subset of "Southern" trends and contrasts differences between the AI discourse "for and from" the Global South and the dominant AI governance discourse.

Brief overview of the Global South AI discourse

Discourses from the Global South around AI are inherently plural. Many of these discourses expand on long-standing work developing digital infrastructures that align with the needs and concerns of low- and middle-income countries, traditionally marginalized populations, and ecosystems. These considerations are routinely neglected by empowered decision-makers who work within a dominant status quo. "Southern" AI discourses tend to draw from anti-hegemonic practices and engage with the downstream effects of imperial histories, as well as constructive critiques of capitalist structures that scale exploitative, unsustainable, unequal and harmful practices. These AI governance discourses do not originate exclusively from the Global South, but also from institutions and communities in the Global North which are yet to be integrated at scale in international AI governance processes.

There are contrasting thematics between the Global North and Global South around the normative frameworks, issue framing and risk assessment surrounding AI. Given the politics of technology (Winner, 1980), scholarship centering Global Souths or marginalized communities with regard to the harms of AI includes the areas of postcolonial computing (Irani et al., 2010), decolonial computing (Ali, 2016), data extractivism (Couldry and Mejias, 2019; Ricaurte, 2019; Crawford, 2021), culturally sensitive AI and human rights (Mhlambi, 2020; Kak, 2020), data colonialism (Birhane, 2020), Indigenous data sovereignty (Rainie et al., 2019), feminist design practices (d'Ignazio and Klein, 2020),

design justice (Costanza-Chock, 2020) and data justice (Milan and Treré, 2019; Taylor, 2017). Communities for transnational solidarity and collective action are also emerging, including the 2017 AI and Inclusion Conference, Article 19, the Web Foundation, Tierra Común, the Non-Aligned Technologies Movement, the Global Data Justice Project, the Justice Tech Lab, Big Data Sur, Black in AI and the Digital Asia Hub, among others.

Importantly, Global South-centered AI discourses have surfaced the physicality and human labor components of AI—the weight of the cloud, so to speak. Based on their action and advocacy work, Pollicy, a Ugandan institute, succinctly identifies areas which they term "digital extractivism." In addition to natural resource mining and cheap digital labor, they point to "illicit financial flows, data extraction, infrastructure monopolies, digital lending, funding structures, beta testing and platform governance" (Iyer et al., 2021). Given that many of these areas are neglected by the dominant AI governance discourse, the involvement of Global South constituents is highly relevant for more comprehensive risk assessment and governance within AI R&D, infrastructure, supply chains, deployment and regulation.

Concrete concerns from the Global South

This section outlines concrete Southern concerns, including cultural differences, Western infrastructure and regulatory dominance, exclusionary ownership, contextual incompatibilities, sustainability and extraction, beta-testing, and workers' rights.

Western infrastructure, regulation and exclusionary ownership

The production and ownership of technological infrastructure by Global South countries is essential for these countries to accrue gains from AI development (Rayment, 1983; Mbembe and Nuttall, 2004). Global South adoption of "infrastructural and regulatory landscapes and histories of Euro-America" (Raval et al. 2021) and China raise concerns of power consolidation across Western European, North American and Chinese governments and industries. Sampath (2021) outlines critical areas of reform within the dominant AI governance discourse that would mitigate dependency-extraction dynamics between South and North. These areas include "models of manufacturing, procurement, development and pricing of technologies" (Sampath, 2021), first mover advantages in trade, and unequal public-private partnerships.

Intellectual Property (IP) legislation is also critical for the Global South. The South Centre (2020), an intergovernmental organization between 54 developing countries, points to the monopoly of IP protections held by companies and countries in the developed world which preclude the autonomy of Global South countries. Rectifications to such monopolies are highly complex, but initial steps include capacity-building and guidelines development within regional IP offices based on South-South cooperation and developing private-sector incentives that are congruent with the needs of developing countries.

In terms of technology procurement, technology transfer (Gopalakrishnan and Santoro, 2004) and technical assistance have been effective short-term solutions, but do not lend themselves to providing Global South governments and their people with comprehensive access to the economic gains of technologies. Specifically, patent rights can "severely reduce technology transfer since they bring high licensing fees and can thus impede the knowledge adaptation to local conditions" (Kane, 2010).

Analogous patent monopolies exist for the COVID vaccination, where rich industrialized countries such as the United States, the United Kingdom and certain European Union countries have blocked the global scaling of vaccine production, prioritizing pharmaceutical profits through the obstruction of the World Trade Organization Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) which would suspend patents on COVID-19 medical tools, necessary for global herd immunity (Vawda, 2021).

Cultural differences and contextual incompatibilities

The import into the Global South of AI systems designed and trained on datasets collected in Western environments—and thus reflecting specific compositions of demography, identity, kinship, religion, culture, socio-politics, regulation and infrastructure—unsurprisingly results in unintended consequences when applied in novel contexts. A learning algorithm trained on North American datasets is unable to be directly implemented in Central America, Africa or Asia without risking contextual incompatibilities (Neupane and Smith, 2017). These unintended, or rather unanticipated, consequences can be circumvented by co-developing AI technologies and policies with local experts and impacted communities. Participatory data collection (Graham et al., 2015) could increase the need for locally relevant datasets in developing economies (Quinn et al., 2014) and allow for the training of AI models containing "unique features not present in other environments" (Lee et al., 2020).

Areas of concern emerging from the Global North are commonly articulated as "bias and fairness, accountability, transparency, explainable AI, and responsible AI" (Singh, 2021). The meaningful import, translation and relevance of these concepts across socio-political, economic, cultural values, linguistic and infrastructural contexts in the Global South is being investigated and contested (Raval et al. 2021). In his keynote at the African Information Ethics Conference, Pretoria in 2007, Professor Rafael Capurro described an "information ethics for and from Africa" which highlights the monopoly of (heterogeneous) Western ethical traditions in the ethics, governance and regulation of information communication technologies and automated information systems.

The Institute of Electrical and Electronics Engineers' Classical Ethics Committee (IEEE, 2019) posits that Western monopolies adversely affect the development of globally relevant AI standards, which is an "inherently value-laden project, as it designates the normative criteria for inclusion to the global network" (Wong, 2016). The blanket application of Western cultural and political values can result in the "delegitimization of the plausibility of [responsible innovation] based on local values, especially when those values come into conflict with the liberal democratic values" which "do not enable scientists and technology developers to be recognized as members of the global network of research and innovation" (Wong, 2016). As an example, data privacy continues to be viewed through a Western lens (Arora, 2018).

Sustainability and extraction

As Crawford (2021) articulates, "the data economy is premised on maintaining environmental ignorance." While AI can be used to optimize energy use and support development of green technologies, governance frameworks cannot omit the environmental costs of AI and information infrastructure (Parikka, 2015). Environmental concerns are still peripheral within the dominant AI governance discourse, and though they are emerging—such as, for instance, in the Global Partnership on AI's "A Responsible AI Strategy for the Environment" (Clutton-Brock et al., 2021)—miss many concerns raised by the Global South.

Training machine learning models is energy-intensive, and AI systems rely on physical infrastructures (data centers, GPUs, semiconductors), which drives demand for rare earth minerals. In 2020 the European Parliament reported that "the extraction of nickel, cobalt and graphite for use in lithium ion batteries—commonly found in electrical cars and smartphones—has already damaged the environment, and AI will likely increase this demand" (Bird et al., 2020, p. 28; Khakurel et al., 2018). This increased demand for rare earth metals leads to constrained supplies where more complex environments must be accessed, leading to further automation of mining and metal extraction (Khakurel et al., 2018). Furthermore, contracts by Microsoft, Google and Amazon providing tools to the oil and gas industry for extraction optimization (Greenpeace, 2020; Conger et al., 2020) further degrade our environment and ecological services.

In 2021, the Observer Research Foundation hosted a workshop on environmental risks of AI, focusing on risks to marginalized communities and using frameworks of environmental racism. It is well documented that extractive industries first harm racialized, vulnerable and neglected groups through labor exploitation, child labor, state violence against Indigenous communities and increased gender-based violence (Legassick, 1974).

Thus the ethical or responsible deployment and governance of AI systems requires a whole-systems assessment of risks and costs, currently conducted by practitioners, researchers and advocates located in or aligned with the Global South. This ranges from the automation of inequality (Eubanks, 2018; Noble, 2018) to the physical infrastructure of AI systems and their material supply chains (Crawford, 2021).

Workers' rights and ghost work

A whole-systems assessment of risks and costs also recognizes laborers or "ghost workers" (Gray and Suri, 2019) who "annotate the large volumes of data needed to expose the commonsense elements that make the data useful for a chosen task" (Mohamed et al., 2020). Ghost workers, often located in the Global South, are contracted by specialized annotation companies or platforms that indeed provide jobs, but often lack accountability structures or policies to protect people from exploitative industry practices. For instance, employers may withhold remuneration, denying "the rights of workers to safer, dignified working conditions" (Irani and Silberman, 2013), which impacts those who are economically vulnerable, notably in jurisdictions with limited labor laws (Yuan, 2018).

Globalized labor unions and tech worker coalitions such as Turkoptikon and UNI Global Union, or research projects such as Fairwork, are sources of expertise in understanding the harms and risks experienced by the laborers who drive the AI economy. These groups are effective in doing so because they center the knowledge of workers, a necessary precondition for robust guardrails and protective regulations.

Beta-testing

Both ghost work and beta-testing refer to potentially exploitative industry practices. Beta-testing is described as "the testing and fine-tuning of early versions of software systems to help identify issues in their usage in settings with real users and use cases" (Mohamed et al., 2020). There is a well-documented practice of beta-testing technologies where companies outsource product risks to already vulnerable populations. For example, Palantir's deployment of predictive policing in New Orleans or Cambridge Analytica's use of election analytics in the Kenyan and Nigerian elections before employing them in Western democracies (Mohamed et al., 2020). There is a pattern of selecting communities that are systematically less protected or more exposed to risks, or jurisdictions that lack pre-existing safeguards and regulations around data usage, which benefits companies because the mode of testing would violate laws in their home countries (UNCTAD, 2013).

RECOMMENDATION 2: UNDERSTAND AND FORMALIZE THE ROLES OF THE GLOBAL SOUTH AND CIVIL SOCIETY

For Global South actors to be meaningfully integrated into AI governance processes, i.e., co-governance, we must understand and formalize their roles. These roles need to be co-constructed with Global South civil society, industry and state actors and set within reformed structures that ensure this integration into AI governance processes is productive rather than performative.

This chapter identifies four proposed roles for Global South actors:

1. Acting as a challenging function to exclusionary governance mechanisms.
2. Providing legitimate expertise in the interpretation and localization of risks, concerns, demands and issue framing.
3. Providing democratic accountability structures to the state and international governance processes; and
4. Providing a source of alternative governance mechanisms.

A legitimate interpretation and contextualization of risks

Avoidable harms resulting from the implementation of AI systems into any geographic or social context must be defined and framed by those who experience the costs incurred by AI systems. Risks, and appropriate guardrails, cannot be adequately defined by those who are distanced from AI risks by dimensions of power and institutional safety (Ulnicane et al., 2021; Milan and Treré, 2019; Schiff et al., 2021). Global South actors have the legitimate capacity to interpret issues to which they are subject and to advocate for the consideration of risks largely neglected in mainstream AI governance discussions. However, due to emphasis on regional and institutional credibility (state actors or elite institutions largely based in Western Europe, North America and China), civil society and Global South actors are conferred less legitimacy, visibility and influence.

The practice of centering the knowledge of those most vulnerable to risks has been long developed by Participatory Action Research and Critical Development Studies; for instance, this is the case for product risk assessments carried out with impacted communities. The practice is well summarized by the slogan "nothing about us without us," originating from Central European political traditions (Smogorzewski, 1938) and later adopted in the disability rights movement around the development of innovative technologies (Werner and PROJIMO, 1998).

Understanding the legitimate expertise of impacted groups is also essential to counterbalancing the dominant AI governance discourse's tendency to universalize notions of harms in ways that may not be applicable to different cultures, regions and jurisdictions. Hence the proposition of comparable rather than universal global standards within the coordination of global responses. Regions, states and cities "must be able to respond to the specific social, economic, and cultural demands of their citizens" (Abdala at al., 2020). "Universalization" often entails hegemonic, and locally incompatible, impositions particular to information-mature economies in North America and Western Europe (Mignolo, 2012). In the context of digital privacy, these limitations are highlighted by Arora (2018): "As technology companies expand their reach worldwide, the notion of privacy continues to be viewed through an ethnocentric lens. It disproportionately draws from empirical evidence on Western-based, white, and middle-class demographics." She, among others, argues for "Southern" perspectives, where privacy regulation "dignifies those at the margins, by giving their privacy its contextual integrity."

A challenging function and a source of democratic accountability

The role of the Global South within the AI governance discourse is, at the very least, a challenging function to exclusionary governance and legal processes at an institutional level which neglect or harm marginalized communities (Marchetti, 2016). Challenges and interventions emerge from state actors and from a growing "political economy of resistance" led by civil society activism (Taylor, 2017; Torres, 2017; Milan and Treré, 2019). "Data activism" describes "new forms of political participation and civil engagement in the age of datafication" that aim to truly mitigate avoidable harms from AI systems (Milan and Velden, 2016).

As a "nongovernmental and noncommercial space of association and communication" (Jaeger, 2007), civil society is well positioned as an accountability structure to states and global governance bodies. Governance bodies often "lack formal mechanisms of democratic accountability that are found in states"; instead, "executive councils of global regulatory bodies are mainly composed of bureaucrats who are far removed from the situations that are directly affected by the decisions they take," illustrating the inaccessible and opaque nature of dominant AI global governance processes (McGlinchey et al., 2017). Civil society at the international level "is predominantly focused on building political frameworks with embedded democratic accountability," upholding, reinforcing and applying reformist pressure to legislation, regulatory guardrails and rights-based frameworks. Civil society has also played the role of "broadcasting" and cooperatively reinforcing concrete concerns from the Global South (Marchetti, 2016) that are excluded from the transnational AI governance agenda.

Global civil society has ensured accountability by increasing the public transparency of global governance operations, monitoring and reviewing global policies, seeking redress for mistakes and harms attributable to global regulatory bodies, and advancing the creation of formal accountability mechanisms for global governance (Scholte, 2004). Given that all sectors of society benefit from this work, global AI governance should direct resources and compensation to such work, which is often under-resourced and whose producers are often asked to participate in consultations on a voluntary basis (McGlinchey et al., 2021).

The Tech Equity Coalition (TEC) by ACLU Washington is an example of how civil society has provided democratic accountability, protecting the rights and civil liberties of marginalized communities in the face of increasingly powerful technologies. The TEC utilizes policy, research, organizing and litigation. Where existing laws are unjust or violated, strategic litigation is used, relying on courts, legislatures and communities to ensure the law is upheld by private actors or government-funded agencies. The TEC also co-develops policy proposals with directly affected communities, advocates for community-centered policy and laws that create safeguards around AI technologies and data and, where necessary, advocates for the cessation of the use of evidently harmful technologies—for example, surveillance as a tool for over-policing.

A source of alternative governance mechanisms

There is a distinct call from the Global South and civil society to identify new forms of political organization, political objectives and action repertoires in order to safely and equitably govern AI technologies (Milan and Velden, 2016). Before pushing for the representation of Global South actors within AI governance processes, we must first interrogate the effectiveness of these processes in achieving this goal, as dominant AI governance processes may simply not have the structural capacity to do so.

> We need to change the institutions that have historically been set up as tools of advancement and control for some, to the exclusion of many, and not just tweak it or look for ways to create space. Without this, advancement of science and technology will continue to benefit those who governed historically at the expense of those who were excluded. (Sampath, 2021)

Future designs of AI technologies, and their governance, are being re-thought by Indigenous-led groups, such as the Global Indigenous Data Alliance, and others aligned with the Global South. These calls for reformist or alternative governance mechanisms understand how administrative, cultural, economic and epistemological legacies of European colonialism are integrated within global governance (Quijano, 2000; Sampath, 2021).

South-South cooperation *fora* such as the UN South-South Initiative, Group of 77 and Non-Aligned Movement, which were key in decolonization and independence movements, are examples of alternative governance mechanisms. Although they are state-led, and grapple with state-civil society tensions, these bodies act as a platform for Global South countries to build and protect collective interests, promote South-South geopolitical cooperation, assert self-determination through multilateral action, and engage in capacity-building in ways that erode dynamics of dependency on states and industry in the Global North (Weiss, 2016).

RECOMMENDATION 3: IDENTIFY AND RESOLVE BARRIERS TO GLOBAL SOUTH PARTICIPATION

Taxonomy of traps: Barriers to inclusive AI governance

During the Chatham House Inclusive AI Governance Seminar launching their 2021 Inclusive Governance Report, public attendees were asked what the greatest barriers to inclusive participation in global governance were. Crowdsourced answers included: "geopolitics, funding, capacity, power imbalances, elites, democracy deficit, language, lack of mechanisms for inclusion." Issues such as "poor national governance, neocolonial mindset, covert lobbying, fragmented efforts, mistrust, securitization, legal barriers, racism, rigid mechanisms, and digital divide" were also raised (Chatham House, 2021). Many of these publicly visible issues are not yet integrated into mainstream inclusive AI governance efforts.

Efforts to address systematic transnational and cross-sectoral exclusion in the global governance of emerging technologies are not unprecedented (Ulnicane et. al, 2021). To responsibly mitigate exclusionary dynamics and their harmful downstream effects, and to ensure that the benefits of AI are equitably distributed, it is imperative to understand the systemic barriers or "traps" preventing Global South civil society actors from accessing different forms of power. These traps include, but are not limited to, organizational culture, exclusionary normative logics, use of broad language, insufficient technological literacy, co-option, and adverse financial incentives. The usage of "traps" here is borrowed from Selbst et al. (2020), who define them as "failure modes that result from failing to properly account for or understand the interactions between technical systems and social worlds."

Organizational culture

Schiff et al. (2021) note that "researchers would do well to focus on the organizational or sectoral contexts that are shaping AI ethics, and how these contexts might guide ethical priorities and actions." AI governance processes exhibit exclusionary organizational structures and norms, reflecting wider interpersonal, geopolitical and historic power inequalities (Wilson, 2000). Organizational incentives shaping AI governance—such as "competitive advantage, strategic planning, strategic intervention, signaling social responsibility, signaling leadership" (Schiff et al., 2020)—can be in tension with inclusion. Funding structures, and their evaluation metrics, also create individual incentives which can be tied to careerism and political opportunism.

AI governance initiatives are often opaque and invitation-based and predicated on institutional credibility and language compatibility (usually English), and favor agendas set by powerful state or industry actors. Even once invited, Global South or civil society actors face "institutional filters" (McGlinchey, 2021) and are rendered peripheral to end-point decision-making. New institutional structures are also "continually emerging and the challenge in terms of integration is therefore endlessly renewed" (Marchetti, 2016), requiring additional effort and resources in order to adapt.

Normative logics

Closely related to organizational culture are the normative logics embedded in institutions and bureaucracies and shaping governance processes. These logics prescribe ways a person ought or ought not to reason, and what political positions are appropriate within a social space. Within AI governance, we see the replication of interventionist impulses and deficit models of the Global South (Latonero, 2019; Vinuesa et al., 2020), including the conflation of resource capacity with the inherent capability of marginalized stakeholders. Though Global South stakeholders do indeed exhibit systemic failures, Taylor (2019) explains that "a root cause of failure of developmental projects lies in default attitudes of paternalism, technological solutionism and predatory inclusion."

Diplomatic norms within global AI governance serve as a necessary friction-minimizer between stakeholders and incentivize performative and hierarchical behavior. What is considered diplomatic and polite "often reflects existing power structures, and reinforces existing patterns of interaction [...] Politeness generally reflects and favors the dominant power structures" (Roberts, 2018). If Global South and civil society actors serve the function of challenging AI governance processes, it is no surprise that challenges are characterized as non-diplomatic, even rude, since they interrupt dominant norms. Diplomatic norms can obfuscate power dynamics and create justifications for excluding those who, by advocating for the marginalized, "rock the boat" (Kazmi, 2012; McConnell et. al., 2012). In order to survive, public-interest organizations must often operate "as a subsystem of world politics rather than opposing the system from outside" (Jaeger, 2007).

Use of broad language

Broad language in AI governance is increasingly under scrutiny. Looking at "AI for good" initiatives, Green (2019) argues that "good isn't good enough," referring to limited and vague definitions of what "social good" means. Both socio-cultural and technical literacy are essential to creating useful definitions.

Comparing private, NGO and public-sector engagement with ethical issues of AI, Schiff et al. (2021) identify that the NGOs' and the public sector's strategies have "more ethical breadth in the number of topics covered, are more engaged with law and regulation, and are generated through processes that are more participatory." These more specific strategies attend to the social and ethical-political impacts of AI. In contrast, the broadness of language adopted by private entities or international organizations allows interpretation that conveniently protects the interests of richer governments and industry actors. As with logics of diplomacy, broad language favors the dominant power structure, in contrast to the socio-political specificity demanded by more critical or radical approaches.

For example, the High-Level Expert Group on Artificial Intelligence (AI HLEG) recently published some of the most tractable draft laws to date for AI regulation. Members of civil society have identified that these draft laws do not meet standards of fundamental digital rights protection, and house loopholes which leave citizens exposed to misuse and malicious uses of AI. The broad nature of draft laws leaves a large range of discretion for the technology industry to regulate itself: "many industry groups expressed relief the regulations were not more stringent, while civil society groups said they should have gone further" (Satariano, 2021).

Co-option/Recuperation

The use of seemingly agreed-upon terms such as "human rights," "sustainability," "interdependence," "sovereignty," or even "inclusion" and "participation" can vary drastically between stakeholder groups. These terms also risk co-option by vested interests (Ulnicane et al, 2020) and redefinition by actors

to serve the priorities of the private sector or richer countries; these actors are conferred "power to do so on the grounds of their elite status, specialist knowledge, or potential ability to threaten essential commitments or goals" (Selznick, 2015).

Co-option is used alongside the term "recuperation" to describe "the process by which politically radical ideas and images are twisted, co-opted, absorbed, defused, incorporated, annexed or commodified [...] interpreted through a neutralized, innocuous or more socially conventional perspective" (Downing et al., 2001). Co-option is countered by its revolutionary counterpart, "*détournement*"—"a subversive plagiarism" (Downing et al., 2001) entailing a process of turning expressions of dominant systems against themselves and from their usual purpose, in service of protest or mobilization.

Participation

There are many instances where inclusion is procedural and numeric, utilized for beneficent marketing, virtue signaling, or "optical inclusion." Inclusion and participation can certainly exist while the structural enablement of harm persists, with "little evidence of the long-term effectiveness of participation in materially improving the conditions of the most vulnerable people or as a strategy for social change" (Cleaver, 1999). This dual reality, as explained earlier, is described as the "paradox of participation" (Cleaver, 1999; Bliss and Neumann, 2008; Williams, 2004; Ahmed, 2012).

It is necessary, therefore, to understand that the purpose of inclusion is for structural reform—redistributing resource allocation, agenda-setting and decision-making power (Fraser, 2005). Based on the spectrum in inclusion-exclusion dynamics provided by Marchetti (2016), which covers ostracization, exclusion, co-option, inclusion and integration—to which critical views would add structural reform and alterity—inclusion is only the first positive step away from harmful exclusion.

Interdependence

The term "interdependence" has grown in popularity over recent decades to describe "international cooperation in the face of an increasingly complex and globalizing world order" (Keohane and Nye, 1977; Coate et al., 2015). The 2020 UN Secretary-General Roadmap for Digital Cooperation locates us in the "Age of Digital Interdependence." But who is benefitting from these interdependencies? It is clear that the dominant discourse does not adequately address power asymmetries within existing interdependencies. Economies of the Global North depend on the continued extraction of natural resources and labor from the Global South, and Global South dependencies on the Global North are systematized across consumer goods, digital infrastructure, trade, financial regulation and more.

Human rights

Latonero (2018) states that "in order for AI to benefit the common good, at the very least its design and deployment should avoid harms to fundamental human values. International human rights provide a robust and global formulation of those values." These include the rights to social security, work, freedom of expression, privacy and social security, and the right against discrimination (Arun, 2020), implemented through policy or at a software or hardware level.

In practice, however, civil society organizations highlight the long-standing co-option of human rights frameworks (Peck, 2011), as well as questioning who in practice has the right to human rights; these concerns have carried over into the discourse on AI and human rights. The integration of human rights into the development and deployment of AI systems by both governments and private companies has been in large part lip service. Profit incentives and international market competition come at the cost of individual and collective rights.

In the context of AI technologies being largely developed and regulated by the Global North and exported to the Global South, strong consideration should be given to the "pitfalls associated with human rights, particularly focusing on the criticism that these rights may be too Western, too individualistic, too narrow in scope and too abstract to form the basis of sound AI governance" (Smuha, 2020). These critiques are expressed generatively by work such as Mhlambi's (2020) *Ubuntu as an Ethical and Human Rights Framework for AI Governance*.

Financial incentives

Tensions between prosocial (people-oriented) and economic (profit-oriented) goals (Schiff et al., 2021) result in asymmetrical dynamics between the Global North and South. As such, policymakers should be astute about the ways funding shapes governance agendas and about participatory versus exclusionary dynamics. Current AI governance initiatives appear to reward the centralization of power, which Gurumurthy (2021) describes as "hegemonic discourses of AI that serve neoliberal capitalism."

The ability for a civil society organization to participate meaningfully in AI governance processes is contingent on resource availability (financial, personnel, political capital, time, expertise, networks), access or visibility to governance networks, and the ability to negotiate norms, language, protocols and priorities under an imbalanced power dynamic (Marchetti, 2016; Milan and Gutiérrez, 2015).

Within governance spaces, notably in the UN, phrasing such as "the future of multilateralism is multi-stakeholderism" promotes the inclusion of non-state stakeholders in global governance processes. Corporate actors assuredly have a legitimate role in AI governance processes. Much of AI governance is internal within companies, and their AI products are procured by governments in both the Global North and South. The outcome of multi-stakeholderism, however, has been corporate stakeholders gaining more political influence and attraction to global governance processes, e.g., Microsoft's office within the United Nations. As a United Nations Department of Economic and Social Affairs report identifies, "An important force shaping governance at national and international levels is big corporations, which lobby for laws and policies that serve their interests" (UN DESA, 2014).

Big Tech also influences scholarship and academic advisory used by AI governance. As Abdala et al. (2020) highlight, "Big Tech can actively distort the academic landscape to suit its needs." Strategies include influencing "decisions made by funded universities," "the research questions and plans of individual scientists" (Abdala et al., 2020), and manipulating academia to avoid regulation (Ochigame, 2019).

AI regulation is outsourced to privatized standards bodies such as the European Committee for Electrotechnical Standardization, the IEEE, *Verband Deutscher Elektrotechniker*, the International Organization for Standardization and others. These bodies also receive heavy lobbying from industry, which can push them to "significantly drift from essential requirements" (Veale and Borgesius, 2021) such as human rights. Civil society organizations, with their empirical understanding of human rights issues on the ground, are essential to standards-setting where human rights expertise is severely lacking (Cath, 2020; ten Oever and Cath, 2017). Again, notable barriers to civil society participation in standards setting include funding; not only are their goals often incommensurable, but Big Tech's resources far outweigh those of civil society and Global South stakeholders.

Global South and civil society limitations

It is not enough to identify barriers to inclusive AI governance within dominant AI governance *fora*. Though Global South and civil society actors are fundamental to the protection of civil and human rights and to accountability in AI governance, they also face limitations to their ability to effectively mitigate harms. These limitations include infrastructure and connectivity constraints, lack of technological literacy, the depoliticization of civil society, and tensions between governments and civil society.

The benefits of civil society as an accountability mechanism remain highly contextual (Grimes, 2008) and should not be understood as a given but instead actively worked towards as part of robust and effective protective governance. Citing Gramsci, Sassoon (2014) reminds us that "above all [...] we must not idealize civil society," and that we must acknowledge the heterogeneity of goals, priorities, and incentives civil society organizations follow. Beyond institutionalized civil society, there are also social movements, activist groups, unregistered grassroots organizations and so forth, which are closer to the impact of automation on rights and equality.

The "blunting of political goals" to survive within elite governance spaces is described as "depoliticization" (Jaeger, 2007), who outlines a "double movement" wherein civil society organizations perform both roles of "depoliticization as much as politicization." Organizations operate "inside as well as outside the political system of world society" (Jaeger, 2007), within "the established order" and sometimes as a "counter-hegemonic bloc" (Katz, 2006).

Further, narratives describing civil society's co-optation by dominant AI processes are incomplete. First, "these models risk adopting unduly simplistic assumptions of passive victimhood on the part of institutions within liberal democratic societies" (Pils, 2019). Second, civil society demonstrates agency in varied ways, by contesting (resisting, dismantling, building solidarity), collaborating, complying or being complicit with overarching financial or political power. It is therefore important to collect "subtle accounts of representation and [...] highlight contestatory practices" (Dryzek, 2012) specific to AI governance, so that evidence-based redesign can ensure effective protections for the most vulnerable.

It is important to remember that the inclusion of Global South governments in AI governance processes will not always materially benefit the broader population of a country, especially the populations that are the most vulnerable within the context of growing intranational inequalities. Tensions between civil society and states or governments are well understood. It is reductionist to assume that Global South state actors and civil society organizations are aligned, that their goals can be conflated, or that all Global South actors adopt postcolonial praxis or narratives. State repression and violence exist in countries of both the Global North and South. Though central, the achievement of self-determination and restructuring for equitable distribution of benefits (trade, ownership, geopolitical influence and so on) at an international level does not resolve domestic issues contested by civil society. For example, in Kenya, the Nubian Rights Forum and the Kenya Human Rights Commission initiated a successful court case contesting the government's National Integrated Identity Management System, on the basis that it violated "the right to privacy, equality, and non-discrimination enshrined in Kenya's constitution" (Mahmoud, 2019).

RECOMMENDATION 4: HISTORICALLY CONTEXTUALIZE POWER IMBALANCES IN AI GOVERNANCE

Historic power imbalances

The traps of inclusive governance outlined in Section 3.1—such as organizational culture, normative logics, the use of broad language, co-option and financial incentives—exist within wider institutional and geopolitical systems of power and inequality. Geopolitical inequalities within AI governance, such as the discourse being led out of the Global North (within the EU and tech companies out of Europe and North America) and China are historically contingent. Thus, the work to increase the representation of Global South stakeholders must be also embedded within a "broader analysis of power and political dynamics or tensions" (ÓhÉigeartaigh et al., 2020). By looking more closely at the repetitive dialectic

of inclusion and exclusion, we can "better understand the politics of global public policy making, including its power dynamics" (Pouliot and Thérien, 2017). To do so, we must ask how these have become systematized over time.

Coloniality of power in AI governance

We cannot understand present AI inequalities, or anticipate their futures, without looking at their historic trajectories. The first-mover advantages and exclusionary path dependencies we see today are, in part, living relics from our colonial histories. The lack of deeper engagement with the historic roots of exclusion within AI governance spaces is explained in part, according to Sampath (2021), by "techno-centric explanations of progress and industrialization" which "are deeply entrenched in a wider social context that encourages us to ignore the historical roots of current inequalities—which, in fact, are not amenable to a technological solution alone."

The contemporary remnants of European colonialism in contemporary geopolitical and interpersonal power inequalities are described by the concept of "coloniality" (Quijano, 2000). The coloniality of power is a necessary framework for understanding the distribution of harms and benefits of AI systems across the Global South and North, and is central to emerging scholarship around data colonialism and data capitalism that recognizes continuities of colonial exploitation, extraction and dispossession in the use of labor, material resources and data in AI industries (Thatcher et al., 2016; Ricaurte, 2019; Couldry and Mejias, 2019; Birhane, 2020; Zuboff, 2019; Irani et al., 2010; Ali, 2016).

The "Fourth Industrial Revolution" (4IR)

McKinsey has estimated that AI could fuel "additional economic output of around US$13 trillion by 2030, increasing global GDP by about 1.2% annually" (Bughin et al., 2018). Though the digital and AI economy may indeed benefit the Global South, trade and infrastructural initiatives or partnerships under the 4IR do not adequately recognize how first-mover advantages and exclusionary path dependencies persist, as a continuous pattern of historic power imbalances. "Those best positioned to profit from the proliferation of artificial intelligence (AI) systems are those with the most economic power" (Chan et al., 2021). The economic benefits of AI sometimes fit with "hyperbolic claims that big data and the data economy are the new 'frontier of innovation,' with 'cost-effective,' 'profit-generating' properties for all" (Sampath, 2021), and do not recognize the selective enrichment, or "Matthew effect" (Fernández-Villaverde et al., 2021), which the capitalist economic system supports along the boundaries of Global North and South.

The First Industrial Revolution was driven by the extraction and exploitation of labor, knowledge and natural resources of European colonies, which was made possible by the military-led efforts by Western European and consequently North American colonial governments. Colonial regimes were structured around unequal legal, political, trade and racialized systems which persist today. It is particularly pertinent to note that the use of "Industrial Revolution" and "Global South" terminology points directly to colonial histories, yet colonial histories are not acknowledged in dominant AI governance discourses, in part because their impacts do not overtly damage rich industrialized countries.

Ghost work and beta testing are practices that articulate historic continuities of extraction and exploitation between ex-colonial states and ex-colonies (Mosco and Wasko, 1988; Agrawal et al., 2019; Keskin and Kiggins, 2021). Beyond shaping the treatment of workers and exposing marginalized populations to the risks of beta-testing, the capitalist mode of production perpetuates the North-South economic divide (Arrighi, 2008) in a way that needs to be acknowledged and dialectically grappled with in discussions of North-South inequalities. We can understand the Fourth Industrial Revolution as bringing tangible benefits (in healthcare, communications, agriculture, labor market and education) as well as techno-imperialism (Sampath, 2021), racialized capitalism and surveillance capitalism.

Historicizing sovereignty

"Now is the time for Europe to be digitally sovereign," states a joint letter authored to the European Commission by European political leadership (ERR News, 2021). Digital sovereignty, here, refers to data ownership, usage and storage, and "increasing Europe's technological capacity and its ability to establish values and rules in a technology-centered world that is becoming dominated by other countries" (European Union, 2021).

Demands for digital sovereignty are also coming from the Global South, with emerging economies wanting to benefit from their own data. In Africa, for example, critical infrastructure (submarine cables, terrestrial fiber-optic networks, and data centers) are largely owned by non-African telecom companies. Sensitive population data are largely hosted on servers abroad, such as Ireland, given that many African countries do not have national data centers. Initiatives such as Smart Africa and the African Tax Administration Forum are developing privacy and taxation policies to mitigate tech giants' unfettered access to and monetization of national data to the detriment of growing local data economies (Velluet, 2021; Elmi, 2020).

Movements concerned with the collection, ownership and application of Indigenous data such as the Indigenous Data Sovereignty movement advocate for the "right of Indigenous peoples to control data from and about their communities and lands, articulating both individual and collective rights to data access and to privacy" (Rainie et al., 2019).

Given this variability of notions of sovereignty, understanding which are legitimized in AI governance discourse and which are sidelined (and why) is crucial. The defense of territorial and digital self-determination differs greatly between European, African and Indigenous governance, and has been significantly shaped by European colonialism. Kovacs and Ranganathan (2019) thus caution against any uncritical operationalization of sovereignty and remind us that "it is important to ask under what conditions it becomes possible to reclaim sovereignty despite these violent roots."

Modern concepts of sovereignty are argued to be grounded in the 17th-century Peace of Westphalia treaties, "when a new political order was recognized" and peace achieved for the Holy Roman Empire after longstanding violence (de Graaf and Kampmann, 2018). De Graaf and Kampmann also argue that this prompted economic and technological development and seeded European "security culture." Reduced neighboring threats also made it possible for European imperial expansion based in a hierarchical and racialized international system. We see this hierarchy play out in the continued extraction, exploitation and dispossession of the Global South in the modern data economy—cheap labor, illicit financial flows, data extraction, natural resource mining, infrastructure monopolies, funding structures, beta testing and more (Iyer et al., 2021; Birhane, 2020).

CONCLUSION

> AI governance initiatives seeking to integrate civil society and Global South stakeholders in order to materially mitigate unequal risk distribution must earnestly examine historic-political exclusionary trends, better acquaint themselves with existing work led by the Global South and civil society, understand internal barriers or traps to meaningful inclusion, and move beyond the paradox of participation. Effective inclusion requires the structural redistribution of power, which current governance institutions are not incentivized towards. The just distribution of transformational AI benefits is only possible with alternative epistemological, development and governance models from the South. A shift towards participatory co-governance is necessary in order to mitigate "new power asymmetries [...] and more drastic degrees of exclusion" (Sampath, 2021).

REFERENCES

Abdala, M. B., Ortega, A. and Pomares, J. 2020. Managing the transition to a multi-stakeholder artificial intelligence governance. G20 Insights. https://www.g20-insights.org/policy_briefs/managing-the-transition-to-a-multi-stakeholder-artificial-intelligence-governance/

Agrawal, A., Gans, J. and Goldfarb, A. 2019. Artificial intelligence: The ambiguous labor market impact of automating prediction. *National Bureau of Economic Research*. Vol. 3, No. 2, pp. 31–50.

Ahmed, S. 2012. *On Being Included: Racism and Diversity in Institutional Life*. Durham, NC: Duke University Press.

Ali, S. M. 2016. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students*. Vol. 22, No. 4, pp. 16–21.

Arora, P. 2018. Decolonizing privacy studies. *Television & New Media*. Vol. 20, No. 4, pp. 366–378.

Arrighi, G. 2008. Historical perspectives on states, markets and capitalism, East and West. *The Asia-Pacific Journal*, Vol. 6, No. 1.

Arun, C. 2020. AI and the Global South. Dub, Pasquale and Das (eds.). *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press, pp. 587–606.

Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E. and Weitkamp, A. 2020. *The Ethics of Artificial Intelligence: Issues and Initiatives*. European Parliament Scientific Foresight Unit. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

Birhane, A. 2020. Algorithmic colonization of Africa. *SCRIPT-ed*, Vol. 17, No. 2, pp. 389–409.

Bliss, F. and Neumann, S. 2008. Participation in international development discourse and practice: "State of the art" and challenges. *INEF-Report*. Vol. 94. https://www.participatorymethods.org/sites/participatorymethods.org/files/Participation%20in%20International%20Discourse%20and%20Practice.pdf

Bughin, J., Seong, J., Manyika, J., Chui, M. and Joshi, R. 2018. *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy*. McKinsey Global Institute.

Capurro, R. 2009. Intercultural information ethics: Foundations and applications. *Signo y Pensamiento*, Vol. 28, No. 55, pp. 66–79. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-48232009000200004

Cath, C. 2020. What's wrong with loud men talking loudly? The IETF's culture wars. https://hackcur.io/whats-wrong-with-loud-men-talking-loudly-the-ietfs-culture-wars/

Chan, A., Okolo, C. T., Terner, Z. and Wang, A. 2021. The limits of global inclusion in AI development. *arXiv:2102.01265 [cs]*. https://arxiv.org/abs/2102.01265

Chatham House 2021. Reflections on building more inclusive global governance. https://www.chathamhouse.org/sites/default/files/2021-04/2021-04-15-reflections-building-inclusive-global-governance.pdf

Cleaver, F. 1999. Paradoxes of participation: Questioning participatory approaches to development. *Journal of International Development*, Vol. 11, pp. 597–612. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.461.2819

Clutton-Brock, P., Rolnick, D., Donti, P.L. and Kaack, L.H. 2021. *Climate Change and AI*. Global Partnership on AI Report.

Coate, R. A., Griffin, J. A. and Elliott-Gower, S. 2015. Interdependence in international organization and global governance. *Oxford Research Encyclopedia of International Studies*.

Comaroff, J. and Comaroff, J. L. 2016. *Theory from the South, or, How Euro-America is Evolving Toward Africa*. London; New York: Routledge, Taylor & Francis Group.

Conger, R., Robinson, H. and Sellschop, R. 2020. Inside a mining company's AI transformation. https://www.mckinsey.com/industries/metals-and-mining/how-we-help-clients/inside-a-mining-companys-ai-transformation

Connell, R. 2007. The Northern theory of globalization. *Sociological Theory*, Vol. 25, No. 4, pp. 368–385.

Costanza-Chock, S. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: MIT Press.

Couldry, N. and Mejias, U. A. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford, CA: Stanford University Press.

Crawford, K. 2021. *Atlas of AI*. New Haven: Yale University Press.

de Graaf, B. and Kampmann, C. 2018. The Peace of Westphalia also had its dark side. Press release of the Religion and Politics Cluster of Excellence from 19 September 2018. University of Münster. https://www.uni-muenster.de/Religion-und-Politik/en/aktuelles/2018/sep/PM_Westfaelischer_Frieden_hatte_auch_Schattenseiten.html

d'Ignazio, C. and Klein, L. F. 2020. *Data Feminism*. Cambridge: MIT Press, pp. 97–123.

Downing, J. 2001. *Radical media: rebellious communication and social movements*. Thousand Oaks, Calif.: Sage Publications.

Dryzek, J. S. 2012. Global civil society: The progress of post-Westphalian politics. *Annual Review of Political Science*, Vol. 15, No. 1, pp. 101–19. https://doi.org/10.1146/annurev-polisci-042010-164946

Elmi, N. 2020. Is Big Tech setting Africa back? *Foreign Policy*. https://foreignpolicy.com/2020/11/11/is-big-tech-setting-africa-back/

ERR News 2021. Estonia, EU countries propose faster "European digital sovereignty." *ERR News*. https://news.err.ee/1608127618/estonia-eu-countries-propose-faster-european-digital-sovereignty

Eubanks, V. 2019. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press.

European Union. 2021. European Digital Sovereignty Conference. https://eu-digitalsovereignty.com/

Fernández-Villaverde, J., Mandelman, F., Yu, Y. and Zanetti, F. 2021. The "Matthew effect" and market concentration: Search complementarities and monopsony power. CAMA Working Paper No. 22/2021. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787787

Fraser, N. 2005. Reframing justice in a globalizing world. *New Left Review* No. 36.

Glissant, É. and Dash, J. M. 1999. *Caribbean Discourse: Selected Essays*. Charlottesville: University Press of Virginia.

Gopalakrishnan, S. and Santoro, M. D. 2004. Distinguishing between knowledge transfer and technology transfer activities: The role of key organizational factors. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1495508

Graham, M., de Sabbata, S. and Zook, M. A. 2015. Towards a study of information geographies: (Im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, Vol. 2, No. 1, pp. 88–105.

Gray, M. and Suri, S. 2019. *Ghost Work: How How to Stop Silicon Valley from Building a New Global Underclass.* New York: Houghton Mifflin Harcourt.

Green, B. 2019. "Good" isn't good enough. In Proceedings of the NeurIPS Joint Workshop on AI for Social Good, Vancouver, Canada.

Greenpeace. 2020. *Oil in the Cloud How Tech Companies are Helping Big Oil Profit from Climate Destruction.* https://www.greenpeace.org/usa/reports/oil-in-the-cloud/

Grimes, M. 2008. Contestation or Complicity: Civil Society as Antidote or Accessory to Political Corruption. https://www.semanticscholar.org/paper/Contestation-or-Complicity%3A-Civil-Society-as-or-to-Grimes/45e98ef85f3abad31de4f473f522de7154cf849d

Gurumurthy, A. 2020. How to make AI work for people and planet. *openDemocracy*. https://www.opendemocracy.net/en/oureconomy/how-make-ai-work-people-and-planet/

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. Classical Ethics in A/IS. https://standards.ieee.org/industry-connections/ec/autonomous-systems/

Irani, L. and Silberman, S. 2013. Turkopticon: Interrupting worker invisibility in Amazon mechanical turk. CHI 2013: Changing Perspectives, Paris, France. http://crowdsourcing-class.org/readings/downloads/ethics/turkopticon.pdf

Irani, L., Vertesi, J., Dourish, P., Philip, K. and Grinter, R.E. 2010. Postcolonial computing. Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010.

Iyer, N., Achieng, G., Borokini, F. and Ludger, U. 2021. Automated Imperialism, Expansionist Dreams: Exploring Digital Extractivism in Africa. Pollicy. https://archive.pollicy.org/wp-content/uploads/2021/06/Automated-Imperialism-Expansionist-Dreams-Exploring-Digital-Extractivism-in-Africa.pdf

Jaeger, H.-M. 2007. "Global civil society" and the political depoliticization of global governance. *International Political Sociology*, Vol. 1, No. 3, pp. 257–277.

Jang, J., McSparren, J. and Rashchupkina, Y. 2016. Global governance: Present and future. Palgrave Communications, Vol. 2, article 15045. https://www.nature.com/articles/palcomms201545

Kak, A. 2020. The Global South is everywhere, but also always somewhere. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Kane, C. 2010. The relationship between IP, technology transfer, and development. *Intellectual Property Watch.* August 30.

Katz, H. 2006. Gramsci, hegemony, and global civil society networks. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, Vol. 17, No. 4, pp. 333–348. https://www.jstor.org/stable/27928042

Kazmi, Z. 2012. Polite anarchy and diplomacy. *Palgrave Macmillan History of International Thought*. New York: Palgrave Macmillan. https://link.springer.com/chapter/10.1057%2F9781137028136_7

Keohane, R.O., and Nye, J. S. 1977. *Power and Interdependence.* Boston: Little Brown, Cop.

Keskin, T., and Kiggins, R., D. 2021. *Towards an International Political Economy of Artificial Intelligence.* Cham, Switzerland: Palgrave Macmillan.

Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A. and Zhang, W. 2018. The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies* Vol. 6, No. 4, article 100. https://doi.org/10.3390/technologies6040100

Kovacs, A. and Ranganathan, N. 2019. Data sovereignty, of whom? Limits and suitability of sovereignty frameworks for data in India. Data Governance Network, Working Paper No. 03.

Latonero, M. 2019. Stop surveillance humanitarianism. *New York Times*, June 11. https://www.nytimes.com/2019/07/11/opinion/data-humanitarian-aid.html

Lee, J., Gamundani and Stinckwich, S. 2020. Blog: The Inaugural AI Expert Consultation Meeting recap: What's next for AI in Africa? United Nations University Institute in Macau, China. https://cs.unu.edu/news/news/ai-expert-consultation-meeting.html

Lee, K.-F. 2019. *AI Superpowers: China, Silicon Valley, and the New World Order.* New York: HarperCollins.

Legassick, M. 1974. South Africa: Capital accumulation and violence. *Economy and Society*, Vol. 3, No. 3, pp. 253–291.

Mahler, A. G. 2017. Global South. Oxford Bibliographies Online Datasets.

Mahmoud, M. 2019. Stopping the digital ID register in Kenya – A stand against discrimination. *Namati* (blog). April 25. https://namati.org/news-stories/stopping-the-digital-id-register-in-kenya-a-stand-against-discrimination/

Marchetti, R. 2016. Global civil society. Excerpt from *International Relations*. Bristol: E-International Relations. https://www.e-ir.info/2016/12/28/global-civil-society/

Mbembe, A. and Nuttall, S. 2004. Writing the world from an African metropolis. *Public Culture*, Vol. 16, No. 3, pp. 347–372.

McConnell, F., Moreau, T. and Dittmer, J. 2012. Mimicking state diplomacy: The legitimizing strategies of unofficial diplomacies. *Geoforum*, Vol. 43, No. 4, pp. 804–814.

McGlinchey, S., Walters, R. and Scheinpflug, C. 2017. *International relations theory*. Bristol: E-International Relations. https://www.e-ir.info/publication/international-relations-theory/

McGlinchey, S., Waters, R. and Scheinpflug, C. 2021. Global civil society as a response to transnational exclusion. Bristol: E-International Relations. https://socialsci.libretexts.org/@go/page/11133

Mhlambi, S. 2020. From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance. Carr Center Discussion Paper Series (2020-009). https://carrcenter.hks.harvard.edu/publications/rationality-relationality-ubuntu-ethical-and-human-rights-framework-artificial

Mignolo, W. 2012. *Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking*. Princeton, N.J.; Oxford: Princeton University Press.

Milan, S. and Gutiérrez, M. 2015. "Citizens" media meets big data: The emergence of data activism. *Mediaciones*, Vol. 11, No. 14, pp. 120–133.

Milan, S. and Treré, E. 2019. Big Data from the South(s): Beyond data universalism. *Television & New Media*, Vol. 20, No. 4, pp. 319–335.

Milan, S. and van der Velden, L. 2016. The alternative epistemologies of data activism. *Digital Culture & Society*, Vol. 2, No. 2.

Mohamed, S., Png, M.-T. and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, Vol. 33.

Mosco, V. and Wasko, J. 1988. *The Political Economy of Information*. Madison, Wisconsin: The University of Wisconsin Press.

Müller, M. 2018. In search of the Global East: Thinking between North and South. *Geopolitics*, Vol. 25, No. 3, pp. 1–22.

Neupane, S. and Smith, M. L. 2017. *Artificial Intelligence and Human Development.* Ottawa, Canada: International Development Research Centre. https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949

Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press.

Ochigame, R. 2019. How Big Tech manipulates academia to avoid regulation. *The Intercept*, December 20. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

ÓhÉigeartaigh, S. S., Whittlestone, J., Liu, Y., Zeng, Y. and Liu, Z. 2020. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology*, Vol. 33, No. 4, pp. 571–593.

Parikka, J. 2015. *A Geology of Media*. Minneapolis; London: University of Minnesota Press.

Peck, J. 2011. *Ideal Illusions: How the U.S. Government Co-opted Human Rights*. New York: Henry Holt; Godalming.

Pils, E. 2019. The risks of complicity in transnational civil society repression: An argument for institutional responses. Workshop for the European Consortium for Political Research: Advancing Political Science, April 8-12. https://ecpr.eu/Events/Event/PaperDetails/44842

Pouliot, V. and Thérien, J.-P. 2017. Global governance in practice. *Global Policy*, Vol. 9, No. 2, pp. 163–172.

Quijano, A. 2000. Coloniality of power and Eurocentrism in Latin America. *International Sociology*, Vol. 15, No. 2, pp. 215–232.

Quinn, J., Frias-Martinez, V. and Subramanian, L. 2014. Computational sustainability and artificial intelligence in the developing world. *AI Magazine*, Vol. 35, No. 3, p. 36.

Rainie, S. C., Kukutai, T., Walter, M., Figueroa-Rodríguez, O. L., Walker, J. and Axelsson, P. 2019. Indigenous data sovereignty. African Minds and the International Development Research Centre (IDRC). https://researchcommons.waikato.ac.nz/handle/10289/12918

Raval, N., Kak, A. and Calcaño, A. 2021. A new AI lexicon: Responses and challenges to the critical AI discourse. AI Now Institute. https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-responses-and-challenges-to-the-critical-ai-discourse-f2275989fa62

Rayment, P. B. W. 1983. Intra-"industry" specialisation and the foreign trade of industrial countries. Stephen F. Frowen (ed.), *Controlling Industrial Economies*. The Vienna Institute for Comparative Economic Studies. London: Palgrave Macmillan, pp. 1–28.

Ricaurte, P. 2019. Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, Vol. 20, No. 4, pp. 350–365.

Roberts, A. 2018. Politeness as process, manners as method. OECD Observatory of Public Sector Innovation. July 26. https://oecd-opsi.org/politeness-as-process-manners-as-method/

Sampath, P. G. 2021. Technology and inequality: Can we decolonise the digital world?1. *South Views*, Vol. 215. https://www.southcentre.int/wp-content/uploads/2021/04/SouthViews-Sampath.pdf

Sassoon, A. S. 2014. *Gramsci and Contemporary Politics: Beyond Pessimism of the Intellect.* London: Routledge.

Satariano, A. 2021. Europe proposes strict rules for artificial intelligence. *The New York Times*, April 21. https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html

Schiff, D., Borenstein, J., Biddle, J. and Laas, K. 2021. AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, Vol. 2, No. 1, pp. 31–42.

Scholte, J. A. 2004. Civil society and democratically accountable global governance. *Government and Opposition*, Vol. 39, No. 2, pp. 211–233.

Selznick, P. 2015. *TVA and the Grass Roots: A Study of Politics and Organization*. New Orleans, LA: Quid Pro Books.

Singh, R. 2021. Mapping AI in the Global South: A new project to identify sites and vocabularies of digital IDs and AI. *Data & Society*, January 26. Blog. https://points.datasociety.net/ai-in-the-global-south-sites-and-vocabularies-e3b67d631508

Singh, R. and Lara Guzmán, R. 2021. Parables of AI in/from the Global South. *Data & Society*. Workshop call for participation. https://datasociety.net/announcements/2021/07/13/call-for-participants-parables-of-ai-in-from-the-global-south/

Smogorzewski, K. 1938. Poland's foreign relations. *The Slavonic and East European Review*, Vol. 16, No. 48, pp. 558–571.

Smuha, N. A. 2020. Beyond a human rights-based approach to AI governance: Promise, pitfalls, plea. *Philosophy & Technology*, Vol. 34, pp. 91–104.

South Centre. 2020. Submission by the South Centre to the Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence (WIPO/IP/AI/2/GE(20/1). https://www.southcentre.int/wp-content/uploads/2020/02/Submission-by-SC-to-the-Draft-Issues-Paper-on-IP-Policy-and-AI.pdf

Taylor, K.-Y. 2019. *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership.* Chapel Hill: University of North Carolina Press.

Taylor, L. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, Vol. 4, No. 2. doi:10.1177/2053951717736335

ten Oever, N. and Cath, C. 2017. Research into human rights protocol considerations. Internet Research Task Force, article 19. https://datatracker.ietf.org/doc/html/rfc8280

Thatcher, J., O'Sullivan, D., and Mahmoudi, D. 2016. Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space,* Vol. 34, No. 6, pp. 990–1006. https://doi.org/10.1177/0263775816633195.

Torres, G. 2017. Taking a look at institutional resistance to citizen empowerment. *DATACTIVE*. Blog. https://data-activism.net/2017/02/blog-taking-a-look-at-institutional-resistance-to-citizen-empowerment-through-data/

Ulnicane, I., Eke, D.O., Knight, W., Ogoh, G. and Stahl, B.C. 2021. Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews*, Vol. 46, No. 1-2, pp. 71–93.

Ulnicane, I., Knight, W., Leach, T., Stahl, B.C. and Wanjiku, W.-G. 2020. Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, Vol. 40, No. 2, pp. 1–20.

UNCTAD. 2013. *Information economy report 2013: The cloud economy and developing countries*. Geneva: United Nations Conference on Trade and Development.

UN DESA. 2014. *Global Governance and Global Rules for Development in the Post-2015 Era*. Department of Economic and Social Affairs, Committee for Development Policy. https://www.un.org/en/development/desa/policy/cdp/cdp_publications/2014cdppolicynote.pdf

Vawda, Y. 2021. The TRIPS COVID-19 waiver, challenges for Africa and decolonizing intellectual property. *South Centre Policy Brief*, Vol. 99. https://www.southcentre.int/policy-brief-99-august-2021/

Veale, M. and Borgesius, F. Z. 2021. Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*. Vol. 22, No. 4, pp. 97–112. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3896852

Velluet, Q. 2021. Can Africa salvage its digital sovereignty? *The Africa Report.* April 16. https://www.theafricareport.com/80606/can-africa-salvage-its-digital-sovereignty/

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M. and Nerini, F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, Vol. 11, No. 1.

Weiss, T. G. 2016. Rising powers, global governance, and the United Nations. *Rising Powers Quarterly*, Vol. 1, No. 2, pp. 7–19.

Werner, D. and PROJIMO. 1998. *Nothing About Us Without Us: Developing Innovative Technologies For, By and With Disabled Persons*. Palo Alto: Healthwrights.

Williams, G. 2004. Evaluating participatory development: Tyranny, power and (re)politicisation. *Third World Quarterly*, Vol. 25, No. 3, pp. 557–578. https://www.jstor.org/stable/3993825?seq=1

Winner, L. 1980. Do artifacts have politics? *Daedalus*, Vol. 109, No. 1, pp. 121–136. http://www.jstor.org/stable/20024652

Wong, P.-H. 2016. Responsible innovation for decent nonliberal peoples: A dilemma? *Journal of Responsible Innovation*, Vol. 3, No. 2, pp. 154–168.

Yuan, L. 2018. How cheap labor drives China's A.I. ambitions. *The New York Times*. November 25. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html

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

DEMOCRATIZE THE DEVELOPMENT OF AI POLICIES

STEFAN RIEZEBOS

Research Lead, Innovation for Policy Foundation. He holds a Master's degree in policy economics from the Erasmus University Rotterdam and has over seven years' experience as senior advisor for the Dutch government, specializing in the regulation of the platform economy (GAFAs).

TIM GELISSEN

Director of Advisory, Innovation for Policy Foundation. He holds a Master's degree in international economics from Tilburg University and has worked 10 years for the Dutch government. With i4Policy he has been supporting several co-creation processes in different countries, most recently in Rwanda and Nigeria.

RAASHI SAXENA

AI Program Specialist, Innovation for Policy Foundation. She holds a bachelor's degree in engineering in telecommunication from Visvesvaraya Technological University, India, and has extensive experience as a technologist, social impact innovator and consultant.

for the Innovation for Policy Foundation

SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

DEMOCRATIZE THE DEVELOPMENT OF AI POLICIES

ABSTRACT

Artificial Intelligence (AI) is becoming the veiled decision-maker of our times. Governments are recognizing the uptake of AI and its potential economic and societal impact. More than 50 countries have published policy documents aimed at harnessing the benefits of AI while safeguarding public interest. The implementation and growth of AI systems is surrounded by risks to human rights and leads to values-driven questions and dilemmas. We stress that curbing the impact of AI systems cannot be left up to a single stakeholder. A multistakeholder participatory approach is needed for the design of AI policies, involving a wide array of stakeholders such as policymakers, civil society actors, citizens, scholars, the technical community, and the private sector. This will enhance the quality of decision-making, create the space for learning and deliberation and help us move away from a monolithic view of AI. Building on country case studies, workshops, expert interviews, and practical experience in over a dozen countries, we highlight five lessons for the design of inclusive AI policies.

INTRODUCTION

Artificial Intelligence is increasingly becoming the veiled decision-maker of our times. Its impact cuts across different economic and social domains, and it affects our daily lives in ways that most of us are not even aware of. AI systems are no longer only of interest to researchers or science-fiction fans; its uptake in society is increasing rapidly. Spotify's AI-driven Wrapped 2022 will spark lively debates about one's music taste while at the same time doctors are using AI to help diagnose illnesses and another AI system can autonomously determine one's eligibility for a personal loan.[59]

59. One could blame Spotify's algorithm for some of the guilty pleasures in the "most listened songs of the year" list.

Even though AI as a research field has existed since the 1950s, it has only recently left the lab and moved into society (Netherlands Scientific Council for Government Policy, 2021). There are three main factors driving this development. First, an expansion in computing power over the past years has made it possible to perform more complex calculations (often referred to as Moore's law). Second, the amount of available data has increased immensely, a development that goes hand-in-hand with decreasing costs of data storage. Third, several scientific breakthroughs have made it possible for AI to discern patterns in different layers of data. This ability to look deeper into data is also known as "deep learning" and paved the way for the AI applications that we use today.

Comparing AI to other general-purpose technologies such as the steam engine, electricity, and the computer makes it clear that AI has substantial potential. An estimate by McKinsey Global Institute (2018) shows that AI can have a positive effect on the global economy of US $13 trillion by 2030. This amounts to a yearly increase in growth of around 1.2%. ITU (2018) notes that, if delivered, that would compare well with the effect of other general-purpose technologies through history. Behind these numbers lies the potential of AI to transform industries and improve societal outcomes, for example by supporting the provision of food, health, water, and energy services and enhancing the transition towards carbon neutrality (Vinuesa et al., 2020).

This is not to say that the impact of AI on the world will solely be positive. Vinuesa et al. (2020) conclude that the use of AI can play an inhibiting role in the achievement of 59 targets under the Sustainable Development Goals (SDGs). This potential negative impact has multiple dimensions, ranging from widened income inequality on a national and international level to exacerbated market concentration and job polarization. The use of AI systems, such as facial recognition, pattern recognition, and deep fakes, can elicit serious privacy concerns, spread disinformation, and lead to the infringement of human rights and individual freedoms.

Governments are recognizing the uptake of AI and its potential implications. Canada was the first country to publish a national AI Strategy in 2017. Many others have followed since and drafted strategies, laws, and policy documents related to AI. As of now, more than 50 such documents have been published.[60] Even though each country defines its own priorities and focus areas, the objective of most AI policies is identical: to harness its benefits while safeguarding the public interest.[61]

AI policies and ethics guidelines are often presented as products of a wider consensus. They are, however, overwhelmingly produced by economically developed countries and underwhelmingly inclusionary (Crawford, 2021). This becomes apparent from the distinct difference in involvement between the Global North and Global South: the African continent is still in the early phases of AI policy development (Smart Africa, 2021); Gwagwa et al. (2020, p.16) note that "one characteristic of certain AI policy discussions [...] has been the marginalization or exclusion of Global Southern inputs"; and AlgorithmWatch (2020) found that the overwhelming majority of ethics guidelines come from Europe and the US.

Only a handful of countries have used a participatory process to reach consensus, inform the public, and deliberate on potential policy solutions. This stands in sharp contrast to the main challenges surrounding the coming-of-age of AI systems. They lead to values-driven dilemmas and complex problems that require trade-offs. These dilemmas and challenges are too important and complex to be decided upon by a single set of stakeholders.

60. For an overview of national AI strategies, see the Future of Life Institute (https://futureoflife.org/ai-policy/), the OECD.AI Policy Observatory (https://oecd.ai/en/dashboards) and the Globalpolicy.ai initiative.

61. We use the term "AI policy" as an umbrella term, which includes AI strategies, action plans, and concrete policy proposals such as laws or other official policy measures.

This chapter centers on the question of how we can democratize the development of AI policies. We believe that true multistakeholder approaches are part of the answer. Involving policymakers, civil society actors, individual users and citizens, academics, the technical community, and the private sector is necessary in order to properly deliberate on the dilemmas and challenges that AI systems bring and to reach an outcome that works for all. It is time to shift our attention to the process instead of focusing solely on outcomes.

Building this argument, section 1 discusses the need for regulating AI applications and technologies. Section 2 introduces the concept of multi-stakeholderism and deliberation; section 3 explains how it applies to the design of AI policy. Section 4 lies out general principles for successful multi-stakeholder processes, while section 5 illustrates these with two case studies. Finally, section 6 is most practical and highlights key lessons for the design of multi-stakeholder approaches.

The work presented is the result of a cooperation between the Innovation for Policy Foundation (i4Policy) and UNESCO, and builds on the output of five workshops, expert interviews, an analysis of over 20 case studies, and practical experience with multi-stakeholder approaches in over a dozen countries.

---

1. RISKS SURROUNDING THE IMPLEMENTATION OF AI SYSTEMS

In this section, we dissect the risks surrounding AI systems. We build on input from two expert workshops in September and October 2021 and an additional literature review. The workshops focused on the impact of AI on human rights and fundamental freedoms. We show that the potential negative impact can be substantial. Failing to channel the impact of AI systems through dedicated policies will result in a need for correcting policies to mitigate the effects of the uncurbed spread of AI into societies. We will discuss three of these risks in detail below: the disparity between the Global North and the Global South, the disproportionately negative effect of AI on historically marginalized groups, and privacy concerns and surveillance. These risks are the impetus for dedicating time and effort to the design of inclusive participatory policies.

Disparity between the Global North and the Global South

First, great disparity exists between the Global North and the Global South in terms of the development, deployment, and use of AI. The digital divide between developed and developing countries remains high and is a recurrent challenge for development (UNCTAD, 2021). This divide translates into the role that the Global South plays in conversations about AI and its development. A meta-analysis by Jobin et al. (2019) shows that 67% of AI ethics principles contain a heavy influence of US and Western values. African, South and Central American, and Central Asian countries (except for India) are not even represented in their data. Similarly, the vast majority of AI applications in use in sub-Saharan Africa are not made by or in Africa (Oxford Insights, 2020). Birhane (2020) explains that this is problematic since value systems vary from culture to culture, including what is considered a problem and what is understood as a successful solution. She points at a striking example from Black and Richmond (2019), who found that early breast cancer detection practices that worked well in the West were not effective in Sub-Saharan Africa because of patients' lower average age, more advanced stages of the disease, and limited availability of treatment options.

In recent years, theories of data colonialism and data capitalism have emerged in literature, recognizing the role of data as a material resource that is being exploited (Mohamed et al., 2020; Kwet, 2019). Moreover, scholars working in the field of AI militarization have warned that the Global South is severely underrepresented in the conversation. Yet, risks to peace and security are felt first in conflict zones in developing countries. This is illustrated by a publication by the United Nations Security Council which

finds that in March 2020, lethal autonomous weapons systems were used to attack a logistics convoy and retreating soldiers in Libya. The report emphasizes that the weapon systems were programmed to attack targets without requiring an active data connection to an operator (Garcia, 2019; United Nations Security Council, 2021, p. 20).

UNESCO (2021) and Gwagwa et al. (2021) underline that, while there is a high level of diversity in AI deployment in Africa in terms of problems being addressed, people, and countries working on AI, the vast majority of African countries do not have dedicated AI policy frameworks. The absence of policy frameworks risks leaving African countries out of the conversation about AI standards and ethics and hampers the creation of a local enabling environment for the development and use of AI.

Disproportionate negative impact on marginalized groups

Second, AI can have a disproportionate negative impact on historically marginalized groups. AI based on non-representative data can exacerbate existing social and economic inequities. This is because AI will reproduce gaps or biases that exist in data on which it is trained. For instance, think of a company where most employees are male. When an AI hiring tool is trained based on the historical hiring data of the said company, it will develop a preference for hiring male candidates and discriminate against women. This is exactly what happened in the case of Amazon (Polli, 2019). Besides being data-driven, biases of AI systems can also be programmer-driven. Since human programmers are responsible for framing the problem and for determining the validity of the output, their personal bias can be integrated in the system (UNESCO, 2019; Barocas and Selbst, 2016).

Both biases are illustrated by Prince and Schwarcz (2020), who argue that AI can result in proxy discrimination. This effect occurs when a seemingly neutral variable is included in a model as a proxy for a variable of which the use is prohibited because of known discriminatory effects. Programmer-driven bias occurs when such a proxy is included intentionally (also known as "masking"). However, in the case of AI, the risk of unintentional proxy discrimination also lingers. This is a data-driven bias. AI is trained with large datasets and often allowed the freedom to derive relations between variables themselves, based on the expected output. This increases the chance of AI making use of and finding new proxy variables that are correlated with marginalized groups, potentially resulting in discriminatory effects. An example of this would be a zip code. At first glance, zip codes seem to be a neutral variable, simply containing one's location. However, zip codes are closely correlated to socio-economic status and ethnicity. Murray (2013), for example, illustrated this by introducing the term "super zip" to refer to the wealthiest and most influential zip codes in the United States. The existence of this correlation means that including zip codes in datasets can result in AI unintentionally discriminating against minorities.

Privacy concerns and surveillance

Third, our right to privacy, a fundamental human right, is impacted by AI. The protection of the right to privacy is broad, covering not only substantive information contained in communications but also metadata, since metadata, when analyzed, may also give insight into an individual's behavior (OHCHR, 2018). Moreover, it is not just the examination or use of data by programmers and algorithms that can affect privacy. As Bernal (2016, p. 249) argues, "many of the key risks occur when data are gathered—the existence of data creates risk." Hence, our right to privacy can be affected at several points in the process, while data is being collected, when it's being analyzed, and when output is evaluated.

To illustrate, AI surveillance technology is an area in which great care needs to be taken to protect the right to privacy. This type of AI system is spreading at a faster rate than commonly understood. Feldstein (2019) shows this with the AI Global Surveillance (AIGS) Index, compiling data on AI surveillance use for 176 countries. Valid questions have been raised by scholars, activists, and NGOs about the consequences of the use of facial recognition technology for our privacy and our freedom of expression (see, for example, Moraes et al., 2021 and Mudongo, 2021). The Office of the UN High

Commissioner for Human Rights concluded in 2018 that many states continue to engage in mass surveillance and intercepting communications (OHCHR, 2018). And even though many states claim that mass surveillance is necessary for national security reasons, this use of facial recognition technology is not permissible under international human rights law (Privacy International, 2019).

In summary, AI systems pose a pronounced risk to human rights and fundamental freedoms when allowed to spread uncurbed. It is imperative that we safeguard the public interest by instating dedicated AI policies. In the next two sections we argue that a deliberative multi-stakeholder approach is well suited to do this.

---

2. MULTI-STAKEHOLDERISM: ITS ORIGINS AND CONTEXT

Stakeholder participation has been an element of public governance strategies for a long time. Hofmann (2016) shows that, traditionally, stakeholder participation was mainly associated with international topics such as labor conditions and environmental standards. She illustrates this by pointing out the tripartite structure of the International Labour Organization (ILO), comprising government, employer, and worker representatives. The ILO, founded in 1919, is until this day the only United Nations (UN) agency with a tripartite structure.

The term "multi-stakeholder" was coined in the 1990s. It has gained a lot of traction since then, in particular related to global governance. Scholte (2020) shows that global multi-stakeholder approaches became an alternative to international multilateralism because they were increasingly seen as an answer to dealing with complex and uncertain problems that affected actors and agencies on a global scale.[62] He adds that the motivating intuition behind multi-stakeholderism is that blending diverse pools of information and insight can yield more effective global problem-solving and more resources can be pooled and used to address the problem at hand.

Simultaneously a "deliberative wave" is being observed in national policy development, as noted by the OECD (2020). These national initiatives are often referred to as deliberative or participatory democracy. In similar fashion as Scholte, the OECD argues that the increasing complexity of policymaking and the failure to find solutions to some of the most pressing policy problems have prompted politicians, policymakers, civil society organizations, and citizens to reflect on how collective public decisions should be taken in the 21st century.

The idea behind multi-stakeholder governance is that it serves to create a platform for dialogue that can build consensus around a shared set of goals and values. It is grounded in Habermas's theory of discourse ethics, which argues that morals and norms emerge from a process where those with opposing views engage with each other. Hence, when all parties rationally consider each other's arguments, together they should achieve a greater understanding. This in turn leads to parties reassessing their position, a process that continues until all parties involved reach a universally agreeable decision (Habermas, 1989; Martens et al., 2019).

The main benefits that follow from this normative approach are that multi-stakeholder approaches improve inclusiveness, create understanding about the concerns and interests of other stakeholders and lead to a higher quality of decision-making. First, multi-stakeholder approaches improve inclusiveness because they open the door to a much more diverse group of people to participate, such as youth, the

62. See also Dingwerth (2008), Brockmyer and Fox (2015) and Gleckman (2018) for further reading on the development of multi-stakeholder governance.

disadvantaged, women, or other minorities (Adam et al., 2007). Second, participatory processes create space for learning, deliberation, and the development of informed recommendations, leading to a better understanding of the concerns and interests of other stakeholders (Faysse, 2006). And third, a higher quality of decision-making is the expected result of adding greater expertise and more diversity into decision-making processes and of encouraging consensus-building (Souter, 2017).

This is not to say that multi-stakeholderism is a panacea. Several scholars show that multi-stakeholder initiatives do not always meet expectations in practice, pointing at a lack of trust (Sloan and Oliver, 2013), issues of legitimacy (Bäckstrand, 2006), or the amount of time and resources involved (Moog et al., 2014). Additionally, power asymmetries can arise when parties are not able to contribute equally in terms of knowledge, finances, and access to information (Fransen and Kolk, 2007). The impact of multi-stakeholderism on decision-making can be particularly problematic for short-term processes, where links to formalized decision-making processes tend to be unclear (Faysse, 2006).

Therefore, before adapting a multi-stakeholder process, one needs to identify whether the nature of such a process is applicable and how to address the weaknesses in the process. In the next section we argue that the development of AI policies fits well with a multi-stakeholder approach, while the remainder of this chapter is concerned with the optimal design of a multi-stakeholder approach.

3. WHY A MULTI-STAKEHOLDER APPROACH IS NEEDED FOR THE DEVELOPMENT OF AI POLICIES

Buhmann and Fieseler (2021) posit two reasons why communicative and deliberative approaches can offer fitting solutions for AI policy. First, they argue that the far-reaching societal ramifications of AI systems and their rapid proliferation in all public and private spheres of human life should make them a central object of broad political concern. Second, the opaqueness and lack of accountability of AI systems require the epistemic power of deliberation to improve knowledge and feedback through self-correcting learning processes among empowered actors. Additionally, we consider a third factor. As the OECD (2020) shows, deliberative processes are best suited when the topic at hand contains value-driven dilemmas, complex problems that require trade-offs, and long-term issues that go beyond the short-term incentives of electoral cycles. We discuss these factors below.

Far-reaching societal ramifications of AI

The potential impact of AI systems is best explained by looking at it through the lens of a general-purpose technology. The Netherlands Scientific Council for Government Policy (2021) shows that the uptake of AI can be seen as such a technology, thereby adding AI, and its potential impact, to the same league as the steam engine, electricity, the combustion engine, and the computer.

General-purpose technologies can be characterized by three factors: pervasiveness, continual improvement and innovational complementarities. First, pervasiveness refers to the way in which the technology spreads to different sectors, production processes, and final products and services. According to the discussion in the workshops, it is becoming increasingly difficult for citizens to opt out of engaging with AI systems as businesses and public institutions are embedding AI in their everyday products and services. Second, continual improvement relates to the technological rate of advancement. This holds as AI is not constant, but it continues to develop and improve, driven by computational improvements, decreasing costs of data generation and storage, and ongoing scientific research. Third, innovational complementarities imply that, as businesses and governments embed AI in processes

or services, connected technologies and processes will also become more efficient, leading to productivity gains. Or, to put it in the words of Trajtenberg (2018, p. 176), AI can bring about "a wave of complementary innovations in a wide and ever-expanding range of applications sectors."

The impact of AI systems is hence large and should be the object of broad political concern, requiring the involvement of a broad set of stakeholders who are affected by the increasing presence of AI systems.

Need for transparency and accountability

AI systems often lack transparency and explicability (meaning whether they are explainable and interpretable). These factors are, however, necessary preconditions to ensuring trust, for legal regimes to work properly, and for the ability to evaluate and potentially challenge outcomes. For this argument, we distinguish between two components here, one of a technical nature and one relating to communication.

From a technical point of view, AI systems can be complex, and their working can be difficult and time-consuming to explain. This is particularly relevant for deep learning models as they evaluate data on deeper layers than we as humans are capable of. However, as Rudin (2019) stipulates, it is a myth that there is necessarily a trade-off between the accuracy and interpretability of these models. She argues that this myth has led researchers to forgo their efforts to produce an interpretable model and adds that there is a strong commercial incentive for the private sector to keep models hidden. It is important, though, to stress that, technically, there are alternatives, and in most cases, no inherent difficulties to providing transparency (see also: Hall and Gill, 2019; Guidotti et al., 2018).

The second component is of a communicative nature. The understanding of AI in society is scarce. The vocabulary used by those that are knowledgeable about it, including terminology such as "black box," "machine learning" and "big data," is difficult to grasp for most. This can derail conversations on AI. However, it is important for civil society actors and users to have a general understanding of the systems they engage with in order to understand their consequences, to identify room for improvement, to engage in debate, and to potentially challenge outcomes when these actors and users are being adversely affected. Deliberative processes simultaneously contribute to this process of learning and generate feedback regarding the working of AI systems.

Values-driven dilemmas and trade-offs

Deliberative processes are best suited when the topic at hand contains value-driven dilemmas, complex problems that require trade-offs, and long-term issues that go beyond the short-term incentives of electoral cycles. Increasingly intelligent AI systems create such dilemmas and trade-offs (see, for example, Winfield, 2019).

An example is the complex issue of online hate speech. AI systems are currently the primary method employed by tech companies to find, categorize and remove online harms at scale (see, for example, Gorwa et al., 2020). However, in practice, they are beset with methodological, technical, and ethical challenges. In many cases, they are used in scenarios where the decision requires the protection of freedom of speech and safeguarding users from harm whilst simultaneously respecting users' right to privacy. In addition to these challenges, tech companies also need to be able to explain the rationale for decisions made by these systems and they are responsible for mitigating harms stemming from the social biases encoded into their AI systems (Llansó et al., 2020).

This section has shown that all the ingredients that call for a multi-stakeholder approach are present. The design and use of AI systems create moral dilemmas, and they have a vast and long-term impact on society. Continued learning and deliberation are needed to improve transparency and accountability.

4. GENERAL PRINCIPLES FOR SUCCESSFUL MULTI-STAKEHOLDER PROCESSES

Designing a policy framework for AI requires striking a balance between supporting innovation and mitigating the risks it poses. One of the biggest challenges of regulating a general-purpose technology is determining when and how strictly to regulate. The Collingridge dilemma captures the difficulty of this task. It holds that in the initial phases of development, the nature and impact of a new technology are still difficult to assess, making it difficult to regulate. However, by the time undesirable consequences of the technology are discovered, it is often so intertwined with our economic and social systems that it is, again, very difficult to regulate (Collingridge, 1981). This calls for an agile and flexible AI policy in order to cope with its continuous development. It implies, for example, that it is wise to choose policy options with low error costs and increases the importance of effective monitoring and evaluation. Engaging stakeholders can be a valuable way to achieve this.

A wide body of literature exists on stakeholder engagement. Lessons learned from different sectors, regions and cultures have been drawn up over the past years. For example, Renn et al. (2020) and Ambole et al. (2021) discuss experiences with stakeholder participation in the energy sector; Mustalahti and Rakotonarivo (2014) provide an analysis of community participation in Tanzania to reduce emissions from deforestation and forest degradation; García-López and Arizpe (2010) provide an example on the use of participatory processes to address conflicts over soy production in Paraguay and Argentina; and Hoogesteger (2012) looks at making water management more democratic through the participation of water users in the Ecuadorian Andes. Common elements in these analyses are, first, that multi-stakeholder approaches enable decision-makers to prioritize pressing issues and make informed, data-driven decisions; second, that these approaches foster long-term growth and sustainability by reinforcing the influence and representation of marginalized groups; and third, that local communities are often inadequately resourced to anchor and manage their own projects. An important general takeaway is therefore that local communities are helped by cooperation with external experts that can aid with co-design, facilitation, and financial and logistical support.

Additional lessons can be drawn from previous technological developments. Van der Spuy (2017) provides a comprehensive overview of the evolution of multi-stakeholder participation in internet governance. Her analysis shows that participatory processes need to exhibit a number of values if they are to be effective in developing consensus and improving decision-making. She finds that multi-stakeholder approaches need to be inclusive, diverse, collaborative, transparent, equal, flexible and relevant, safe and private, accountable and legitimate, and responsive. We argue that these generic principles can be considered a baseline for a multi-stakeholder approach for the design of AI policies.

Catering specifically to the topic at hand, Buhmann and Fieseler (2021) suggest four communicative principles for deliberation on AI. These are slightly more practical in nature and are based on a theoretical study by Nanz and Steffek (2005). The first principle signifies the need for institutionalized access to deliberative settings. Everyone with the competence to speak and act, especially those that may potentially suffer the negative effects of processes and decisions of algorithms, should have equal access to an open forum that aims to spotlight issues and facilitate conversation. This principle is underlined by Bondi et al. (2021), who stress the need for a community-based approach in evaluating the success of AI projects. An important precondition to note here is that, especially in developing countries, raising awareness and building capacity on the topic is a crucial aspect of a successful deliberative process. This will encourage and enable more people to contribute to the debate, improving both the level of inclusion and diversity of the deliberative process.

The second principle states that those who participate in the deliberative process should have access to as much information as possible about the issues at stake, potential solutions, and their consequences. Making information openly accessible to all actors would be a good example in this respect. However, that alone is not enough to comply with this principle. Different stakeholders do not

possess the same level of information or access to it. And while it is utopic to aim for similar levels of knowledge, this principle does entail responsibility for those who possess more knowledge and information (i.e., technical community, private sector, and policymakers) to ensure that other stakeholders, generally civil society actors, have the opportunity and resources to improve their knowledge about the subject matter.

The third principle holds that all possible arguments should be included and considered in the process. This precondition safeguards the rationality of discourse and deliberation and ensures that the outcome is balanced. Lastly, the authors stress that a deliberative process needs to be responsive to stakeholders' concerns and suggestions. Even if all of the previous conditions are met, participation cannot affect the outcome if decision-makers are not open to input and do not allow influence on the decision-making process.

To help policymakers implement these principles, the University of Washington and the Université de Montréal have developed two useful how-to guides. The first guide, titled "Diverse Voices," helps policymakers improve diversity in the policy process and contains a method to include under-represented groups (Magassa et al., 2017; see also Young et al., 2019). The second guide explains in simple but accurate terms what AI is, how it relates to other technological concepts, and why deliberation on AI is necessary (Dilhac et al., 2020). In addition, the Westminster Foundation for Democracy and the newDemocracy Foundation analyzed the new deliberative wave of democratic processes in Africa, Asia, and Latin America. These examples and lessons on deliberative democracy can also be helpful and act as a source of inspiration (WFD, 2021 and Kimaili, 2021).

5. CASE STUDIES: CHILE AND INDIA

Since the first national AI strategy was published in 2017 in Canada, many countries have started working on their own strategies, action plans, or policies. Initiatives have been proposed or implemented in most of the OECD countries by now, but in many countries in the Global South, the work has yet to start. For example, Egypt and Mauritius are the only African countries that currently have a dedicated AI strategy, although developments are also ongoing in Rwanda, Kenya, Ghana, Nigeria, South Africa, Tunisia, and Uganda (Effoduh, 2020). Up-to-date overviews of developments worldwide are published by the Future of Life Institute,[63] the OECD.AI Policy Observatory (OECD.AI, 2021), and the OECD Globalpolicy.ai initiative.[64]

This section delves deeper into the approach taken by Chile and India. Three phases of the policy process are distinguished: agenda-setting, drafting, and implementation and evaluation. Steps taken in each phase are discussed by highlighting key choices and decisions and paying special attention to the nature of participation.

63. https://futureoflife.org/ai-policy/

64. https://globalpolicy.ai/en/

Case study: Chile

| FIGURE 1 |

Key moments in the development of the AI Strategy Chile.

Chile started developing its AI Strategy in 2019. The Chilean government noted that advancement in AI created a need to act preventively in the face of societal changes that AI could spur. After a two-stage process of stakeholder participation the strategy was finalized and published in October 2021 (MCTCI, 2021b). The strategy is structured around three axes—enabling factors, development and adoption of AI, and ethical, regulatory and socioeconomic aspects—and calls for the development and use of human-centered AI which is safe, inclusive, globalized, and at the service of society. The final publication includes 70 priority actions for the short term (Action Plan) and 180 initiatives to be developed over the period of 2021 to 2030 (AI strategy).

The Chilean approach is an excellent example of a multi-stakeholder participatory process. First, in the agenda-setting phase, policymakers carried out a comparative analysis of the AI strategies and policies of other countries. The results of this analysis were presented to the Presidency of Chile in August 2019. The Presidency mandated the Ministry of Science, Technology, Knowledge and Innovation to develop a national AI strategy, guided by a committee of experts and representatives from various ministries. They were tasked with creating a draft strategy to be published for input from the general public.

However, after the 2019 period of social unrest in Chile, the experts and policymakers modified the linear, top-down approach to a bottom-up, participatory multi-stakeholder one. Instead of creating a draft strategy, the experts compiled a list of relevant AI policy topics. This list guided the first phase of the multi-stakeholder participatory process, launched in February 2020. This phase consisted of three elements: an open call for self-convocated roundtables (including a blank online feedback form), the organization of regional roundtables by the ministry, and online webinars held by experts to raise awareness and build capacity. The process was facilitated by a public participation manual, with civil servants offering presentations at roundtables when required and with public sponsorship of these roundtables.

The unique nature of the participatory process becomes apparent when we consider the number of stakeholders involved. During a period of six months, over 1,300 persons and organizations self-convened roundtables and provided input online, and a total of 69 regional roundtables were organized with 400 participants. The webinars reached 6,600 people; half of the experts hosting these were men

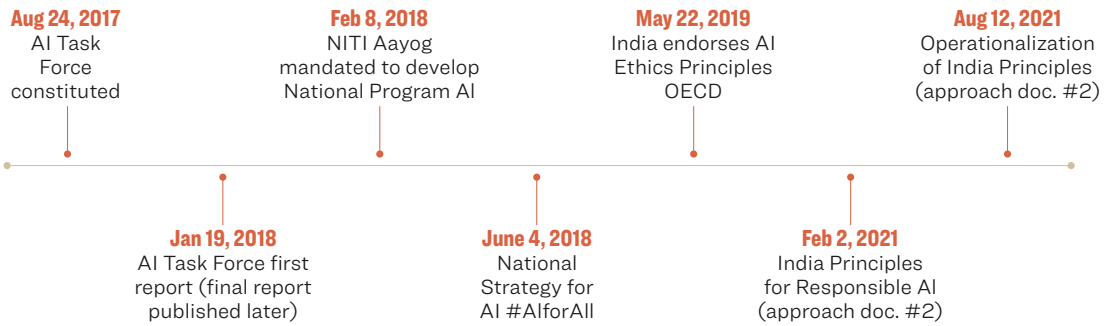
and half were women. Participation in the process was also diverse: 36% of the responses online originated from civil society, and several participants indicated that they had not contributed to policy development before.

Based on these inputs, experts and policymakers developed a first draft of the strategy. A second phase of participation started in December 2020, when the public-input draft was published online for public consultation. In this process, participants provided new questions and comments and weighed their level of agreement with the objectives and specific aspects of the AI Policy. The consultation process indicated an average acceptance of over 80% for the objective and principles of the draft. Qualitative feedback showed that participants valued both the bottom-up process and the educational benefits it provided (MCTCI, 2021a). After processing these inputs, the drafting stage was completed in June 2021, and the phase of political adoption commenced. Five months later, on October 28, the Chilean AI strategy and action plan were published.

Case study: India

| FIGURE 2 |

Key moments in the development of the AI Strategy India.



The policy process in India started with the constitution of an AI task force (see Figure 2). Following the report of this task force, the public policy think-tank NITI Aayog was mandated to draft a National AI Strategy. The strategy was published in the summer of 2018 and was branded "#AIforAll," aiming for inclusive technology leadership (NITI Aayog, 2018). Since 2018, discussions on the way to transform the strategy into public policy have been ongoing. After more than a year of extensive consultations with experts, civil society, and the private sector, NITI Aayog recently released two approach documents. These serve as a roadmap for the development of the AI ecosystem in India and contain the latest information on the policy process (NITI Aayog, 2021a; NITI Aayog, 2021b).

Because the policy process in India is still ongoing, this section highlights a number of interesting elements instead of describing the full process in detail. First, we take a closer look at the AI task force. It was tasked with analyzing the state of AI in India and providing recommendations on the role of the government. The task force presented its findings in January 2018 (Kamakoti, 2018). Two aspects stand out. The task force comprised 18 members from diverse backgrounds: members from the field of AI technology, civil services, healthcare, law and finance. Diversity in an AI task force is crucial, as it lays the groundwork for the approach to and perspective on AI. The second aspect that catches the

eye is that the task force launched a website to solicit public opinion on AI related issues. This is a good practice, especially in the agenda-setting phase. Engaging the public early not only helps provide context on the state of the AI landscape, but it also provides valuable information on the perception of AI in civil society, which is crucial to determining the potential of AI in terms of uptake and understanding.

The second element that deserves mention is the National AI portal INDIAai, which was launched in 2020.[65] It aims to provide stakeholders with one single place to find all information related to AI and to strengthen the AI ecosystem in India. It is financed jointly by the government and the private sector and has started several noteworthy initiatives. Examples are education programs for youth, the launch of an AI chatbot to combat misinformation about COVID-19, and a National Mission on Language Translation. The latter project aims to remove the language barrier that not possessing a high level of English poses in India. This is particularly relevant as India has 22 official languages and at least a thousand more unofficial languages and dialects (Census of India, 2011).

Third, we draw attention to the publication of two handbooks that are the outcome of engagement between public bodies and stakeholders in India—specifically, a handbook on data protection and privacy targeting developers of AI and one on mitigating bias in AI for startups. They were drafted and published respectively under the coordination of GIZ India (2021) in close cooperation with the Data Security Council of India (DSCI), and by INDIAai (2021). The handbooks contain practical tips and guidance for developers and entrepreneurs based on academic research, globally recognized ethical principles and the regulatory landscape for India. They are an excellent way to disseminate recent insights to the target audience of AI developers and entrepreneurs and could also prove useful for policymakers in other countries.

The last element is India's global AI Summit, RAISE (Responsible AI for Social Empowerment). This virtual summit was held in 2020. Organized by the Ministry of Electronics and IT (MeitY), it brought together policymakers, AI experts, thinkers, influencers, practitioners, and youth from India and abroad. It consisted of 48 sessions, lasted 85 hours, included several hundred speakers and engaged 79,000 participants from 147 countries. India used RAISE to reiterate its commitment towards responsibly embracing artificial intelligence on a global scale and provided the international AI community with a platform to exchange ideas, an important component of multi-stakeholder engagement.

6. LESSONS LEARNED

The previous sections discussed a number of criteria for successful participatory processes, shared evidence from multi-stakeholder approaches in practice, and reviewed the approaches taken by Chile and India. Several lessons can be drawn when we combine these sources of information.

First, clarity in the agenda-setting phase of the policy process should be a building block of an inclusive multi-stakeholder process for AI policy. It is the responsibility of governments to take the lead on the development of AI policy and they need to provide clarity to the private sector, academics, and civil society on the process that will be employed. Regulatory predictability is always important, but as AI policies are very much in a developing phase, their contents are difficult to predict. Clarity on the process to be followed is therefore extra important as it provides a degree of comfort to stakeholders that their voices will be heard and their interests considered. A good example can be seen in Chile, where policymakers announced early on that a two-stage participation process would be followed. This provided clarity to all stakeholders on how and when they would be engaged.

65. https://indiaai.gov.in/

Second, both case studies show that employing an expert group or task force early in the policy process can be beneficial. Experts are generally tasked with analyzing the current state of AI, determining strategic priorities and potential competitive advantages in research, development, and deployment of AI. Crucial to such an approach is the selection process. To obtain a balanced view of the opportunities, weaknesses, and priorities of AI, the selection process needs to be inclusive. Experts should represent a variety of perspectives and interests. An inclusive process and outcome can only exist when an active effort is made to include stakeholders who are usually neglected, such as representatives of disadvantaged groups and youth organizations. For example, the task force of India was diverse in terms of educational backgrounds.

Third, developing an AI strategy and AI policies requires a clear mandate. A single ministry or public institute should ultimately be in charge and coordinate the development of the overarching AI strategy or policy framework. The reason for this is that AI affects the policy fields, and therefore responsibilities, of practically all ministries and public institutions. This lesson is illustrated by the case studies, as both NITI Aayog and MCTCI operated with a presidential mandate. However, having a mandate does not entail unilateral decision-making: multistakeholder approaches are ineffective when decisions are made disregarding the input from stakeholders. Rather, the mandate can be used to incentivize relevant government agencies and institutions to join the deliberation.

Fourth, moving on to the drafting phase, employing an open and inclusive drafting process is key. Consultation processes have become common in many countries, but they are often short in duration and follow a rigid structure, and their effectiveness can be questionable. The Innovation for Policy process and the Chilean approach show that there is room to innovate these processes.[66] For example, stakeholders can be provided with the means to annotate draft texts, to respond to specific sections or to ask detailed questions about key elements, and the level of stakeholders' agreement with the draft could be surveyed. This ensures that policymakers obtain more relevant feedback, that stakeholders feel more involved and will enable policymakers to be responsive to participants' feedback and suggestions.

Fifth, and lastly, we consider the implementation and evaluation phase. The lesson to be learned here is the necessity and importance of a short-term action plan. An AI strategy presents the overarching framework and gives direction but moving in this direction requires concrete actions. A good practice is to draft both an AI strategy and a short-term action plan. This creates ownership and a sense of urgency as it requires all stakeholders to define concrete actions, allocate budgets and agree on the distribution of responsibilities. Additionally, short-term actions will help to keep stakeholders engaged and to present tangible short-term progress to decision-makers, politicians, and civil society.

66. See https://participedia.net/method/6426

CONCLUSION

In this chapter we showed that the principle of multi-stakeholder governance extends to the design of AI policy and presented some building blocks for inclusive policy design. AI can be seen as a general-purpose technology that is transforming the way we work and live. It is becoming increasingly difficult to opt out; we are only just beginning to understand the impact it will have on our lives; and most importantly, AI is constantly changing, learning, and advancing. Its implementation and growth create risks to human rights and will lead to value-driven questions and dilemmas. We have shown that multi-stakeholder approaches fit with developments entailing these ingredients. Society needs to act to ensure that it harnesses the benefits of AI in a responsible, sustainable way and negate the risks it poses to rights and freedoms. We are confident that a deliberative multi-stakeholder approach is key to designing AI policy for several reasons. First, deliberation will ensure that the resulting policy framework is based on a widely shared set of goals and values. Second, the policy framework will allow for quick adjustments through feedback from participants and an increased sense of ownership among stakeholders. Third, the participatory process increases awareness and builds capacity on AI. And fourth, new connections are forged between stakeholders, thereby positively contributing to further conversations on AI policy and to its development.

Finally, while participation is relevant at every stage of the policy process, we urge policymakers, civil society organizations, academics, the technical community, the private sector, and interested citizens to pay particular attention to stakeholder engagement at the start of policy processes. Knowledge of AI is scarce, the language in use is vague, and its impact is wide-ranging; for these reasons, increasing awareness, building capacity, and demystifying overly optimistic or negative images of AI should be key elements of any policy related to AI, as well as to any participatory policy process.

This chapter has highlighted five lessons for multistakeholder AI development. It is a condensed version of UNESCO and i4Policy (2022) which outlines ten building blocks for inclusive policy design. We hope these publications will equip the reader to push for the democratization of AI policy development.

REFERENCES

Adam, L., James, T. and Wanjira, A. M. 2007. *Frequently Asked Questions about Multi-Stakeholder Partnerships in ICTs for Development*. Association for Progressive Communications (APC). https://www.apc.org/sites/default/files/catia_ms_guide_EN-1.pdf

AlgorithmWatch. 2020. In the realm of paper tigers: Exploring the failings of AI ethics guidelines. April 28. https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/

Ambole, A., Koranteng K., Njoroge, P. and Luhangala, D. L. 2021. A review of energy communities in sub-Saharan Africa as a transition pathway to energy democracy. *Sustainability,* Vol. 13, No. 4, pp. 21-28. https://doi.org/10.3390/su13042128

Bäckstrand, K. 2006. Multi-stakeholder partnerships for sustainable development: Rethinking legitimacy, accountability and effectiveness. *European Environment*, Vol. 16, No. 5, pp. 290–306. https://doi.org/10.1002/eet.425

Barocas, S., and Selbst, A. D. 2016. Big Data's disparate impact. *104 California Law Review* No. 671. https://doi.org/10.2139/ssrn.2477899.

Bernal, P. 2016. Data gathering, surveillance and human rights: Recasting the debate. *Journal of Cyber Policy*, Vol. 1, September, pp. 1–22. https://doi.org/10.1080/23738871.2016.1228990

Birhane, A. 2020. Algorithmic colonization of Africa. *SCRIPTed*, Vol. 17, No. 2, pp. 389–409. https://doi.org/10.2966/scrip.170220.389

Black, E. and Richmond, R. 2019. Improving early detection of breast cancer in sub-Saharan Africa: Why mammography may not be the way forward. *Globalization and Health*, Vol. 15, No. 1, p. 3. https://doi.org/10.1186/s12992-018-0446-6

Bondi, E., Xu, L., Acosta-Navas, D. and Killian, J. A. 2021. Envisioning communities: A participatory approach towards AI for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 425–36. Virtual event, USA: ACM. https://doi.org/10.1145/3461702.3462612

Brockmyer, B. and Fox, J. A. 2015. Assessing the evidence: The effectiveness and impact of governance-oriented multi-stakeholder initiatives. Transparency and Accountability Initiative, September 20. https://papers.ssrn.com/abstract=2693379

Buhmann, A. and Fieseler, C. 2021. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, Vol. 64, February, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

Census of India. 2018. Census of India 2011: Language. New Delhi: Office of the Registrar General. https://www.jagranjosh.com/current-affairs/language-census-2011-surge-in-hindi-speakers-south-indian-language-and-urdu-speakers-decline-1530869001-1

Collingridge, D. 1981. *The Social Control of Technology*. New York: Palgrave Macmillan.

Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Dilhac, M.-A, Mai, V., Mörch, C., Noiseau, P. and Voarino, N. 2020. *Responsible Artificial Intelligence: A Guide for Deliberation*. Montreal, QC: Algora Lab – Mila. https://observatoire-ia.ulaval.ca/en/responsible-artificial-intelligence-a-guide-for-deliberation/

Dingwerth, K. 2008. Private transnational governance and the developing world: A comparative perspective. *International Studies Quarterly*, Vol. 52, No. 3, pp. 607–34. https://doi.org/10.1111/j.1468-2478.2008.00517.x

Faysse, N. 2006. Troubles on the way: An analysis of the challenges faced by multi-stakeholder platforms. *Natural Resources Forum*, Vol. 30, August, pp. 219–29. https://doi.org/10.1111/j.1477-8947.2006.00112.x

Feldstein, S. 2019. *The Global Expansion of AI Surveillance: Working Paper.* Carnegie Endowment for International Peace. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf

Fransen, L. W. and Kolk, A. 2007. Global rule-setting for business: A critical analysis of multi-stakeholder standards. *Organization*, Vol. 14, No. 5, pp. 667–84. https://doi.org/10.1177/1350508407080305

Garcia, E. 2019. The militarization of artificial intelligence: A wake-up call for the Global South. https://doi.org/10.2139/ssrn.3452323

García-López, G. A., and Arizpe, N. 2010. Participatory processes in the soy conflicts in Paraguay and Argentina. *Ecological Economics*, Vol. 70, No. 2, pp. 196–206. https://doi.org/10.1016/j.ecolecon.2010.06.013

GIZ India. 2021. *Handbook on Data Protection and Privacy for Developers of Artificial Intelligence (AI) in India.* New Delhi. https://www.dsci.in/content/privacy-handbook-for-ai-developers

Gleckman, H. 2018. *Multistakeholder Governance and Democracy: A Global Challenge*. London: Routledge.

Gorwa, R., Binns, R. and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, Vol. 7, No. 1, 2053951719897945. https://doi.org/10.1177/2053951719897945

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D. and Giannotti, F. 2018. A survey of methods for explaining black box models. *ArXiv:1802.01933 [Cs]*, June. http://arxiv.org/abs/1802.01933

Gwagwa, A., Kachidza, P., Siminyu, K. and Smith, M. 2021. Responsible artificial intelligence in sub-Saharan Africa: Landscape and general state of play. International Development Research Centre (IDRC). https://idl-bnc-idrc.dspacedirect.org/handle/10625/59997.

Gwagwa, A., Kraemer-Mbula, E., Rizk, N., Rutenberg, I. and De Beer, J. 2020. Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions. *The African Journal of Information and Communication*, No. 26, December, pp. 1–28. https://doi.org/10.23962/10539/30361

Habermas, J. 1989. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Translated by Thomas Burger. Studies in Contemporary German Social Thought. Cambridge, MA, USA: MIT Press.

Hall, P. and Gill, N. 2019. *An Introduction to Machine Learning Interpretability*. 2nd ed. O'Reilly Media. https://www.oreilly.com/library/view/an-introduction-to/9781098115487/

Hofmann, J. 2016. Multi-stakeholderism in internet governance: Putting a fiction into practice. *Journal of Cyber Policy*, Vol. 1, No. 1, pp. 29–49. https://doi.org/10.1080/23738871.2016.1158303

Hoogesteger, J. 2012. Democratizing water governance from the grassroots: The development of Interjuntas-Chimborazo in the Ecuadorian Andes. *Human Organization*, Vol. 71, No. 1, pp. 76–86. https://doi.org/10.17730/humo.71.1.b8v77j0321u28863

INDIAai. 2021. *Mitigating Bias in AI: A Handbook for Startups.* https://indiaai.s3.ap-south-1.amazonaws.com/docs/AI+Handbook_27-09-2021.pdf

ITU. 2018. *Assessing the Economic Impact of Artificial Intelligence*. 1st Issue Paper on Emerging Trends. International Telecommunications Union. http://handle.itu.int/11.1002/pub/81202956-en

Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol. 1, No. 9, pp. 389–99. https://doi.org/10.1038/s42256-019-0088-2

Kamakoti, V. 2018. *Report of the Artificial Intelligence Task Force*. Government of India. https://dpiit.gov.in/whats-new/report-task-force-artificial-intelligence

Kimaili, K. 2021. Lessons from deliberative democracy in Africa. Westminster Foundation for Democracy. August 25. https://www.wfd.org/commentary/lessons-deliberative-democracy-africa

Kwet, M. 2019. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, Vol. 60, No. 4, pp. 3–26. https://doi.org/10.1177/0306396818823172

Llansó, E., Van Hoboken, J., Leerssen, P. and Harambam, J. 2020. Artificial intelligence, content moderation, and freedom of expression. *The Transatlantic Working Group Papers Series*. Annenberg Public Policy Center of the University of Pennsylvania. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf

Magassa, L., Young, M. and Friedman, B. 2017. *Diverse Voices: A How-to Guide for Facilitating Inclusiveness in Tech Policy*. Tech Policy Lab, University of Washington. https://techpolicylab.uw.edu/wp-content/uploads/2017/10/TPL_Diverse_Voices_How-To_Guide_2017.pdf

Martens, W., van der Linden, B. and Wörsdörfer, M. 2019. How to assess the democratic qualities of a multi-stakeholder initiative from a Habermasian perspective? Deliberative democracy and the Equator Principles framework. *Journal of Business Ethics*, Vol. 155, No. 4, pp. 1115–33. https://doi.org/10.1007/s10551-017-3532-4

McKinsey Global Institute. 2018. *Notes from the AI Frontier: Modeling the Impact of AI on the World Economy*. Discussion Paper. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

MCTCI, Ministerio de Ciencia, Tecnología, Conocimiento e Innovacíon. 2021a. Consulta Pública de Inteligencia Artificial Informe de Resultados. Chile Government. https://minciencia.gob.cl/areas-de-trabajo/inteligencia-artificial/politica-nacional-de-inteligencia-artificial/proceso-de-elaboracion/

———. 2021b. *Política Nacional de Inteligencia Artificial*. Chile Government. https://minciencia.gob.cl/areas-de-trabajo/inteligencia-artificial/politica-nacional-de-inteligencia-artificial/

Mohamed, S., Png, M.-T. and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, Vol. 33, No. 4, pp. 659–84. https://doi.org/10.1007/s13347-020-00405-8

Moog, S., Spicer, A. and Böhm, S. 2014. The politics of multi-stakeholder initiatives: The crisis of the Forest Stewardship Council. *Journal of Business Ethics*, No. 128, pp. 469–493. https://doi.org/10.1007/s10551-013-2033-3

Moraes, T. G., Almeida, E. C. and Pereira, J. R. L. 2021. Smile, you are being identified! Risks and measures for the use of facial recognition in (semi-)public spaces. *AI and Ethics*, Vol. 1, No. 2, pp. 159–72. https://doi.org/10.1007/s43681-020-00014-3

Mudongo, O. 2021. Africa's Expansion of AI Surveillance: Regional Gaps and Key Trends. Policy Brief. Cape Town: Research ICT Africa. https://researchictafrica.net/publication/africas-expansion-of-ai-surveillance-regional-gaps-and-key-trends/

Murray, C. 2013. *Coming Apart: The State of White America, 1960–2010*. Illustrated edition. New York: Crown Forum.

Mustalahti, I. and Rakotonarivo, O. S. 2014. REDD+ and empowered deliberative democracy: Learning from Tanzania. *World Development*, Vol. 59, July, pp. 199–211. https://doi.org/10.1016/j.worlddev.2014.01.022

Nanz, P. and Steffek, J. 2005. Assessing the democratic quality of deliberation in international governance: Criteria and research strategies. *Acta Politica*, Vol. 40, No. 3, pp. 368–83. https://doi.org/10.1057/palgrave.ap.5500118

NITI Aayog. 2018. National Strategy for Artificial Intelligence #AIForAll. https://www.niti.gov.in/national-strategy-artificial-intelligence

———. 2021a. Approach Document for India. Part 1: Principles for Responsible AI. Responsible AI #AIForAll. https://indiaai.gov.in/research-reports/responsible-ai-part-1-principles-for-responsible-ai

———. 2021b. Approach Document for India. Part 2: Operationalizing Principles for Responsible AI. ResponsibleAI#AIForAll.https://indiaai.gov.in/research-reports/responsible-ai-part-2-operationalizing-principles-for-responsible-ai

OECD. 2020. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*. OECD. https://doi.org/10.1787/339306da-en

OECD.AI. 2021. National AI policies and strategies. https://oecd.ai/en/dashboards

OHCHR. 2018. *The Right to Privacy in the Digital Age.* A/HRC/39/29. United Nations High Commissioner for Human Rights. https://www.ohchr.org/EN/Issues/DigitalAge/Pages/ReportDigitalAge.aspx

Effoduh, J. O. 2020. 7 ways that African states are legitimizing artificial intelligence. *Open AIR*. October 20. https://openair.africa/7-ways-that-african-states-are-legitimizing-artificial-intelligence/

Oxford Insights. 2020. *Government AI Readiness Index 2020*. Ottawa: IDRC. https://www.oxfordinsights.com/government-ai-readiness-index-2020

Polli, F. 2019. Using AI to eliminate bias from hiring. *Harvard Business Review.* October 29, 2019. https://hbr.org/2019/10/using-ai-to-eliminate-bias-from-hiring

Prince, A. and Schwarcz, D. 2020. Proxy discrimination in the age of artificial intelligence and Big Data. *Iowa Law Review,* Vol. 105, No. 1257, p. 63.

Privacy International. 2019. *Guide to International Law and Surveillance 2.0.*

Renn, O., Ulmer, F. and Deckert, A. 2020. *The Role of Public Participation in Energy Transitions.* Cambridge, MA: Academic Press.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *ArXiv:1811.10154 [Cs, Stat]*, September. http://arxiv.org/abs/1811.10154

Scholte, J. A. 2020. *Multistakeholderism Filling the Global Governance Gap?* The Global Challenges Foundation. https://globalchallenges.org/multistakeholderism-filling-the-global-governance-gap/

Sloan, P. and Oliver, D. 2013. Building trust in multi-stakeholder partnerships: Critical emotional incidents and practices of engagement. *Organization Studies*, Vol. 34, No. 12, pp. 1835–68. https://doi.org/10.1177/0170840613495018

Smart Africa. 2021. *Blueprint: Artificial Intelligence for Africa*. Kigali: Smart Africa, GIZ and GFA Consulting. https://smartafrica.org/knowledge/artificial-intelligence-for-africa/

Souter, D. 2017. *Inside the Information Society: Multistakeholder Participation, a Work in Progress*. Association for Progressive Communications. https://www.apc.org/en/blog/inside-information-society-multistakeholder-participation-work-progress

The Netherlands Scientific Council for Government Policy. 2021. *Mission AI*. The New System Technology (English Summary). The Hague: WRR. https://www.wrr.nl/publicaties/rapporten/2021/11/11/opgave-ai-de-nieuwe-systeemtechnologie

Trajtenberg, M. 2018. Artificial Intelligence as the Next GPT: A Political-Economy Perspective. *NBER Chapters*. National Bureau of Economic Research. https://econpapers.repec.org/bookchap/nbrnberch/14025.htm

UNCTAD. 2021. *Digital Economy Report 2021*. New York: United Nations Publications.

UNESCO. 2019. *Steering AI and Advanced ICTs for Knowledge Societies: A Rights, Openness, Access, and Multi-Stakeholder Perspective*. Vol. 14. UNESCO Series on Internet Freedom. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000372132

———. 2021. *Artificial Intelligence Needs Assessment Survey in Africa*. Paris: UNESCO.

UNESCO and i4 Policy. 2022. Multistakeholder AI development: 10 building blocks for inclusive policy design. United Nations Educational, Scientific and Cultural Organization (UNESCO) and the Innovation for Policy Foundation (i4Policy). https://unesdoc.unesco.org/ark:/48223/pf0000382570

United Nations General Assembly. 2015. *Resolution Adopted by the General Assembly on 25 September 2015*. A/RES/70/1. United Nations. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/291/89/PDF/N1529189.pdf

United Nations Security Council. 2021. Letter dated 8 March 2021 from the Panel of Experts on Libya Established pursuant to Resolution 1973 (2011) addressed to the President of the Security Council. https://digitallibrary.un.org/record/3905159

Van der Spuy, A. 2017. *What If We All Governed the Internet?: Advancing Multistakeholder Participation in Internet Governance*. UNESCO Series on Internet Freedom 11. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000259717

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M. and Nerini, F. F. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, Vol. 11, No. 1, Article No. 233. https://doi.org/10.1038/s41467-019-14108-y

WFD (Westminster Foundation for Democracy). 2021. An Introduction to Deliberative Democracy for Members of Parliament. https://www.wfd.org/wp-content/uploads/2021/09/WFD_newDemocracy_An-introduction-to-deliberative-democracy-for-members-of-parliament_2021.pdf

Winfield, A. 2019. Ethical standards in robotics and AI. *Nature Electronics*, Vol. 2, No. 2, pp. 46–48. https://doi.org/10.1038/s41928-019-0213-6

Young, M., Magassa, L. and Friedman, B. 2019. Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, Vol. 21, No. 2, pp. 89–103. https://doi.org/10.1007/s10676-019-09497-z

OWNERSHIP AND MANAGEMENT OF LEARNING BEHAVIOR INFORMATION FOR AIED

SHITANSHU MISHRA

PhD, Information Technology Officer at UNESCO Mahatma Gandhi Institute of Education for Peace and Sustainable Development.

DAN SHEFET

Lawyer at the Paris Court of Appeal, France.

ANANTHA KUMAR DURAIAPPAH

PhD, Director at UNESCO Mahatma Gandhi Institute of Education for Peace and Sustainable Development.

SDG4 - Quality Education
SDG5 - Gender Equality
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

OWNERSHIP AND MANAGEMENT OF LEARNING BEHAVIOR INFORMATION FOR AIED

ABSTRACT

AI in Education (AIED) refers to the artificial intelligence (AI)-based tools, techniques, and methodologies used to automate processes that need to be carried out by computers in a technology-enhanced learning environment. In the context of AIED systems' development, "learner data" has become an indispensable commodity. Controlling this commodity's supply chain amounts to controlling AIED. But this dynamic begs the question of whether AIED is a common heritage of humankind. If so, then learner data will invariably be a common heritage of humankind. This issue raises the need to constitute a commons – the governance framework or mechanism – to manage learners' data as a collective heritage. This chapter presents insights into how the development of AIED systems is dependent on data. Furthermore, it argues why educational data should be regarded as a common heritage of humankind and proposes a structure of a commons that can manage such heritage.

INTRODUCTION

On November 10, 2021, UNESCO released its Futures of Education report at its 41st General Conference. The report urges for a new social contract on education founded on the principles of non-discrimination, social justice, respect for all life, human dignity, and cultural diversity. Calling for such a social contract implies citizens giving up some of their natural freedoms to the state or even an international or intergovernmental entity in return for specified services or goods agreed upon in the contract (Castiglen, 2015). In this case, the service rendered would be the provision of high-quality education in an equitable manner to all citizens as a "common good."

However, education will be fundamentally different from what it has been for the past 300 years when existing education systems emerged to meet the needs of the industrial revolution and economic growth. As we move into the digital world, now more commonly called the metaverse, future education systems will be predominantly digital. AI has the potential to be the primary "agent" to provide guidance for learners to improve their learning and henceforth their potentiality. Interestingly, current discussion

on AI is accompanied by conversations on ethics in AI. “Ethics in AIED” (Holmes et al., 2021) generally discusses risks of monitoring students, data privacy, informed consent, data interpretation, data ownership, data access, accountability, etc. However, in this chapter, we focus on the ownership and management of educational data.

If this new social contract on education in the metaverse is to ensure quality education in an equitable manner, a number of key parameters need to be addressed. One which we shall discuss in detail in this chapter is the acquisition process of learners’ “learning behavior information” (LBI), managing this data for equity, efficiency, and effectiveness while maintaining the highest privacy protocols of that data.

This LBI is distinct from the knowledge resources that learners use. For example, a mathematics module (a knowledge resource) used in classrooms falls under the latter category of resources that can be either privately owned or part of the open-access resources database. However, the learning experience by the learner going through the module is what we call LBI. By LBI, in this chapter, we refer to all kinds of learner data collected from different modalities, such as system-log, gaze data, physiological data, texts, images, videos, etc. This includes, for instance, tracking student performances, strategies, attention sequences, misconceptions, time spent on the question, the number of times reverting back to previous sections of the module, the number of attempts on section and quizzes, and, more recently, the emotions the learner is experiencing.

LBI is presently owned by the entity that owns and manages the learning platform on which the module is offered. This LBI is the “gene” pool from which AI algorithms linked to the learning platform draw to help learners with a more effective and efficient learning experience. This gene pool has the LBI data from other learners who have undergone similar learning experiences. As with all AI algorithms, the more data that are used for its training, the better the efficiency and effectiveness of the learning interventions that the AI might suggest.

The issue of equity arises when the quality of the pool of data used by entities owning and managing the data differs. This implies that the quality of learning by learners will vary depending on the entity owning the best pool of data. This, in turn, suggests that entities that have better data pools can offer a better learning experience but at a cost that might exclude a certain group of learners, particularly those from the marginalized and disadvantaged groups. Cost then determines who can and cannot access the benefits of pooled data, which is by our earlier definition a common heritage.

The challenge arises with the question of sharing the LBI among the various learning platforms and striving for equity. In addition to the four common arguments for sharing data – i.e., replicating research, making public assets available to the public, leveraging research investments, and facilitating research and innovation – as highlighted by Borgman, the use of AI in education suggests a fifth: improving learning by offering a personalized learning experience (Borgman, 2015; Margolis et al., 2014; Wilkinson et al., 2016). The option to offer multiple learning trajectories rather than just one is at the core of sharing data and using AIED.

The LBI is, in principle, owned by each individual. These data become valuable only when used in a collective manner when pooled with similar data from as many individuals as possible and collected over time. A large pool of data is a prerequisite for creating accurate AI models and, thereby, effective AIED systems. Each individual datum by itself is of little value. Moreover, the owner of the data – the individual – also does not benefit unless their information is pooled with many others before gains to learning can be reaped.

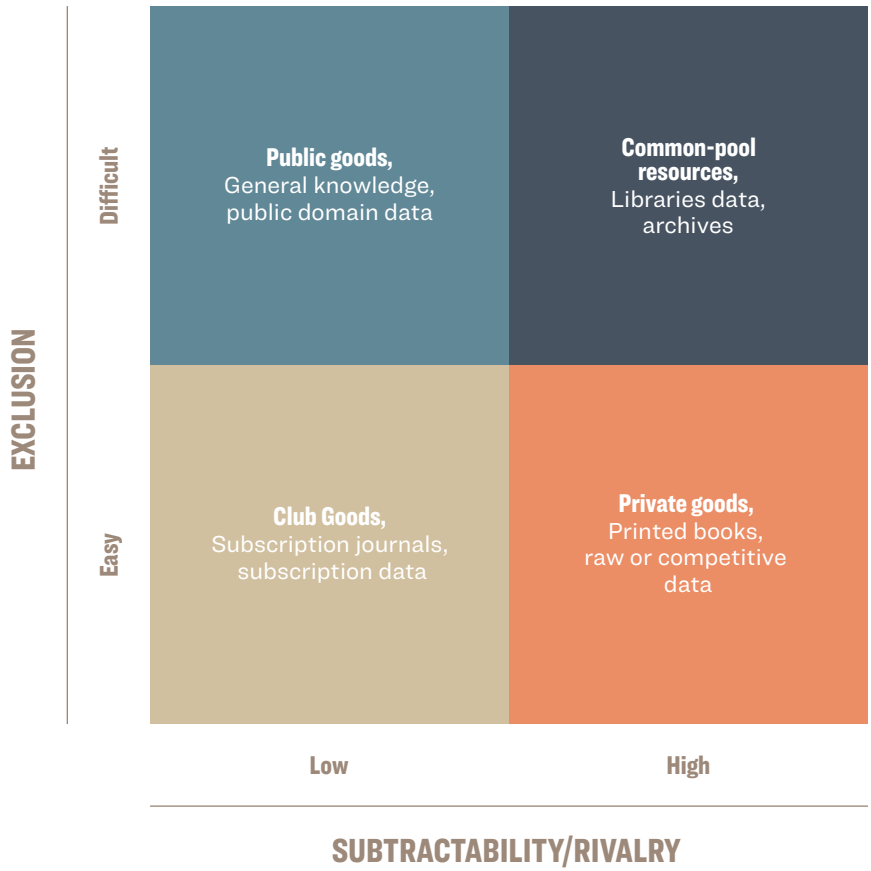
This, therefore, leads us to the quintessential question of how to bring all the information generated by learners distributed across space and time such that AI algorithms can use them to benefit the learners. The LBI generated by each learner, as mentioned earlier, is in principle owned by the learner. However, the collective LBI of all learners on a specific learning platform is presently owned by the entity that

owns the learning platform. This same entity might own the AI algorithm used to support learners or might have acquired licenses from AI companies to use specific AI algorithms to build a profitable AIED learning platform.

In addition to the factors such as quality, reliability, skewness (that leads to bias) of data in the database, the volume of the database, as mentioned before, determines the effectiveness of AIED-based learning platforms. The larger the LBI database, the better the learning intervention the respective AI algorithms offer. In addition to the LBI data, the quality of learning interventions will also depend on the quality of the AI algorithms. We now ask these two questions in this chapter; first, should the collective pool of LBI be owned by entities offering the learning platforms as a private good, a club good offered by a few entities, a public good, or a common-pool good; second, should the algorithms that are used to provide learning interventions also be a common heritage of humankind or a private, public, club, or common-pool resource?

| FIGURE 1 |

Types of goods. Source: Adapted from Ostrom and Ostrom, 1977; Hess and Ostrom, 2007.



Using the Hess-Ostrom classification shown in Figure 1, the present system of assigning ownership to the learning platforms essentially classifies LBI as club goods available to users who subscribe to the learning platform they are members of. This option produces sub-optimal outcomes as the data pool is restricted to those only on the learning platforms and usually leads to monopolies in the sector, producing inequitable outcomes, as highlighted earlier in the paper. Monopolies essentially collude to fix prices to capture the market and therefore exclude others from offering more competitive options to the consumer, who in this case is the learner. If education is to be a social contract, then relegating it to monopolies might not serve society as intended. The level of efficiency again becomes an issue if the LBI is treated as a private good, as defining the ownership of the information generated between learners and the learning platforms becomes a non-trivial task. This leaves us with the option of treating the LBI as a public or common-pool resource. In the former, governments might offer the service, but again problems associated with only governments providing a good with, well-intentioned regulations in an efficient manner, is questionable due to high transaction costs and the high probability of regulation capture, especially when key players in the sector are private, for-profit entities (Beales et al., 2017).

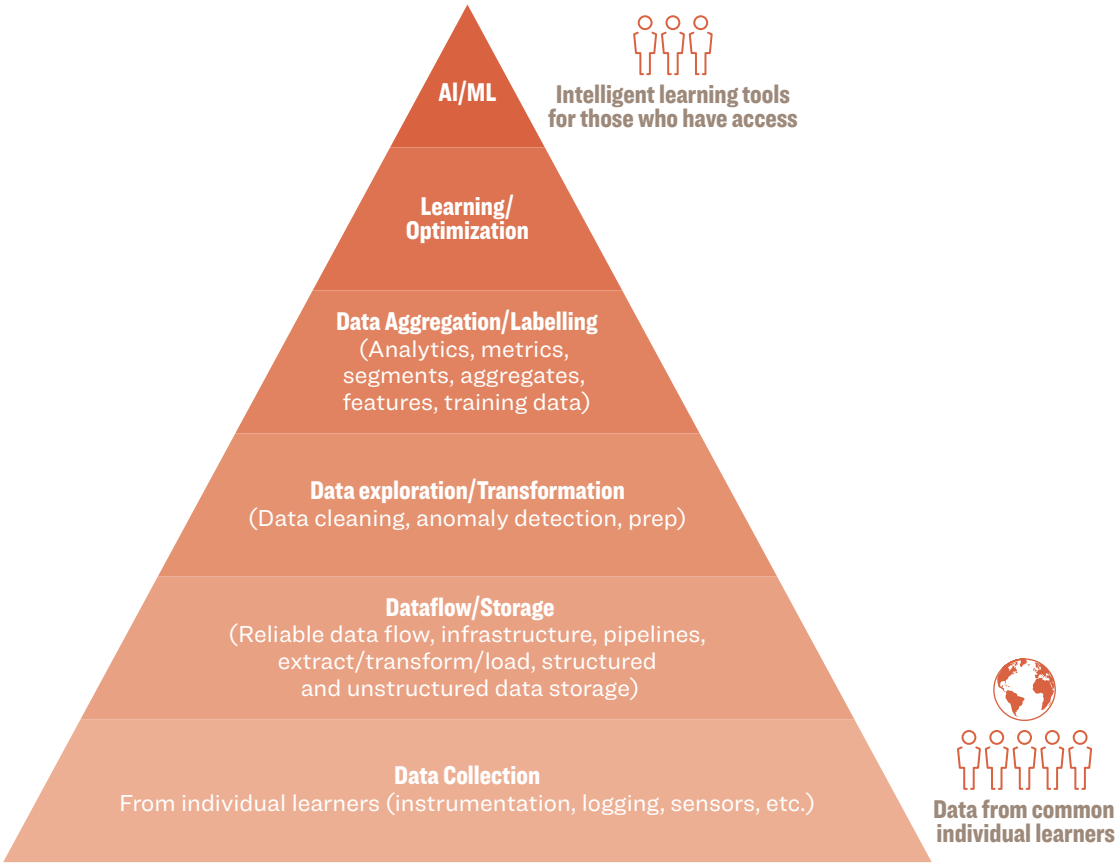
This requires establishing a governance structure that involves the learners, governments as well as learning platform entities that might transcend national boundaries and make the LBI globally pooled to the advantage of every learner across the globe. A common governance structure that allows every party that is a potential owner and user of the service lends itself well, based on the common-pool nature of LBI. The following section provides an in-depth analysis of the nature of LBI, how it is consumed in an AI-based learning platform, and the rationale for treating it as a common-pool resource.

LBI DATA AND AI IN EDUCATION

AIED refers to the AI tools, technologies, and methods used to automate the processes that computers need to perform in a technology-enhanced learning environment. Figure 2 illustrates the dependency of any AIED system on the data collected from the individual learners (LBI data).

| FIGURE 2 |

The data science hierarchy of needs (adapted from Rogati, 2017).



In Figure 2, we see that data is at the bottom of the hierarchy of needs of data science (Rogati, 2017). To discuss how the LBI data in AIED is used and should be used, it becomes imperative to unbox and understand how AI systems are built. This would provide a clearer picture of the contribution of LBI data in the process and the challenges that AIED system developers need to address while using the data to build any AIED-based learning platforms. The three major challenges in AI system development and implementation are lack of data access, lack of infrastructure, and lack of talents with the skill set needed to ensure effective and successful AI development (Ernst et al., 2019). This also applies to the AIED systems. Out of these, data availability seems to be a perpetual challenge. Availability of diverse and high-volume data is needed to ensure that an AI (or AIED) system is efficient and produces fair and equitable outcomes.

Any AI system development process typically starts with understanding and defining the problem that AI needs to solve. This is followed by ensuring the availability of data, which requires administering rich data collection processes. After ensuring sufficient data availability, AI developers need to ensure that the data are well stored and organized, such that they can be easily accessed for further processes. This step is followed by data exploration and preprocessing. Data exploration is needed to verify if the data represent the transpired events correctly, thereby allowing the evaluation of AI's data assumptions and understanding. The data cleaning process is crucial, as uncleaned data can lead to inaccurate training, producing wrong decisions, conclusions, and poor analysis, especially if the huge quantities of big data are in the picture. The data cleaning involves removing or updating incomplete, incorrect, duplicated, or irrelevant information. It also involves addressing data skewness and normalization of data (Jeni et al., 2013) to make it appear similar across all records and fields. Data cleaning maximizes the dataset's accuracy without necessarily tampering with the data available.

The data cleaning is followed by data aggregation and feature-engineering (Zheng and Casari, 2018), which involves extracting, generating, aggregating, reducing data features to better represent the underlying problem (e.g., being able to predict whether or not a learner is interested in reading further on a learning platform at any given time) for further machine learning. After the data cleaning and aggregation, AI developers develop and/or apply machine learning algorithms to train software models and systems that can intelligently support learning-teaching of any topic. After developing and testing an AIED system, one needs to work on its deployment to make it operational and accessible for the end-users (learners) and ensure regular maintenance and upgrades.

CONTRIBUTORS TO AN AIED SYSTEM

While developing an AIED system, the developers need to choose the AI approach to be implemented. This involves choosing between data-driven AI and model-driven AI. The choice determines the relationship between the data and the intelligence of an AI system after training. The data-driven way focuses on building a system that can identify the right answer based on having "seen" a large number of examples of question-and-answer pairs and "training" it to get to the right answer. This kind of AI is data-hungry. Some AI systems are powerful enough to generalize from limited training data and find actionable feature sets and decision criteria on their own, but many machine learning approaches (including deep learning) require very large data to produce meaningful results, and some demand their own type of experts to set them up.

Contrary to data-driven AI, which depends almost entirely on the collection and analysis of data to inform its decision-making, model-driven AI captures knowledge and enables decision-making via clear representation and rules that are informed by the knowledge and science of the problem domain for which the AI system is being developed. However**,** models (the science of learning in the case of AIED systems) continuously evolve. often using research carried out on empirical data collected from learning

contexts that again come from a large number of individual learners. Moreover, the knowledge of learning sciences is contributed not just by the private entity that develops (or owns) an AIED system, but is a result of a long-sustained effort of the communities of practitioners, researchers, and scientists in the education, cognitive sciences, and learning science domains. We should note that the knowledge-building process encompasses collaborative efforts at the scale of the entire scientific enterprise, such as community knowledge-building (Hong and Scardamalia, 2014; Scardamalia and Bereiter, 2006) bringing insights from multiple disciplines to bear on a complex problem.

Looking at the development process of an AIED system discussed above, it is clear that an AIED system cannot be built devoid of LBI data that better fits to be a common-pool good. Moreover, many of the AIED systems (especially model-driven systems) rely on models and knowledge from the domain(s), which may also be largely regarded as a common-pool good. However, we also see that an AIED system cannot be built without the contribution in the form of data collection, data management, data engineering, algorithm development, and deployment and maintenance of the AIED system. Therefore, in addition to the ownership of the pool of LBI data, it would be crucial to discuss the ownership of the AIED algorithms and systems, and the contributions of private and public entities in their development. However, in this chapter, we further restrict our discussion to the ownership and management of LBI data only.

MANAGING LBI AS A COMMON HERITAGE: KEY PRINCIPLES AND A GUIDING FRAMEWORK

Much of the research on knowledge commons is limited to the medical and public health sectors (Chatterjee et al., 2022). There is little research on the knowledge commons in the field of education. There is even less literature and experience in the field of LBI and the governance of this data. The present trajectory suggests a distorted market emerging in the face of monopolies and very little ownership of this data by the learners themselves and the use of this data for their benefit.

Gyuris (2014), while defining the term "knowledge commons," argues that "knowledge as a shared resource" requires information to be accessible and should allow potential recipients to internalize the information as knowledge. Therefore, knowledge cannot be a shared resource without a complex set of institutions and practices that provide potential recipients with the opportunity to acquire the necessary skills and preparations. Similarly, to treat LBI data as a common-pool heritage of humankind, we need to set up a governance structure that involves the learners as well as the learning platform entities that might transcend national boundaries and make the LBI globally pooled to the advantage of every learner around the globe. In the subsequent discussions, we propose directions, guidelines and questions that would be useful in setting up this governance structure.

Principles for managing a commons

Drawing from the work of Ostrom and colleagues (Hess and Ostrom, 2007; Ostrom, 2005), an international regime managing the LBI data as a common heritage and AI algorithms would need to ensure the following eight basic principles to ensure the three E's of Efficiency, Effectiveness, and Equity (Hess and Ostrom, 2007; Ostrom, 2005):

1. Define clear group boundaries.
2. Match rules governing the use of common goods to local needs and conditions.
3. Ensure that those affected by the rules can participate in modifying the rules.
4. Make sure the rule-making rights of community members are respected by outside authorities.
5. Develop a system for monitoring members' behavior, carried out by community members.
6. Use graduated sanctions for rule violators.
7. Provide accessible, low-cost means for dispute resolution.
8. Build responsibility for governing the common resource in nested tiers from the lowest level up to the entire interconnected system.

The strength of using Ostrom's framework is that it recognizes scale, a multitude of actors, a participatory process among the various stakeholders, and a graduated system of sanctions to ensure accountability and responsibility to minimize misuse of the data pooled. It is important to emphasize that commons does not refer to the resource but to the governance of a resource and in particular to a shared resource. In this case, sharing the pooled data offers the best solution for maximizing learners' learning experience.

In addition to the Ostrom principles framework, Frischmann et al. (2014) provide a useful guide to exploring and establishing a commons governance structure for what they call the knowledge commons, which in our case would be the LBI commons. Their knowledge commons framework builds on the Institutional Analysis and Development (IAD) framework (Ostrom, 2005; Hess & Ostrom, 2007), and they propose the following steps. For each step, we present lists of relevant questions or considerations that may be crucial in establishing a governance structure for LBI commons.

The definition of LBI

- What is the background context (legal, cultural, economics) of the LBI commons? How is the present LBI being collected, pooled, and used? Are there differences in how the LBI is collected, pooled, and used across different countries because of socio-economic and cultural disparities?
- What is the present ownership status of the LBI (patented, copyright, open, or other)?

The attributes of the resource

a. What is the nature of the resources to be pooled and how are they created or obtained?

b. What are the characteristics of the LBI? Are they rival or nonrival, excludable or non-excludable, tangible or intangible? Is there a shared infrastructure?

c. What are the technologies and skills needed to create, obtain, maintain, and use the LBI?

The community

d. Who are the members and their roles?

e. What is the degree and nature of openness for each type of community member and the general public?

Goals and objectives

f. What are the goals and objectives of the LBI commons and its members, including obstacles or dilemmas to be overcome?

g. What is the past history of the use of the LBI and if there are any special governance structures overseeing the pooling and use of the LBI?

Outcomes

h. What are the benefits for members and others (for example, innovation and creative outcomes, production, sharing and dissemination to a wider audience, social interactions resulting from the commons)?

i. What are the costs and risks associated with the commons, including negative externalities?

Governance

Conceptualizing the governance of the commons requires that we answer several questions in numerous dimensions. This includes thinking about the relevant action areas and how they relate to the goals and objectives of the commons and the relationships among various types of participants and with the general public. Another important aspect is the governance mechanism that includes constituting membership rules, LBI contribution or use standards and requirements, conflict-resolution mechanisms, sanctions for a rule violation, etc. For example, putting in place multilateral or bilateral agreements between countries where data is collected, and therefore very often according to local data privacy laws stored, and the jurisdiction or institution under which the pool of data will be based and the country where it will be utilized by the ultimate user (school, university, etc.).

The decision-makers and their selection process are essential aspects of a governance structure. Moreover, institutions and technological infrastructures that structure and govern decision-making are also crucial to the sustenance of the commons. For instance, an institution's management of the pool will, to the most considerable extent possible, be based on an algorithm or AI (which is rule-based), thereby reducing cost and enhancing quick turnaround. From an operational viewpoint, the main cost element will be maintenance.

Another important aspect that needs to be sorted out is the rules that will enable the usage of the LBI. For example, any recipient of data, whether a subscriber or an ad-hoc "client," must sign a contract restricting the use of the data to educational purposes. Such use must be free of charge for the learner, and the educational facility and a clear prohibition against any direct or indirect commercial use must be included. A dissuasive penalty clause should be included alongside arbitration.

The governance structure should also include rules describing how non-members interact with the commons. What institutions govern those interactions? An example could be the ICANN model (Christie, 2002), an organization not owned by one particular group, government, or corporation. Financing of the (low-cost) services could be achieved by a subscription model or on a transactional basis, i.e., a fee for the transfer of data that could be calculated against data volume.

Other important questions that a governance structure should address include, but are not limited to, informal norms that govern the commons and legal structures. For example, how do intellectual property, subsidies, contract, licensing, tax, antitrust apply in this structure?

STATE OF PLAY

The European Commission's initiative is currently the most important in terms of data sharing. The Proposal for a Data Governance Act, which was adopted by the European Commission on November 25, 2020 (European Commission, 2020), is currently considered to be the most ambitious draft regulation on the scope and obligations of data sharing. It deals extensively with the principle of data sharing and the procedural and institutional framework allowing efficient sharing to take place. The basic idea is that of creating data-sharing pools and regulated data-sharing intermediaries which will ensure that the rights of European "data subjects" are met when these operations take place. Unless there are strong reasons for the opposite, the data will be shared in an anonymized fashion.

Notwithstanding the importance of the draft regulation, it does not deal with data-sharing from private sources. It is limited to the public sector, and it is very much inspired by the idea of open government. Leaving out private sources is obviously the main drawback, but it is explained by the major economic consequences of compulsory data-sharing, which could not be achieved without compensation.

The Proposal for a Data Governance Act may in this respect be seen as the first step. In the United States, we do not find similar federal initiatives and the same is the case for India, where there is no data-sharing act. The Personal Data Protection Bill, now the "Data Protection Bill" (Lok Sabha, n.d.), which will most probably be passed shortly in India, does not include provisions similar to the EU draft.

CONCLUSION

> The discussion on the management of LBI has yet to be addressed formally by most organizations working on AI and the ethics of AI. Most of the focus has been on creating a code of ethics on developing algorithms and the use of data by respective data suppliers. However, the issues relating to the privacy of the LBI and the ownership of this information and its use in a collective manner has yet to be addressed. The present system of each learning platform providing these services offers a sub-optimal outcome and has the potential of increasing inequitable outcomes for learners. A commons approach might offer a solution to reap the three E's of efficiency, effectiveness and equity in the education sector.

REFERENCES

Beales, H., Brito, J., Davis, J. K. Jr., DeMuth, C., Devine, D., Dudley, S., Mannix, B., and McGinnis, J. O. 2017. *Government Regulation: The Good, The Bad, & The Ugly*. The Regulatory Transparency Project of the Federalist Society. https://regproject.org/wp-content/uploads/RTP-Regulatory-Process-Working-Group-Paper.pdf

Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: The MIT Press.

Castiglen, D. 2015. Introduction to the logic of social cooperation for mutual advantage: The democratic contract. *The Political Studies Review*. Vol. 13, No. 2, pp. 161–175.

Friend, C. n.d. Social Contract Theory. Internet encyclopedia of Philosophy. https://iep.utm.edu/soc-cont/

Chatterjee, A., Kuiper, M., and Swierstra, T. 2022. Dealing with different conceptions of pollution in the Gene Regulation Knowledge Commons. *Biochimica et Biophysica Acta (BBA) – Gene Regulatory Mechanisms*, Vol. 1865, No. 1. https://doi.org/10.1016/j.bbagrm.2021.194779

Christie, A. 2002. The ICANN Domain-Name Dispute Resolution System as a model for resolving other intellectual property disputes on the internet. *Journal of World Intellectual Property*, Vol. 5, No. 105, pp. 105–117.

Ernst, E., Merola, R. and Samaan, D. 2019. Economics of artificial intelligence: Implications for the future of work. *IZA Journal of Labor Policy*, Vol. 9, No. 1, pp. 1–35.

European Commission. 2020. Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (*Data Governance Act*). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

Frischmann, B. M., Madison, M. J. and Strandburg, K. J. (eds.). 2014. *Governing Knowledge Commons.* Oxford: Oxford University Press.

Gyuris, F. 2014. Basic education in communist Hungary: A commons approach. *International Journal of the Commons*, Vol. 8, No. 2, pp. 531–553.

Hess, C. and Ostrom, E. 2005. A framework for analyzing the knowledge commons. In: *Understanding Knowledge as a Commons: From Theory to Practice*., eds. C. Hess and E. Ostrom, 2007. Cambridge, MA: MIT Press.

Hess, C., and Ostrom, E. (eds.). 2007. *Understanding Knowledge as a Commons: From Theory to Practice*. Cambridge, MA: MIT Press.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., (...) and Koedinger, K. R. 2021. Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pp. 1-23. https://doi.org/10.1007/s40593-021-00239-1

Hong, H. Y., and Scardamalia, M. 2014. Community knowledge assessment in a knowledge building environment. *Computers & Education*, No. 71, pp. 279–288.

Jeni, L. A., Cohn, J. F., and De La Torre, F. 2013. Facing imbalanced data – Recommendations for the use of performance metrics. *2013 Humaine association conference on affective computing and intelligent interaction,* IEEE, pp. 245–251.

Lok Sabha. n.d. The Personal Data Protection Bill, 2019. Bill No. 373 of 2019. http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J. et al. 2014. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: Capitalizing on biomedical big data. *Journal of the American Medical Informatics Association* Vol. 21, No. 6, pp. 957–958. DOI: https://doi.org/10.1136/amiajnl-2014-002974

Ostrom, E. 2005. *Understanding Institutional Diversity.* Princeton, N.J.: Princeton University Press.

Ostrom, V. and Ostrom, E. 1977. Public Goods and Public Choices. In E. S. Savas (ed.), *Alternatives for Delivering Public Services: Toward Improved Performance*, pp. 7–49. Boulder, CO: Westview Press.

Rogati, M. 2017. The AI hierarchy of needs. Hacker Noon. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Scardamalia, M. and Bereiter, C. 2006. Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (ed.), *Cambridge Handbook of the Learning Sciences,* pp. 97–118. New York: Cambridge University Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., (...) and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, Vol. 3, No. 1, pp. 1–9.

Zheng, A., and Casari, A. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc.

AUTONOMOUS WEAPONS AND DEEPFAKES: THE RISKS OF THE ONGOING WEAPONIZATION OF AI AND THE URGENT NEED FOR REGULATION

BRANKA MARIJAN

Senior Researcher at Project Ploughshares. Board member of the Peace and Conflict Studies Association of Canada (PACS-Can). She holds a PhD from Balsillie School of International Affairs.

WANDA MUÑOZ

International Consultant in human rights and disarmament. She holds a Master of International Affairs from Columbia University and Sciences Po Paris and is a member of the Human Security Network in Latin America and the Caribbean (SEHLAC), the Feminist AI Research Network, and the Global Partnership on AI.

SDG3 - Good Health and Well-being
SDG5 - Gender Equality
SDG9 - Industry, Innovation and Infrastructure
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG13 - Climate Action
SDG15 - Life on Land
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

AUTONOMOUS WEAPONS AND DEEPFAKES: THE RISKS OF THE ONGOING WEAPONIZATION OF AI AND THE URGENT NEED FOR REGULATION

ABSTRACT

Soon, technology powered by artificial intelligence (AI) could independently launch wars and promote hate crimes. Already, more than 130 military systems can now autonomously track targets. AI is increasingly being used in semi-autonomous weapons and sophisticated deep-fake technologies that could escalate conflict and promote global instability. But the real dangers posed by the weaponization of such technologies have been largely ignored at national and international discussions on ethical and responsible applications of AI. And while they are discussed at arms control and disarmament forums, some participants seem more intent on acquiring these weapons than regulating or banning their use.

Fortunately, it's not too late to ensure that AI is used to benefit the majority of the world's people, not only oppressors and autocrats. A growing number of researchers, policy analysts, and members of civil society are eager and able to develop and promote measures that will lead to effective regulation. Voices of AI researchers are particularly needed to ensure the ethical development of new technologies.

What are the top priorities? First, a legally binding instrument prohibiting weapons that cannot be used with meaningful human control, and those that would target human beings; and regulating all other autonomous weapons. Second, technical responses to deepfakes that ensure that manipulated content is flagged. Third, regulations that protect human rights and prohibit applications that promote gender-based violence and other hate crimes.

INTRODUCTION

Drones capable of targeting individuals without a human operator in control. Sophisticated digital identification technologies in the hands of non-state armed groups and human rights abusers. Manipulated videos of political leaders providing statements that they never made.

Even a few years ago, such scenarios would have been, and indeed were, dismissed as science fiction or fear mongering. Yet, these seemingly dystopian uses of new technologies have all occurred in some form and are likely to continue to occur and expand if their unchecked development is not addressed with sound governance responses. Simply put, (Artificial Intelligence) AI is already being weaponized in direct uses by military and security institutions as well as in malicious uses by non-state armed actors. AI systems are increasingly adopted in the defense sphere as well as in intelligence gathering and analysis. Moreover, AI tools to manipulate images, video and audio have become more advanced and accessible, raising concerns about impacts on the erosion of public trust as well as posing a risk to international stability (UNIDIR, 2021). At the moment this chapter was written, in 2021, there is a clear governance gap to address the many risks associated with the weaponization of AI. This gap requires the urgent attention of the AI community, civil society, governments, and international organizations.

Technology has long outpaced regulation, and AI-enabled applications are no exception. Weaponized AI is too often perceived as a more distant threat. The fact that many of the advancements in AI are occurring in the private sector, with leading technological companies at the forefront of the research and development in the field, can also pose a regulatory challenge since they are further away from public scrutiny and government investments. Moreover, the multi-use nature of the technology at times obscures implications of potentially malicious applications. A growing global competition amongst leading militaries has emerged which places AI at the center of future military capabilities and, as a result, increased funds for research and development for military uses (Keller, 2021). In spite of this increasingly acknowledged military AI competition, the military and security applications of AI are generally excluded from broader discussions and commitments on ethical and responsible AI. For instance, the United Nations Educational, Scientific and Cultural Organization (UNESCO), the Organization for Economic Co-operation and Development (OECD) and the European Commission all exclude military applications from the mandates of their work on AI. While there is growing attention in certain international forums and at the United Nations, the calls for regulation and effective policy responses initially came from civil society and technical experts.

Against this background, and continued advancements made in AI and related fields such as robotics, the weaponization of AI deserves much greater attention. In particular, following Burton and Soare (2019, p. 4), the weaponization of AI is focused here on two aspects, "(a) how AI is and might be incorporated into weapons systems and platforms, and (b) how AI technologies may be used with ill-intent to cause harm in the international arena." As such, the first section highlights key concerns regarding the development of autonomous weapon systems, understood as systems where there is no human control over the critical functions of target selection and engagement of targets in the application of force. The second section focuses on deepfakes: synthetic media where images, video, audio, and even text, are manipulated in various ways to convince consumers that they are in fact real. The two dimensions of weaponized AI are highlighted to shed light on the specific issues but also on the possible and necessary responses. Each of these areas of AI development raises similar concerns about conflict escalation, increased threat of the use of force, global instability, and making access to justice for civilian victims even more difficult. In contrast to some of the literature (see Burton and Soare, 2019), certain uses of weaponized AI are not seen as inevitable, rather the chapter highlights the windows of opportunity that still exist to regulate these technologies and prevent misuse.

AUTONOMOUS WEAPONS SYSTEMS

This section examines the concerns raised by the development of autonomous weapon systems. It then describes ongoing diplomatic discussions on this issue, including some of the ways in which the shared understanding of autonomous weapons has advanced as well as challenges that have prevented the adoption of a regulatory instrument. This section also refers to the lack of coherence between the discussions on autonomous weapons and the ongoing efforts on developing ethical AI. Finally, it explains why an international treaty on autonomous weapons systems is a critical part of the multilateral response to this issue.

What are autonomous weapons systems?

AI-enabled autonomy is no longer a mere possibility. Weapons systems that once activated can select and engage targets and apply force on their own without any human control are already being developed. In the "Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions," Christof Heyns highlighted the wide-ranging concerns of such weapon systems, including questions regarding the extent to which they could comply with International Human Rights Law and International Humanitarian Law (IHL) (UN Human Rights Council, 2013).

Questions regarding accountability and the role of human operators in the critical functions of selection and engagement of targets are at the core of the calls for regulation. Existing systems are able to search for specific targets and then engage those targets with explosives (see Figure 1). Other systems, such as the SGR-A1 sentry robots in the demilitarized zone between North and South Korea, are less sophisticated. The sentries have cameras, heat and motion detectors, as well as pattern recognition software that allow the system to recognize an intruder. The SGR-A1 can engage the target with a light machine gun from some 800 meters away. Currently, the loitering munitions and the SGR-A1 (see Figure 1) are all under the control of human operators and fully autonomous weapons do not yet exist. But the role of the human operators exists on a sliding scale, and the type and quality of control they exert over the systems is being diminished by new technical possibilities as well as the push for faster decision-making. As Gould (2021) argues, there is an ongoing datafication of warfare that increasingly relies on remote methods such as drone footage, satellite phones, surveillance, and collection of metadata to label what normal or abnormal behavior in complex situations of conflict that cannot be reduced to such elements.

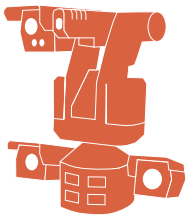
While there is no agreed-upon definition adopted internationally yet, there is an emerging understanding around certain characteristics of autonomous weapons systems. Autonomous weapons would incorporate preprogrammed target profiles and technical indicators that would be recognized through the weapons' sensors (Moyes, 2019). These would generate data based on the environment, instead of a user's input. They would process and analyze such data and determine what actions to take. They would also "apply force" – for instance, fire or launch a missile – if its analysis concludes that certain pre-programmed conditions have been met. Similarly, in 2021 the International Committee of the Red Cross (ICRC) highlighted, that autonomous weapon systems would be those that select and apply force to targets without human intervention, which means that the user does not choose, or even know, the specific target and the precise timing and/or location of the application of force.

| FIGURE 1 |

Examples of existing systems. Source: Pax Netherlands (2019)

SGR-A1

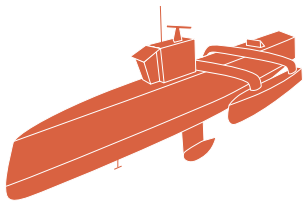
Made by: Hanwha (South Korea)
Sold to: South Korea



This stationary robot, armed with a machine gun and a grenade launcher, operated along the border between North and South Korea. It can detect human beings using infra-red sensors and pattern recognition software. The robot has both a supervised and unsupervised mode available. It can identify and track intruders, with the possibility of firing at them.

SEAHUNTER

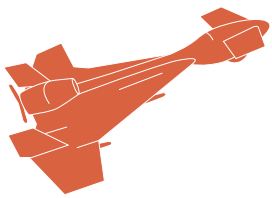
Made by: Pentagon's DARPA (United States)
Sold to: Under development



This 40 m long self-navigating warship is designed to hunt for enemy submarines and can operate without contact with a human operator for 2-3 months at a time. It is currently unarmed. US representatives have said the goal is to arm the Sea Hunters and to build unmanned flotillas within a few years. However, it has been said any decision to use offensive lethal force would be made by humans.

HARPY

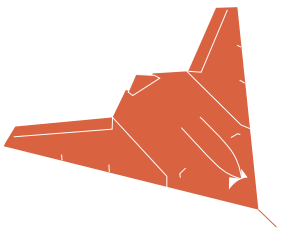
Made by: Israel aerospace industries (Israel)
Sold to: China, India, Israel, South Korea and Turkey



This 2.1 m long "loitering" missile is launched from a ground vehicle. It is armed with a 15 kg explosive warhead. The Harpy can loiter for up to 9 hours at a time, searching for enemy radar signals. It automatically detects, attacks and destroys enemy radar emitters by flying into the target and detonating.

NEURON

Made by: Dassault aviation (France)
Sold to: Under development



This 10 m long stealth unmanned combat aircraft can fly autonomously for over 3 hours for autonomous detection, localization, and reconnaissance of ground targets. The Neuron has fully automated attack capabilities, target adjustment, and communication between systems.

Such a definition is agreed upon by a growing number of calling for a legally binding instrument on autonomous weapons that includes both prohibitions and positive obligations (Campaign to Stop Killer Robots, 2021; Human Rights Watch, 2021). At the same time, there is still much debate on the precise definition of autonomous weapons as well as pushback from some states that do not align themselves with the views of the majority of states and civil society. In fact, the lack of consensus on this matter is one of the issues used by some countries as a reason—arguably, a pretext—to avoid starting a negotiation of a legally binding instrument to restrict the development of autonomous weapons systems (Sauer, 2021). This debate regarding definitions overlooks the fact that an agreed definition is not required prior to launching such negotiations, as has been the case in other disarmament processes (Devoto et al., 2021), including the one that led to the prohibition of cluster munitions.[67]

It is also important to note that although discussions at the Convention on Certain Conventional Weapons (CCW) generally refer to lethal autonomous weapons systems (LAWS) this is not a widely accepted phrasing. Many countries, the ICRC, and civil society organizations suggest that the qualifier of "lethal" should be dropped from this concept. As these countries and groups outline, lethality is not an intrinsic characteristic of a weapon; even without being lethal, the use of certain weapons may violate IHL if they cause unnecessary injuries or civilian damage; and those that are considered "defensive" weapons may also result in violations of IHL. As such, we use the broader term, "autonomous weapon systems" or "autonomous weapons" in our discussion of the issue, only using LAWS when citing the CCW process.

The multiple concerns raised by autonomous weapons systems

The concerns raised by autonomous weapons systems can be analyzed from the following perspectives:

Ethics

As a matter of principle, life-or-death decisions should not be delegated to a machine. Autonomous weapons systems would, by definition, lack the human capacity to analyze cultural contexts and situations of conflict, and to understand what it means to take a human life. Allowing machines to make such decisions undermines human dignity. As Wallach (2013) points out, autonomous weapons should be deemed a *mala in se* – or an evil in itself –, given that they "lack discrimination, empathy, and the capacity to make the proportional judgments necessary for weighing civilian casualties against achieving military objectives. Furthermore, delegating life and death decisions to machines is immoral because machines cannot be held responsible for their actions." This means that, in addition to the fact that the use of autonomous weapons may result in the killing of civilians by accident, such weapons should be subject of regulation because, based on the principle of human dignity, no one's life – including combatants' lives – should be endangered by a machine.

Additionally, a feminist approach to ethics brings to the forefront of any debate the lived experience of persons affected or potentially affected by the topic being discussed (Palmer, n.d.). From this perspective, it is also important to discuss the ethics of autonomous weapons through the perspectives of countries and populations affected by conflict, which would likely be the first to suffer from the use of these weapons. Their priorities and assessment of what is ethically acceptable or not would certainly be quite different to those presented by states that are the main producers of weaponry.

67. For more on how the Convention on Cluster Munitions was negotiated, see: Borrie, J. 2009. *Unacceptable Harm: A History of How the Treaty to Ban Cluster Munitions was Won*. UNIDIR. https://www.unidir.org/publication/unacceptable-harm-history-how-treaty-ban-cluster-munitions-was-won

International Humanitarian Law (IHL)

The use of autonomous weapons would certainly lead to violations of IHL, including the following principles that require human judgment: the principle of distinction between civilians and combatants, and between civilian objects and military objectives; and the principle of proportionality, which requires an assessment of whether an attack may be expected to cause civilian casualties or damage to civilian objects which would be excessive in relation to direct military advantage (ICRC, n.d.).

Additionally, the more autonomy is embedded in a weapon, the more difficult it will be to establish responsibility and accountability to access remedy and reparation for any victims, and to ensure there are consequences for the perpetrators of IHL violations. Liability could be attributed to different stakeholders: the data collectors, the programmers, the commanders, or the final user. However, it is not clear how such liability would be determined when multiple individuals are engaged in the building of systems and their use. If a system is making decisions outside of human control, it is also unlikely that a human could be held accountable for its actions. This would create more challenges for victims to access their rights – for which they already face enormous obstacles.[68] From a humanitarian perspective, we must also consider the psychological and economic impact of being attacked for target populations already traumatized by conflict. Also, the specific impact of an attack by autonomous weapons, considering the well-documented impact of existing means and methods of remote warfare.[69]

Human rights

If autonomous weapons are developed, they could be deployed not only in situations of conflict between states, but also at the national level, through police or national security institutions. This could lead to violations of human rights, such as the right to life, the right to a remedy and the right to privacy. Arbitrary arrests or detention as well as potential infliction of harm against individuals identified by autonomous systems are just some potential scenarios. Concerns about potential misuse and human rights implications of security institutions using facial recognition technologies have resulted in companies such as Amazon, IBM, and Microsoft calling for or establishing moratoriums on their use by police forces (Dastin, 2021; Allyn, 2020; Greene, 2020). The potential misuse of facial recognition technologies by police services has brought to the fore the need for more stringent regulation where risks and misuse worries are acute.

Social bias

It has been well documented that biases permeate AI applications such as facial recognition (Buolamwini and Gebru, 2018). In particular, one study showed that error rates were higher for darker-skinned women than for light-skinned men. In some programs the error rates for the former group was more than 34 percent (Buolamwini and Gebru, 2018). Indeed, AI applications beyond facial recognition have been demonstrated to incorporate and amplify social biases, particularly based on race and gender, in sectors such as education, health, employment, social housing, and tools for policing. There is no reason why such bias would not be translated, as well, into autonomous weapons using similar

68. For more on the accountability challenges posed by autonomous weapons systems, see: Human Rights Watch, 2015 *Mind the Gap: The Lack of Accountability for Killer Robots* Available at: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

69. See for instance Sharkey, 2019 and Safer World on the impact of remote warfare on mental health. SaferWorld (no date). *Warpod episode 8: Remote Warfare: Interdisciplinary Perspectives* https://www.saferworld.org.uk/multimedia/saferworldas-warpod-episode-8-remote-warfare-interdisciplinary-perspectives

technologies – only with life and death consequences (Díaz and Muñoz, 2019; Ramsay-Joynes, 2019).[70] This is particularly the case as Horowitz (2020) has argued that most militaries are likely to adopt "general purpose applications based on related algorithms in the commercial world." Hence, the concerns about bias will then also translate into military contexts.

In addition to potential violations of IHL and IHRL, autonomous weapons that would target humans and whose operation depends on such applications, could have a disproportionate impact among populations that are already amongst the most marginalized, such as people of color, women, persons with disabilities and LGBTIQ+. For example, these tools applied in conflict settings could be more likely to misidentify women of color in the targeting process; or to mislabel assistive devices, such as crutches, as a weapon. A person who uses a wheelchair, a cane, or walker may have a different speed, height, and possible reactions from those of the rest of the population – these and other elements of diversity in the population would probably not be considered in datasets or programming of autonomous weapons systems. Or, due to the speed of decision-making enabled by AI systems, that may result in a determination of suspicious behavior.

It is also important to consider that a large number of the people who die at the hands of the police in the United States are racialized men with some kind of impairment: estimates suggest that this population makes up some one third to one half of all those who die at the hands of the police (Abrams, 2020). This is no coincidence: racism intersects with ableism to create intersectional discrimination. In the case of autonomous weapons, the consequences of this systemic discrimination would be a matter of life and death.

International security

Some states portray autonomous weapons – and other methods of remote warfare – as increasingly clear and precise. Such claims should be analyzed in light of the impact of current methods of remote warfare, such as drone strikes. In the most recent tragedy, an investigation by the New York Times suggests that the United States targeted a humanitarian worker in a drone strike that killed nine more persons, including seven children, based on incorrect analysis that led them to identify his activities as "suspicious moves" (Koettl et al., 2021). The assessment of Knowles and Watson (2018) that remote warfare is certainly not precise nor less horrendous and traumatic for the victims of these weapons is also relevant for autonomous weapons. Furthermore, it leads to the "remote war paradox," which means that countries with autonomous weapons could engage more easily in war, since they face less of their own casualties – disregarding the impact on the victims. Once this technology exists, it could be replicated, used, and expanded by non-state and illegal armed groups. Moreover, hacking and adversarial attacks could be potentially dangerous as the impact of the systems will not be as easily controlled. And, as has been argued by Russel (2021), autonomous weapons could easily become weapons of mass destruction because they require no human supervision: pushing one single button, in itself, could result in launching a mass attack of thousands or millions of networked autonomous weapons systems.

Balance of power

Accepting autonomy in the critical functions of weaponry would also have an impact on the balance of power giving already militarized states first-mover advantage that is the ability to wield greater geopolitical power. As Bengio (2019) has argued, "essentially, AI is a tool that can be used by those in power to keep that power, and to increase it". States that would be among the first to develop these

70. For an analysis of the potential impact of autonomous weapons amongst marginalized populations specifically in Latin America: Díaz and Muñoz (2019): https://bit.ly/ArmasInterseccionalidad

technologies and achieve advantage over others would have a disproportionate control over global security. As Horowitz (2020) notes, one way that such advantage could be achieved is by developing a "general algorithm that could write other algorithms, operate in many domains and avoid the problem of catastrophic forgetting (forgetting previous learning after acquiring new information in a different area)." The appeal of such an advantage is prompting greater investments in developing these technologies among the Great Powers, China, and the US, but also a wider number of countries with smaller militaries (Horowitz, 2020).

Autonomous weapons are not simply a military or technical issue that is to be left to the military technical experts. Although these voices tend to be amplified, the discussion regarding autonomous weapons needs to be more inclusive given the wider impact on global security (Marijan, 2018). Moreover, the prominence of AI and the military advantage states believe the technology could afford them is requiring advocacy outside of these military circles that express fear about being left behind or at a disadvantage. Interestingly, concerns about being left behind or other countries developing the technologies ahead of currently more powerful states is also what is prompting the further investment and push for these technologies resulting in a vicious circle.

The inadequate international response to autonomous weapons systems

It is precisely advocacy by several groups that has resulted in the growing attention on the risks posed by autonomous weapons. These include the Campaign to Stop Killer Robots (with civil society organizations in more than 70 countries), Nobel Peace Prize laureates (Nobel Women's Initiative, 2014), the European Parliament (2018), the United Nations Secretary General (Bugge, 2018), the Alliance for Multilateralism (2019) and thousands of experts in AI, ethics and international security (Future of Life Institute, 2015; International Committee for Robot Arms Control since 2009). Importantly, the Montreal Declaration for Responsible Development of Artificial Intelligence (2010) signed by 187 organizations affirms in its principle 9.3: "The decision to kill must always be made by human beings, and responsibility for this decision must not be transferred to an artificial intelligence system".

In addition, lethal autonomous weapons systems have been discussed at the United Nations Convention on Prohibition or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001 (CCW) since 2014.[71] The High Contracting Parties have been meeting regularly for seven years now, and in 2019 adopted a list of Guiding Principles which summarized the understandings and agreements reached in the past years (CCW, 2019).

While a growing number of countries have called for a prohibition of lethal autonomous weapons systems,[72] others maintain that it would be premature to launch such negotiations given that they do not yet exist. In their view, the direction that the development of these systems will take is not yet clear and, as such, it is too early to craft a regulatory regime applicable to their use. This disregards

71. Although the mandate of the CCW relates to "lethal" autonomous weapons systems, the authors refer to such weapons as "autonomous weapons systems" because lethality should not be used as a criteria to define these weapons. Lethality is not a defining characteristic in IHL and keeping this qualifier in would set a negative precedent and go against IHL. For more: Muñoz, W (2021). It's about more than autonomous weapons systems. International Human Rights Clinic, Harvard Law School. https://humanitariandisarmament.org/2021/08/30/it-is-about-more-than-autonomous-weapons-systems/

72. As of August 2021, 31 countries have called for a ban of lethal autonomous weapons systems. Source: Human Rights Watch. 2021. *Killer Robots: Urgent Need to Fast-Track Talks Shared Vision Forms Sound Basis for Creating a New Ban Treaty* https://www.hrw.org/news/2021/08/02/killer-robots-urgent-need-fast-track-talks

what is already known on the risks of remote warfare; the risks related to bias and challenges of accountability in AI applications; and the precedent of the prohibition of blinding lasers, which was adopted preemptively in 1995.

In the seven years of CCW discussions, only some countries have made reference to the existing advances being made in various AI-related technologies and how these could be used in weapon systems. For example, increasingly autonomous uncrewed ground vehicles, aerial vehicles as well as loitering munitions have received surprisingly scant attention. The current lack of an international regulatory framework means that such developments can continue without a clear guide of what is legal, and what is morally acceptable.

Furthermore, the lack of a regulatory response from the international community is out of step with international and regional frameworks on artificial intelligence, to which several countries have committed so far. These include (Muñoz, 2020):

- The **UNESCO Recommendations on the Ethics of Artificial Intelligence** (2021, p. 8, 10), which indicates that "where decisions are understood to have an impact that is irreversible or difficult to reverse or may involve life and death decisions, final human determination should apply" and that "as a rule, life and death decisions should not be ceded to AI systems."
- The **Resolution 473 of the African Commission on Human and Peoples' Rights** (2021), which "appeals to State Parties to ensure that all AI technologies, robotics and other new and emerging technologies which have far-reaching consequences for humans must remain under meaningful human control to ensure that the threat that they pose to fundamental human rights is averted... The emerging norm of maintaining meaningful human control over AI technologies, robotics, and other new and emerging technologies should be codified as a human rights principle."
- The **G20 Human-centered AI principles** and the **OECD Ethical principles for AI** (2019), which affirm that "AI actors should respect the rule of law, human rights and democratic values. These include freedom, dignity, autonomy, privacy, data protection, non-discrimination, equality, diversity, fairness, social justice, and internationally recognized labor rights."
- The **EU Commission's proposal for new rules and actions for excellence and trust in AI** (2021), which proposes a prohibition of practices including "all those AI systems whose use is considered unacceptable as contravening Union values, for instance by violating fundamental rights."
- Other regional and national references such as the **Charter of Ethics on Emerging Technologies in the Arab Region** (2019), which aims to help identify means to "guide science and technology towards the right course, steering them away from unethical trends and practices that are harmful to humans and the surrounding environment;" and a statement from the **Republic of Korea** on the ethics of AI focusing on promoting harmony leaving no one behind (Republic of Korea, 2020).

In addition, the Global Partnership on AI (GPAI) is carrying out excellent initiatives in areas such as contributions to reducing climate change, to the pandemic response, and to the achievement of the United Nations Sustainable Development Goals. But it does not address the issue of AI in weaponry. Similarly, the European proposal on AI says that "this regulation shall not apply to AI systems developed or used exclusively for military purposes" (European Commission, 2021, p. 39). It seems, then, that the international community is leaving the decision regarding the legality and legitimacy of delegating life and decisions to autonomous functions exclusively to the CCW, even though this issue has much wider ethical implications for humanity as a whole.

The September 2021 meeting, *Safeguarding Human Control over Autonomous Weapons Systems,* organized by the Austrian Minister of Foreign Affairs responded to this shared concern by including in the discussions military and diplomatic experts, as well as ethicists and AI experts from the civil society, private sector, academia – including some involved in the drafting of the European Commission proposal –, scientists, the ICRC, and UNESCO.

THE URGENT NEED FOR AN INTERNATIONALLY LEGALLY BINDING INSTRUMENT ON AUTONOMOUS WEAPONS SYSTEMS

An international treaty is key to address the various concerns raised by autonomous weapons systems. To be effective, such a treaty should include a prohibition of antipersonnel autonomous weapons and those that could be used without meaningful human control; and positive obligations for other uses of autonomous weapon systems. A legally binding instrument at the international level is at the core of the policy responses that will be necessary to address concerns regarding AI-enabled weapon systems. Codes of conduct, declarations, guiding principles, a compendium of good practices or weapons reviews would not be sufficient responses to autonomous weapons systems, as they would not carry the same weight as a legally binding instrument. Crucially, non-binding instruments would not be able to ensure the transparency, accountability, and responsibility required in the case of autonomous weapons. Simply, states would not have a legal obligation to abide by their rules. Furthermore, such instruments would not create the high international standard that was established, for instance, by the Antipersonnel Mine Ban Convention; and which has led to *de facto* compliance with most treaty obligations by 32 countries that are not yet Party to this treaty (ICBL, 2021).

The fact that autonomous weapons could be in the arsenals of any military or police force is already of concern. But once autonomous weapons systems exist, they are likely to end up in the hands of non-state actors and other illegal recipients due to diversion – weapons that are acquired by, or delivered to, unauthorized users – as is the case with other conventional weapons.[73] Consider for example the case of Afghanistan, where the Taliban now control (i) billions of USD worth of weaponry (Cohen and Liebermann, 2021); and (ii) the biometric data of Afghan staff who worked for US and NATO forces, which is already used to "hunt down Afghans who helped US and allied forces," using US equipment and data to do it (Roy and Miniter, 2021). Such a risk of diversion of weapons to non-authorized or unintended users is a common concern related to the global arms trade.

Failing to adopt an international treaty on autonomous weapons systems would mean that the international community accepts *de facto* that the decisions over human life can be delegated to autonomous systems. This could, in turn, negatively impact other regulations related to AI and emerging technologies, for instance regarding decisions over human lives in healthcare settings. Indeed, if the right to life can be ceded to AI applications, why shouldn't we cede other rights as well?

Discussions on autonomous weapons have already taken seven years at the CCW. Some of the same countries that are investing in autonomous weapons research are the ones that say that more discussion and research is needed. According to Human Rights Watch (2020), Israel, Russia, and the United States are some of the countries investing heavily in the development of various autonomous weapons systems. These same countries have called a legally binding instrument on autonomous weapons "premature,"

73. See for instance: Kirkham, Elizabeth. 2017. *International efforts to prevent diversion of arms and dual-use goods transfers: challenges and priorities.* SaferWorld. https://www.saferworld.org.uk/resources/publications/1112-international-efforts-to-prevent-diversion-of-arms-and-dual-use-goods-transfers-challenges-and-priorities

according to Amnesty International (2021). Such a position is not neutral: if countries participating in the CCW meetings agree to continue discussing without a negotiating mandate for a legally binding instrument, the decision would be benefiting those countries and industries that are already working to develop autonomous weapons.

On a positive note, more voices from the South – including from civil society, scientists, and delegations at the GGE-LAWS – are rising in favor of a treaty on autonomous weapons. As such, the Global South, including countries affected by conflict – which would also be the likely testing grounds for autonomous weapons – are expressing their discontent with the lack of progress on this matter, and presenting clear proposals for the international legal framework that is urgently needed. In disarmament and arms control discussions, countries from the Global South have emerged not only as norm-takers but indeed as norm-makers (Bode, 2019).

Arms control and humanitarian disarmament treaties work. They have a direct impact in preventing death, injuries, human suffering, and long-lasting negative effect in the lives and livelihoods of generations of thousands of people.[74]

Yet, the recent CCW meetings have raised serious questions about the possible multilateral response to addressing these systems in that forum and the role of more advanced militaries. Two crucial meetings were held in December 2021: first, the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems, which was unable to agree on a consensus report with recommendations to continue its work with a more ambitious mandate. Though many delegations called for a mandate to negotiate a legally binding instrument, some countries have blocked progress. At the end of the GGE, the initial proposal of the Belgian Chairman to develop "an instrument" (not even a legally binding one, as a compromise for militarized countries) was further watered down. And even then, these delegations would not accept it. The decision was then pushed to the CCW Review Conference held in mid-December 2021 which takes place every five years. The Review Conference result was a weak mandate which only commits the CCW to continue deliberations for ten days in 2022, and without any specific objective, let alone the negotiations of a legally binding instrument.

The key question is whether there is political will to pursue the issue of autonomous weapons in another forum. This was for example the case with the Antipersonnel Mine Ban Convention and the Convention on Cluster Munitions (Herby, 1998). The uncertainty around political will raises a larger set of questions related the multilateral system ability to respond to challenges regarding the militarization of emerging technologies. Particularly, given that "consensus" has been distorted and led to international forums being hijacked by the militarized few. Given the role of the AI research community and industry in developing some of the technology that will be critical to military applications, it remains to be seen how these communities will respond to the lack of international agreement. Certainly, more attention and pressure for agreements and standards would be helpful in preventing malicious uses. Still, states need to pursue agreements that will have the greatest impact on global security.

74. For more on this, see for instance: United Nations, Office of Disarmament Affairs. *Landmines*. https://www.un.org/disarmament/convarms/landmines/. The Convention on Cluster Munitions. *Achievements*. https://www.clusterconvention.org/

DEEPFAKES

Autonomous weapons developments show how AI is already being weaponized and directly being incorporated in weapon systems. However, AI can also be used in different ways to create instability or contribute to conflict escalation that has not been fully grasped. While autonomous weapons have been called the next major revolution in warfare, deepfakes are increasingly recognized as transformative for the overall security and political environment and touch on issues of global and national security, democracy, gender-based violence as well as privacy rights. Deepfakes are of particular concern given the changing character of warfare and the use of what are seen as grey zone tactics, which are activities that fall below the threshold of conflict. Particularly, manipulated content and disinformation campaigns carried out through cyber means.

So, what are deepfakes? Deepfakes are essentially manipulated or manufactured images and video as well as audio. While image and video manipulation has been happening for many years, the accessibility and speed of the technology have greatly changed the realities of who can create manufactured believable content. It is also important to highlight that deepfakes are distinct from so-called "cheapfakes," that is the slowing down or speeding up of footage to make a particular point (Venema, 2020). Advancements in AI have transformed content manipulation capabilities. Deepfakes involve two aspects of machine learning: neural networks and generative adversarial networks (Pantserev, 2020, pp. 39-40). Neural networks ensure that audio and video content generated is as accurate as possible by downloading a number of examples of content it is trying to synthesize. The generated content is then tested by another neural network that aims to determine whether the content is real or fake, called the "discriminator." If the discriminator determines the video is fake, the generator then tries to learn how it was detected and improve on those aspects (Pantserev, 2020, pp. 39-40). As such, the quality and accuracy of the deepfake is constantly improved upon, making it harder to detect with each repetition of this process.

According to the startup Deeptrace, the number of deepfakes on the web increased 330% from October 2019 to June 2020 (Wiggers, 2021). Yet, most governments and businesses are unprepared for the potential widespread impacts of deepfakes on society. As the technology continues to improve, the public is finding it harder to discern real content from manufactured content (O'Brien, 2019). This is leading to greater uncertainty about the veracity of information and is contributing to the undermining of trust and online civic culture, including in democratic societies. Interestingly, the technological possibility is also leading to questioning of real content, with deepfakes being used as a possibility to dismiss actual statements. Critically, deepfakes have tended to target women, with some 90% of deepfakes to date being pornographic content, highlighting the concern that the technology is being used to deploy gender-based violence (Venema, 2020). Additionally, manipulated videos of political leaders have raised concerns about potential for conflict escalation or misunderstandings that could have a widespread impact in conflict-affected societies. Of particular concern for contemporary military engagements is also the centrality of the information and communication technologies in controlling narratives and responses and hence, the impact that a deepfake could potentially have on escalating a crisis. As such, the gap in legislation and policy responses concerning deepfakes requires greater attention from technologists, governments, industry, international organizations, and civil society.

What are the concerns with deepfakes?

Deepfakes are specifically altered using machine learning and deep learning and have become more convincing. As Adee (2020) notes, it used to take Hollywood studios a year with a team of experts to essentially "stich in" an actor into a video or image that they have not been in. Now, deepfake technologies allow for that type of insertion of individuals into images or scenarios much more quickly. There is also a wider proliferation of more basic synthetic creation. For example, smartphone applications such as Avatarify or Zao App allow the user to animate faces or swap faces of individuals in videos and images (Fowler, 2021; Meskys et al., 2021). However, more sophisticated technological capability is also available and individuals who are not technical experts can now create convincing deepfake content from their homes. This a result of the fact that some code is open source and fewer images are needed to create good quality fakes. As systems "learn" from more images, the quality will also improve, raising the issue of how manufactured content will be identified. As Kietzmann et al. (2020) argue, the believability of fake content, as well as our tendency to trust photographic and particularly audio and video evidence, makes it challenging to respond to manipulated content. As some uses of deepfake technology focus on entertainment and the creation of fun content, it can lead to a misperception regarding the potential misuses of the technology. However, there have already been instances of malicious uses and images of individuals, including celebrities but also ordinary individuals, being featured in particularly problematic content without their consent or knowledge. Women, in particular, have been disproportionately impacted with deepfake technology being used to create nonconsensual fake pornography. As such, gendered aspects of deepfake technology need to be addressed by governments. As Venema (2020) points out, women in different parts of the world have been impacted by uses of deepfake pornography leading them to lose their jobs or not being able to find employment. Venema also highlights the case of Indian journalist, Rana Ayyub, who was targeted with a deepfake slander campaign when she advocated for justice in the rape of a young girl. In Ayyub's case, she also experienced "doxing," the release of her personal contact information. Ayyub's case is one of several showing how the deepfake content translates into real-world security concerns for women. Given the differing levels of gender protections in place in various countries, the spread of the technology could result in dire consequences for women whose image is maliciously used. Such context and country-specific impacts need to be carefully considered by tech experts as well as platforms through which the content is shared.

As the technology proliferates there is also a growing recognition about the potential use of deepfakes in disinformation campaigns and in possibly escalating or creating conflicts. In 2018, a video of former United States President Barack Obama published by BuzzFeed and shared on social media contributed to this growing consciousness regarding the use of deepfakes in political campaigns. In the video, Obama makes a quip about then US President Donald Trump that is very out of character for him. Indeed, looking at the camera, Obama says so himself. He goes on to say that it is someone like Jordan Peele, a movie director, who would make those remarks about then President Trump. A split screen then shows the viewer that it is Peele speaking and that his team used AI to make it appear as if it was Obama speaking. The video created by Peele and his team was supposed to be a public service announcement warning about the dangers of deepfake technology, but it left some viewers uneasy. The video was simply too good. Obama's image and videos are readily available, and it did take the machine learning system some 56 hours of training to get it just right (Romano, 2018). Still, the Obama video brought to the forefront concerns about potential fake content that could be used to spread disinformation against political leaders or even manufacture crisis and escalate conflicts.

In already conflict-affected areas, it does not take a stretch of the imagination to realize the potential dangers of circulation of a video or audio of a political leader inciting violence or voicing threats against other communities (Citron and Chesney, 2019). Social media spread of misinformation, for example, in conflict areas has shown how violence called for in the virtual world can lead to violence in the real world. The use of Facebook in inciting violence in Myanmar has received a great deal of attention over

the years (Asher, 2021). Facebook posts were used to target the Rohingya minority community for years including in the 2018 crisis which saw the displacement of some 800,000 Rohingya. In 2014, a viral post targeting the Muslim community in Myanmar resulted in two deaths after a violent mob response to a Facebook post. The case illustrates how manipulated videos and images in fragile contexts can be used to target specific groups and communities.

While many hypothetical scenarios can be thought of, for some countries, these are no longer mere possibilities. Critically, it is in fragile political contexts and in developing countries with lower levels of digital literacy that deepfakes pose potential dangers. Consider the case of Gabon, that in 2019 faced a military coup, prompted by what is believed to be a video of the ailing president (Breland, 2019). Gabon's president had been receiving medical attention out of the country, and as demands for his appearance were growing, the government released a video message. However, the video message appeared to show that the concerns about the president were indeed accurate as the video was described as odd. The coup was ultimately unsuccessful, but the video was deemed to be a deepfake and highlighted that the real dangers of technology will be most felt in fragile contexts and those where the population have been excluded or had not had access to digital literacy. Still, while deepfakes may have disproportionate impacts on marginalized communities, the broader international community is impacted as well. The use of deepfakes could have wide-ranging implications for international stability. Citron and Chesney (2019) offer several examples that have implications for national and global security as well as for diplomatic relations. They note how a "False audio might convincingly depict U.S. officials privately 'admitting' a plan to commit an outrage overseas, timed to disrupt an important diplomatic initiative" (Citron and Chesney, 2019, p. 176). Not all deepfakes would have a destructive aim, but some could be timed to impact the outcomes of diplomatic summits and to constrain the possible negotiations between different countries.

Towards a regulatory response for deepfakes

A number of technological countermeasures in relation to deepfakes have been proposed along with regulatory responses. In terms of regulatory responses, there is an ongoing debate regarding whether deepfake technology should be banned. Some sites, such as Reddit, have banned deepfake pornography and Facebook has also banned deepfakes. Specifically, Facebook policy prohibits videos that are "edited or synthesized" by AI and that the users are not aware have been manipulated by these technologies. Facebook continued to allow content that is for parody or satire. One of the key sticking points in the regulatory discussion is finding a way to prohibit certain uses while allowing for freedom of expression, artistic creations and entertaining content that individuals know are deepfakes.

It is clear that deepfake technology requires a regulatory as well as normative response. Prohibitions on some uses of technology need to become normalized among countries, technologists and civil society. For instance, deepfake non-consensual pornography and any deepfakes that are created without the consent of the individual involved should be banned. Privacy protections in various national and regional jurisdictions need to be enacted. It will be critical to ensure that jurisdictional gaps do not exist, and a wider global agreement is realized that would lead to the adoption of the norm.

Disclaimers and notifications that the content deemed acceptable is manipulated are necessary. However, any disclaimers or notifications need to be clearly communicated and visible to the consumer. As such, an obscured message or simple popup that could be easily ignored is not sufficient. Thought must be given to the design of systems to ensure that users are completely aware of the deepfake and that it is overlaid in the media (Olejnik, 2021).

What next? The proliferation of deepfakes is still in its early stages. Without timely and appropriate regulatory responses, the dangers of the technology which have already been demonstrated in real-world impacts will only continue to grow. Unfortunately, deepfakes tend to disproportionately impact already vulnerable and disenfranchised groups with gendered implications and serious concerns

in fragile political contexts and developing countries. However, the reality is also that no country is immune from the impacts of deepfakes. The concerns about disinformation are prominent in liberal democracies but they are real for all countries. Technical experts have also proposed solutions, such as using AI to detect and eliminate manipulated content, and states could consider a number of regulatory responses, such as requiring the labelling of manipulated content and removal of malicious content (Pantserev, 2020). As a starting point, clarity on the malicious uses of deepfakes and concerns regarding the technology deserve greater attention from all countries. Leading tech companies have a role to play in addressing removal of content in a timely manner and clear legislation is needed on their role and requirements placed on these companies. International discussions on developing norms regarding responsible state behavior in cyberspace could also offer a possible venue to place limits on certain uses of deepfakes, such as electoral interference. The continued use of grey zone tactics and notably of disinformation campaigns also mean that malicious uses of deepfakes could have very real security implications.

CONCLUSION

Autonomous weapons systems and deepfakes point to areas in which AI could be weaponized, raising concerns among stakeholders in and out of government. There are clear ethical, legal, human rights, and humanitarian concerns regarding the development and potential use of increasingly autonomous weapon systems and deepfakes. For instance, potential hacks, accidents and misperceptions that functioning without meaningful human control could bring. The establishment of norms to tackle their development and proliferation could lead to increased international stability. The window of opportunity to respond to this concern is still open, but not for long. Increasing competition among advanced militaries also means that AI systems that are not ready could be rushed and deployed – and, as history demonstrated, countries already in conflict and those in the Global South would probably be the first affected by such weaponization.

The lowering of barriers to access these technologies means that the stakeholders that need to be engaged are much wider than on many previous issues in the realm of global and national security. The wider awareness that is needed among the broader public about the impacts of these technologies also includes an urgent need to build digital literacy. Digital literacy, international legislation based on human rights and humanitarian law, and an informed public are key to addressing the very real threats posed by these technologies. Multilateral diplomatic engagement is necessary to make certain that AI and emerging technologies remain a tool to advance the social good. Ultimately, it is governments that need to and can develop policies to protect their citizens and ensure international stability.

REFERENCES

Abrams, A. 2020. *Black, Disabled and at Risk: The Overlooked Problem of Police Violence Against Americans with Disabilities*. Time. https://time.com/5857438/police-violence-black-disabled/

Adee, S. 2020. What Are Deepfakes and How Are They Created? IEEE Spectrum. https://spectrum.ieee.org/what-is-deepfake

African Commission on Human and Peoples' Rights. 2021. *Resolution on the need to undertake a Study on human and peoples' rights and artificial intelligence (AI), robotics and other new and emerging technologies in Africa* – ACHPR/Res. 473 (EXT.OS/XXXI) 2021. Available in English and French at https://www.achpr.org/sessions/resolutions?id=504

Alliance for Multilateralism. 2019. *Declaration by the Alliance for Multilateralism on Lethal Autonomous Weapons Systems (LAWS)*. https://multilateralism.org/wp-content/uploads/2020/04/declaration-on-lethal-autonomous-weapons-systems-laws.pdf

Allyn, B. 2020. *IBM Abandons Facial Recognition Products, Condemns Racially Biased* Surveillance.NPR. https://www.npr.org/2020/06/09/873298837/ibm-abandons-facial-recognition-products-condemns-racially-biased-surveillance

Altmann, J. 2009. Preventive Arms Control for Uninhabited Military Vehicles. In Capurro, R. and Nagenborg, M. (eds.). *Ethics and Robotics*. Heidelberg: AKA Verlag.

Amnesty International. 2021. *A critical opportunity to ban killer robots while we still can*. https://www.amnesty.org/en/latest/news/2021/11/global-a-critical-opportunity-to-ban-killer-robots-while-we-still-can/

Asaro, P. 2012. On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision-making. Special Issue on New Technologies and Warfare. *International Review of the Red Cross*, Vol. 94, No. 886, pp. 687-709 (Braille translation by Don Winiecki, French Translation). https://www.cambridge.org/core/journals/international-review-of-the-red-cross/article/on-banning-autonomous-weapon-systems-human-rights-automation-and-the-dehumanization-of-lethal-decisionmaking/992565190BF2912AFC5AC0657AFECF07

Asher, S. 2021. *Myanmar coup: How Facebook became the 'digital tea shop'*. BBC News. https://www.bbc.com/news/world-asia-55929654

Bengio, Y. 2019. *AI pioneer: The dangers of abuse are very real.* Nature. https://www.nature.com/articles/d41586-019-00505-2

Bode, I. 2019. Norm-making and the Global South: Attempts to Regulate Lethal Autonomous Weapons Systems. *Global Policy,* No. 10, No. 3, pp. 359-364.

Breland, A. 2019. *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink*. Mother Jones. https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

Bugge, A. 2018. *U.N.'s Guterres urges ban on autonomous weapons.* Reuters. https://www.reuters.com/article/us-portugal-websummit-un-idUSKCN1NA2HG

Burton, J. and Soare, S. R. 2019. *Understanding the Strategic Implications of the Weaponization of Artificial Intelligence*. 11th International Conference on Cyber Conflict: Silent Battle T. Minárik, S. Alatalu, S. Biondi, M. Signoretti, I. Tolga, G. Visky (eds.) NATO CCD COE Publications, Tallinn. https://www.ccdcoe.org/uploads/2019/06/Art_14_Understanding-the-Strategic-Implications.pdf

Buolamwini, J., Gebru, T. 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.* MIT. https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/

Campaign to Stop Killer Robots. 2021. *Clear momentum towards a new legal framework on autonomous weapons*. https://www.stopkillerrobots.org/2021/08/clear-momentum-towards-a-new-legal-framework-on-autonomous-weapons/

CCW. 2019. *Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*. CCW/MSP/2019/9 (see Annex 3 for the Guiding Principles.) https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/343/64/PDF/G1934364.pdf?OpenElement

Citron, D. K., Chesney, R. 2019. *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. California Law Review (107:1753). https://scholarship.law.bu.edu/faculty_scholarship/640

Cohen, Z. and Liebermann, O. 2021. *Rifles, Humvees and millions of rounds of ammo: Taliban celebrate their new American arsenal*. CNN. https://edition.cnn.com/2021/08/21/politics/us-weapons-arsenal-taliban-afghanistan/index.html

Dastin, J. 2021. *Amazon extends moratorium on police use of facial recognition software.* Reuters. https://www.reuters.com/technology/exclusive-amazon-extends-moratorium-police-use-facial-recognition-software-2021-05-18/

Devoto, M., Janssen, E. and Muñoz, W. 2021. A*nálisis de las propuestas y declaraciones de países del Sur Global sobre el marco normativo y operativo en el área de Sistemas de Armas Autónomas.* SEHLAC. https://bit.ly/2WD8ELQ

Díaz, M. and Muñoz, W. 2019. *Los riesgos de las armas autónomas: una perspectiva interseccional latinoamericana.* SEHLAC. https://bit.ly/ArmasInterseccionalidad

Dreifus, C. 2019. *Toby Walsh, A.I. Expert, Is Racing to Stop the Killer Robots*. The New York Times. https://www.nytimes.com/2019/07/30/science/autonomous-weapons-artificial-intelligence.html

European Commission. 2021. *Proposal for a regulation of the European Parliament and of the Council Laying Down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

European Parliament. 2018. *European Parliament resolution of 12 September 2018 on autonomous weapon systems.* 2018/2752 (RSP). https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_EN.html

Fowler, G. A. 2021. *Anyone with an iPhone can now make deepfakes. We aren't ready for what happens next*. Washington Post. https://www.washingtonpost.com/technology/2021/03/25/deepfake-video-apps/

Future of Life Institute. 2015. *Autonomous weapons: an open letter from AI & robotics researchers.* https://futureoflife.org/open-letter-autonomous-weapons/

Gould, L. 2021. *Remote Warfare: Interdisciplinary perspectives*. Safer World (podcast episode). https://www.saferworld.org.uk/multimedia/saferworldas-warpod-episode-8-remote-warfare-interdisciplinary-perspectives

GPAI. 2021. https://gpai.ai/projects/

Greene, J. 2020. *Microsoft won't sell its facial-recognition technology, following similar moves by Amazon and IBM*. The Washington Post. https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/

G20 Trade Ministers and Digital Economy Ministers. 2019. *G20 Ministerial Statement on Trade and Digital Economy*. https://www.mofa.go.jp/files/000486596.pdf

Herby, M. 1998. *An international ban on anti-personnel mines: History and negotiation of the "Ottawa treaty"*. ICRC. https://www.icrc.org/en/doc/resources/documents/article/other/57jpjn.htm

Horowitz, M. C. 2020. *AI and the Diffusion of Global Power.* Centre for International Governance Innovation. https://www.cigionline.org/articles/ai-and-diffusion-global-power/

Human Rights Watch. 2012. *Losing Humanity: The Case against Killer Robots.* https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

––––. 2020. *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control*. https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and

––––. 2021. *Areas of Alignment. Common Visions for a Killer Robots Treaty.* https://www.hrw.org/news/2021/08/02/areas-alignment

ICRC. 2021. *ICRC position on autonomous weapon systems*.

––––. No date. *Fundamental Principles of IHL.* https://casebook.icrc.org/glossary/fundamental-principles-ihl

ICBL. 2021. *Treaty Status*. http://www.icbl.org/en-gb/the-treaty/treaty-status.aspx

IKV Pax Christi. 2011. *Does Unmanned Make Unacceptable? Exploring the Debate on using Drones and Robots in Warfare.* https://paxvoorvrede.nl/media/download/does-u-make-ulowspreads_0.pdf

Keller, J. 2021. *Pentagon to spend $874 million on artificial intelligence (AI) and machine learning technologies next year*. Military Aerospace. https://www.militaryaerospace.com/computers/article/14204595/artificial-intelligence-ai-dod-budget-machine-learning

Kietzmann, J., Lee, L. W., McCarthy, I. P., Kietzmann, T. C. 2020. Deepfakes: Trick or treat? *Business Horizons*, Vol. 63, No. 2, pp. 135-146, ISSN 0007-6813. https://doi.org/10.1016/j.bushor.2019.11.006.

Knowles, E. and Watson, A. 2021. *Remote Warfare: Lessons Learned from Contemporary Theatres*. SaferWorld. https://www.saferworld.org.uk/resources/publications/1280-remote-warfare-lessons-learned-from-contemporary-theatres

Koettl, C., Hill, E., Aikins, M., Schmitt, E., Tiefenthäler, A. and Jordan, D. 2021. *How a U.S. Drone Strike Killed the Wrong Person* (video). The New York Times, September 10. https://www.nytimes.com/video/world/asia/100000007963596/us-drone-attack-kabul-investigation.html.

Marijan, B. 2018. Human-less or human more? *The Ploughshares Monitor*, Vol. 39, No. 2, pp. 5-7.

Meskys, E., Liaudanskas, A., Kalpokiene, J. and Jurcys, P. 2021. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law & Practice,* Vol. 15, No. 01. pp. 24-31. DOI:10.1093/jiplp/jpz167.

Montreal Declaration for a Responsible Development of AI. 2017. https://www.montrealdeclaration-responsibleai.com/the-declaration

Moyes, R. 2012. *Autonomous weapons – the risks of a management by 'partition.'* Article 36. https://article36.org/updates/autonomous-weapons-the-risks-of-a-management-by-partition/

––––. 2019. *Target profiles*. Article 36. http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf

Muñoz, W. 2021a. *Autonomous weapons systems: an analysis from human rights, humanitarian and ethical artificial intelligence perspectives*. https://bit.ly/SEHLACAI-AWS

––––. 2021b. *It's about more than autonomous weapons systems*. Armed Conflict and Civilian Protection Initiative of the International Human Rights Clinic, Harvard Law School. https://humanitariandisarmament.org/2021/08/30/it-is-about-more-than-autonomous-weapons-systems/

Nobel Women's Initiative. 2014. *Nobel Peace Laureates call for Preemptive Ban on "Killer Robots".* https://nobelwomensinitiative.org/nobel-peace-laureates-call-for-preemptive-ban-on-killer-robots/

O'Brien, M. 2019. *Why 'deepfake' videos are becoming more difficult to detect*. PBS NewsHour. https://www.pbs.org/newshour/show/why-deepfake-videos-are-becoming-more-difficult-to-detect

OECD.AI Policy Observatory. n.d. *OECD AI principles overview*. https://www.oecd.ai/dashboards/ai-principles/P6

OECD. 2019. Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Olejnik, L. 2021. *TechLetters #23 – Deepfakes OK? Vulnerable IoTs. SolarWind hacks in Europe. Cyber sanctions, Russia twice. Hacked cheese*. https://techletters.substack.com/p/techletters-23-deepfakes-ok-vulnerable?s=r

Palmer, C. n.d. *Ethical Theory and Philosophical Method Feminist Ethic*s. Lancaster University. https://www.lancaster.ac.uk/users/philosophy/awaymave/401/feminist.htm

Pantserev, K. A. 2020. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. In Jahankhani, H. et al. *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, pp. 37-55. Springer.

Pax Netherlands. 2019. *Killer robots: what are they and what are the concerns.* https://paxforpeace.nl/media/download/pax-booklet-killer-robots-what-are-they-and-what-are-the-concerns.pdf

Ramsay-Jones, H. 2019. *Racism and Fully Autonomous Weapons*. https://www.ohchr.org/Documents/Issues/Racism/SR/Call/campaigntostopkillerrobots.pdf

Reaching Critical Will. 2021. *CCW Report*. https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/gge/reports/CCWR9.4.pdf

Republic of Korea. 2020. *ROK and UNESCO co-organize Virtual Asia-Pacific Consultation on UNESCO Recommendation on the Ethics of Artificial Intelligence.* Ministry of Foreign Affairs. https://www.mofa.go.kr/eng/brd/m_5676/view.do?seq=321173&srchFr=&amp%3BsrchTo=&amp%3BsrchWord=&amp%3BsrchTp=&amp%3Bmulti_itm_seq=0&amp%3Bitm_seq_1=0&amp%3Bitm_seq_2=0&amp%3Bcompany_cd=&amp%3Bcompany_nm=

Roy, S. and Miniter, R. 2021. *Taliban kill squad hunting down Afghans — using US biometric data*. New York Post. https://nypost.com/2021/08/27/taliban-kill-squad-hunting-afghans-with-americas-biometric-data/

Russel, S. 2021. *It's time to ban autonomous killer robots before they become a threat.* Financial Times. https://www.ft.com/content/04a07148-d963-4886-83f6-fcaf4889172f

Sauer, F. 2021. Stepping back from the brink: Why multilateral regulation of autonomy in weapons systems is difficult, yet imperative and feasible. ICRC. https://international-review.icrc.org/articles/stepping-back-from-brink-regulation-of-autonomous-weapons-systems-913

Sharkey, A. 2019. Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology*. Vol. 21, pp. 75–87.

Sharkey, N. 2007. *Robot Wars are a Reality*. The Guardian. https://www.theguardian.com/commentisfree/2007/aug/18/comment.military

UNESCO. 2019. *Charter of Ethics of Science and Technology in the Arab Region*. http://www.umi.ac.ma/wp-content/uploads/2020/01/Charte-Ethique-des-Sciences-et-Technologies-UNESCO-V.Eng_.pdf

––––. 2020. *Outcome document: first draft of the Recommendation on the Ethics of Artificial Intelligence*. SHS/BIO/AHEG-AI/2020/4 REV.2. https://unesdoc.unesco.org/ark:/48223/pf0000373434

––––. 2021. *Elaboration of a Recommendation on the ethics of artificial intelligence*. https://en.unesco.org/artificial-intelligence/ethics

UN Human Rights Council. 2013. Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions. A/HRC/23/47. https://www.refworld.org/docid/51a747c54.html

UNIDIR. 2020. *Modernizing Arms Control.* https://unidir.org/publication/modernizing-arms-control

––––. 2021. *The 2021 Innovations Dialogue: Deepfakes, Trust And International Security*. https://unidir.org/events/2021-innovations-dialogue

Venema, A. E. 2020. Deepfakes as a Security Issue: Why Gender Matters. WIIS Global. https://wiisglobal.org/deepfakes-as-a-security-issue-why-gender-matters/

Wallach, W. 2013. *Terminating the Terminator*. Science Progress.

Wiggers, K. 2021. *Fewer than 30% of business have a plan to combat deepfakes, survey finds*. VentureBeat. https://venturebeat.com/2021/05/24/less-than-30-of-business-have-a-plan-to-combat-deepfakes-survey-finds/

ETHICS OF CARE AND ARTIFICIAL INTELLIGENCE: THE NEED TO INTEGRATE A FEMINIST NORMATIVE APPROACH

PAULINE NOISEAU

AI Ethics Researcher at Algora Lab and at Université de Montréal.

SDG5 - Gender Equality
SDG10 - Reduced Inequalities
SDG11 - Sustainable Cities and Communities
SDG16 - Peace, Justice and Strong Institutions
SDG17 - Partnerships for the Goals

ETHICS OF CARE AND ARTIFICIAL INTELLIGENCE: THE NEED TO INTEGRATE A FEMINIST NORMATIVE APPROACH

ABSTRACT

In recent years, we've witnessed an anthology of instruments relating to artificial intelligence (AI): charters, declarations, a set of ethics principles, etc. These structures are directly inspired by dominant moral philosophies in a similar way to deontologism or consequentialist ethics. The ethical dimension of certain uses of AI is determined by these moral philosophies. However, their premises and representations are rarely called into question. The aim of this article is to disrupt what we call "the ethical saturation of AI" by turning to a truly alternative theory. Therefore, this chapter will explore how the ethics of care can be an answer to the saturation of the dominant theories. The ethics of care invites us to change our outlook and to adopt new critical perspectives on AI. It marks a clear paradigm break by establishing the care given to others as a criterion for morality of action and by acknowledging the interdependence of living things and management of vulnerability as inherent characteristics of the human species (Brugère, 2011). An important characteristic of the care ethics is its feminist nature: it places care at the heart of politics and collective action as the foundation of life. Care activities are traditionally and primarily carried out, for free or very little recognition, by women and people historically marginalized in our society (Gilligan, 2011). This chapter raises the following questions: How can we explain that the ethics of care constitutes a true blind spot in reflections on the ethics of AI? Is AI ethical when viewed through the lens of care ethics? This proposal is important, because it enables us to broaden our ethical and critical perspectives on AI. Furthermore, it provides us with a new perspective to make public policies in the area of AI fairer and more inclusive.

INTRODUCTION

> To understand what is at stake, we must focus less on ethics and more on power. AI is invariably designed to amplify and reproduce the forms of power it has been deployed to optimize. Countering that requires centering the interests of the communities most affected. Instead of glorifying company founders, venture capitalists, and technical visionaries, we should begin with the lived experiences of those who are disempowered, discriminated against and harmed by AI systems.
>
> Kate Crawford (2021, pp. 224-225).

> Mais une éthique du care, avec les exigences morales d'attention et de responsabilité qui l'accompagnent, pourrait permettre de dévoiler la manière dont les puissants tentent de fausser la compréhension des besoins pour maintenir leurs privilèges et leurs positions de pouvoir.
> [But an ethics of care, along with its accompanying moral demands for care and responsibility, may help to reveal how the powerful attempt to distort the understanding of needs to maintain their privileges and positions of power.]
>
> Joan Tronto (2009, p. 198).

We can define artificial intelligence (AI) as the collection of computer systems that enable us to simulate and reproduce certain functions of human intelligence such as memorization, learning or calculation (Boden, 1990). Although AI has existed for a long time, it has been talked about as a particularly striking technique since work began on deep learning (Goodfellow et al., 2016). This latest technical feat makes it possible to use a neural network—directly inspired by the human brain—to achieve levels of complex abilities at unprecedented speed and accuracy. Considering those characteristics, the uses of AI are recognized for facilitating certain actions performed by human beings. It can be used to carry out certain tasks that can free up time and energy for the human being, making them more attainable. In that sense, it can be argued that AI increases the power of human beings, understood here as the ability to do and act. Therefore, the benefits AI brings to human beings are primarily related to operationalization and efficiency, i.e. making a system more productive and consequently gaining both quantitatively and qualitatively.

On the other hand, some uses of AI have been identified as costly on different levels, namely social, political or environmental. Specific examples include the carbon footprint associated with maintaining algorithms, algorithmic biases, big data from surveillance, as well as the violation of human emancipation through nudging and choice orientation processes. This is when ethics emerged as a discipline helping to determine the moral value of certain uses in order to limit or control their development.

The ethics of AI can be defined as the field that "(...) attempts to reflect, identify and propose a use for AI that is in agreement with a common way of being, i.e. a set of values and principles that are specific to a society" (Noiseau et al., 2021). Despite the involvement of ethics to responsibly guide and direct the development of AI, it's clear that these normative proposals have not succeeded in ensuring that the development of AI is aligned with the collective and social issues that we are facing today. There are many similarities in the analyses proposed by the dominant theories of ethics applied to AI, at both the formal and conceptual levels. At the formal level, these ethical analyses take the form of a set of ethical recommendations, principles or values to be applied directly to use cases or existing practices. At the conceptual level, the contents largely converge, highlighting the same issues and the same values (Jobin et al., 2019). This normative homogenization can be explained in two ways.

First, it illustrates the domination of certain moral theories over others, conveying a similar way of perceiving and understanding the world. Second, it only exists because the field of AI ethics does not use a sufficiently alternative moral theory to disrupt the philosophical premises, the form of moral reasoning and the ethical conclusions that follow. Historically, the ethics of AI seem to have stopped at aging structures, directly inspired by modernity, which originate from a liberal conception of the human being,

in other words, rational, autonomous and independent. However, this chapter argues that the reason that the ethics of AI struggles to target or identify the negative ramifications of certain uses, or even to question the relevance and, therefore, the existence of this technology in our global life ecosystem, is because it is failing to consider an essential human activity, namely care.

The activities covered under the umbrella term of *care* encompass a large portion of the practices that enable us to maintain life. Indeed, in her book published under the original title *Moral Boundaries. A Political Argument for an Ethics of Care*, Tronto (1993) distinguishes four phases of care: caring about; taking care of; care giving; care receiving (Tronto, 2009, p. 147). The different phases are described as follows.

The first step is to be able to recognize the existence of a need. Therefore, the first provision and practice of care relates to attentiveness, i.e. the attitude of consciously engaging with the other, of perceiving explicit or non-explicit signs of a need. For example, recognizing a person's need to talk about a traumatic event or to be taken care of by a health professional. Inversely, not paying attention to the other person—not caring—leads to a failure of care because the first step will have been missed.

After becoming aware of and recognizing the other person's need, the second phase is to take care of this need. Taking care may result in the provision of a specific material structure ensuring that the need is met. The care provider must assume the responsibility and determine how to respond to the current situation. For example, a father takes care of his family by working and providing the material resources necessary for the family's survival.

The third phase consists of care giving. In other words, providing a concrete and direct response to the need. Here, we must distinguish between taking care of and care giving. Although the doctor takes care of the patient by administering a treatment, it is the nurses who provide care to the patient by concretely carrying out the treatment and establishing the relationship. Although a father takes care of the family's needs by bringing home a salary, he is not necessarily providing care, because this involves preparing the meals, washing bodies, listening to the children.

Finally, the fourth phase of care consists of care receiving. However, the care is void if it is not recognized as complete by the beneficiary. In other words, it is a question of verifying whether the care provided corresponds to the initial need. The last phase makes it possible to ensure the care giver's accountability. An institution that claimed to provide care to a segment of the population without validating that the care provided meets the identified need in terms of its reception and adequacy would have failed in its caring work.

These four phases enable us to establish what care is and what it is not. In sum, caring is both an ability and it is work (Tronto, 2009, p. 145). Stating that it is work allows us to shed the aesthetics of care as a vocation and its assumed sentimental aspect (Paperman, 2021). Consider the discourse stating that women or immigrants are by nature more inclined to take care of others. On the contrary, to say that care is work enables us to show how those with privilege have discharged their duty of care towards the other members of society by devaluing the activity (Gilligan et al., 2013). Certain life situations are simply confounding: How is it possible nowadays to work more than 50 hours a week for a company and have a family of three children, while being responsible for the lives of others? We naturally ask ourselves the following questions: Who takes care of the children? The home? Who launders the sheets? Who does the housecleaning? Who buys the groceries? Who prepares the meals? In short, who ensures that our life is a life?

We can say that care activities are both everywhere and nowhere. Everywhere, because these are the actions that keep us alive and support the social and economic system (health, education, sanitation, basic needs sectors); nowhere, because care activities are devalued and passed on to certain segments of the population who carry out this work in the shadows while others continue to enjoy the privilege of not caring about others[75].

Since the 1980s, these care practices have been the subject of significant research in various fields, such as the social sciences and humanities. What is now called the ethics of care emerged after the publication of the book *In a Different Voice* by Carol Gilligan (1982) that became famous for its revolutionary scope. The author questions the hierarchy of moral development proposed by psychologists of the day, including Lawrence Kohlberg, for whom the pinnacle of moral maturity is equivalent to a human being's capacity to actively formulate abstract and universal principles of justice. Despite the dominant understanding that moral reasoning should be understood in terms of rights and responsibilities, in the light of a conception of justice outside of one's own living environment, Gilligan (1982) goes far beyond this ethic by distinguishing another form of moral reasoning. This other form of ethical thinking would stem from specific experience, believed to be associated with gender. In this sense, she shows that women do not understand, reflect and respond in the same way as men when it comes to acting for good when faced with a moral dilemma. According to Gilligan, women understand moral behaviour in terms of responsibility to others in order to uphold an ecosystem of relationships. This other form of moral reasoning, which she calls the ethics of care, is not prompted by adherence to abstract and universal rules; rather, it is applied to guarantee an interdependent long-term network of relationships. Therefore, from a care perspective, indifference towards relationships and the needs of others would be a deficiency, and not the symbol of moral maturity.

The ethics of care makes it possible to add a moral dimension to something that was previously ignored, i.e. ordinary practices of care, carried out in silence and amid general indifference, because they are confined to the private, emotional and relational spheres of life (Tronto, 2009). By establishing the maintenance of an interdependent and contextualized ecosystem of relationships to preserve life and its qualities as a criterion for morality of action, the ethics of care radically disrupts moral theories. Thanks to the ethics of care, what was regarded as the highest form of moral reflection, i.e. the coincidence with principles of justice distanced from the material and relational structures of life, suddenly becomes one form of moral reasoning among many. The ethics of care swerves the ethics of justice and its foundations by valuing a particularist, relational and contextual approach to moral action. The ethical person is not the one who acts in accordance with disembodied rules, but rather the one who utilizes care practices to concretely and clearly maintain a living ecosystem.

Although care relates naturally to the practices carried out by loving and caring relatives, it extends to other spheres, as defined by Tronto and Fisher (2009, [1991], p. 40):

> At the most general level, we suggest that 'caring' be considered a generic activity that includes all that we do to maintain, perpetuate and repair our 'world', so that we can live in it as well as possible. This world includes our bodies, ourselves and our environment, all of which we seek to connect in a complex network, in support of life.

75. The COVID-19 pandemic has revealed a great deal about the unequal responsibility of care within our society. Widespread lockdown demonstrated that certain workers are integral to the operation of society, namely health and education workers, cashiers, cleaners, food sector workers, etc. See also the work of Silvia Federici, specifically her well-known book entitled *Le capitalisme patriarcal*, published in 2019 by La fabrique éditions.

    In her book, the author shows how the capitalist economic system was built on the free work done by women in the private sphere, namely through what she calls the invention of the housewife (see p. 125).

It should be pointed out that what may resemble a form of moral essentialism is not the case. Indeed, it is not a question of stating that there is such a thing as women's morality, but rather, that it emanates from political and material conditions specific to a subject (Tronto, 2009). If women have been led to use moral reasoning based on care, it is because they have historically been predisposed to carry out this work of caring for the needs of others within the family.

The ethics of care is part of a feminist perspective because it goes beyond a moral duality and its underlying moral hierarchy (Gilligan, 2010). It is also about recognizing the ethical dimension of the activities carried out, in our society, by women and historically marginalized people. What Joan Tronto calls "le pouvoir des pauvres" [the power of the weak] is essential to the smooth running of institutions and ramping up of operations because without the little hands that take care of us, the whole system would collapse. Care is therefore a brilliant, critical and revolutionary, tool that enables us to set the record straight by identifying a society's real needs, on the one hand, and on the other to see how the powerful and the privileged relieve themselves of the burden of care so as not to be responsible for the care of others (Hamrouni, 2015; Tronto, 2009).

How, then, can we explain the near absence of care ethics-driven critical perspectives on AI[76]? What would we *lose or gain* by questioning certain uses of AI from a care perspective? If we lived in an ethical structure of care, would AI still find its *raison d'être*? In other words, what would an ethical analysis of AI from a care perspective look like? Could AI be fair if we adopted a common way of being based on a conception of naturally vulnerable human beings engaged in an interdependent network of relationships? The following sections will analyze AI-related ethical and social issues using a feminist normative approach that recognizes care as being at the heart of our life as human beings. First, we will see that the ontology underlying the development of AI relates to a liberal conception of autonomy and of the human being. Second, we will propose a way out of what we will call *ethical saturation in AI* by attempting to change perspectives and looking beyond current considerations on the issue. Finally, we will show that the risk-based approach to AI should be left behind to espouse an approach based on responsibility and attention in a world in crisis[77].

AGAINST HUMAN ENHANCEMENT: VULNERABILITY AND INTERDEPENDENCE

One of the first things to consider when it comes to adopting an ethical and critical perspective on any subject is how we define or identify the conception of humanity or life that is supported by this moral theory. All moral philosophy stems from a set of considerations about the human species. We cannot propose a framework for a subject without first defining it. To build a theory on quality of life, a form of life or a way of life, we must begin by defining what we mean by humanity. To speak of ethics, we must hold a conception of humanity, which serves as the initial premise for the reflection or moral reasoning that follows. These representations must be subjected to criticism. Whether it is Aristotle's virtue ethics, Kant's deontological ethics, or consequentialist ethics, each of these stems from a definition of the human being, that we must examine.

76. However, consider the work of Vanessa Nurock. Many thanks to her for our brief discussions, which were very useful when writing this article. Note also the analysis of UNESCO's normative instrument proposed by the Algora Lab—University of Montreal research laboratory and Mila—Institut Québecois d'intelligence artificielle. In this report, we highlighted the lack of an ethical concept of care in the preliminary version of the Recommendation. See point 2.5 on page 18 of the report (Algora Lab, 2020).

77. When we use the terms world in crisis, we are explicitly referring to the climate crisis (IPCC, 2021). See the work of: Debourdeau, 2013; Servigne and Stevens (2015); eco-feminists (Hache, 2016, Starhawk 2019). This research and data on the environment invite us to shift the paradigm in all areas, including ethics. This is where the ethics of care becomes especially relevant today.

While ethical perspectives on AI are flourishing, they are all based on dominant moral theories that originate from the same conception of the human being, i.e. a rational, autonomous and independent being. The ethics of care specifically criticizes the liberal premise on which these moral theories depend. Since the modern age, rationality has been recognized as humans' main characteristic.

In other words, humans are profoundly and fundamentally rational beings. However, reason has been set in opposition to feelings, emotions, relationships, in short, all the things that materially evoke life, aiming, on the contrary, for abstraction and a conceptual representation of reality. Objectivity would therefore be the target of a properly rational disposition where one would deal with an object external to oneself, in a neutral and impartial way. Morality has also been associated with this rationality and impartiality. We can only be moral if we strive for universal good. For Kohlberg, moral maturity is associated with an individual's ability to formulate abstract principles of justice that apply to everyone. Therefore, in order to be morally just, one must use the veil of ignorance, to put it in Rawlsian terms, which means having to desubjectify oneself, to hold what is called "a moral point of view," a political and social non-place, in short, *to be a stranger in the middle of nowhere*. Morality would therefore be considered from the standpoint of an individual who is isolated from the world, autonomous, rational and independent of others.

The ethics of care specifically criticizes this conception of the human being. Brugère (2008) examines the fantasized representation in philosophical literature of the human being who doesn't need anyone to build themselves up, recognizing namely that, "the independent individual is indeed one of the great theoretical fictions of our western mythology" (Brugère, 2008, p. 50). Vulnerability, meaning the capacity of a person to be hurt and injured, is completely eliminated from the dominant moral theories as it is associated with a form of fragility which could affect the formulation of a judgement that is independent and neutral, and therefore fair. However, what care ethicists affirm is the essential and unexceptional nature of human vulnerability (Paperman, 2011). We have all experienced vulnerability to the extent that we needed care to remain alive. Whether the individual is an infant, an adolescent, sick or bereaved, they need care to continue living. Therefore, care is necessary, and not contingent. This ontological recognition allows us to affirm other elements. Indeed, if we intrinsically need care, then we are absolutely linked to others through interdependence or interconnection (Perreault, 2015). The living world as a whole, including non-humans and living territories, is also vulnerable (Laugier, 2012). Thus, the world must be understood as a vast network of variable and particular vulnerabilities that make it possible, through bonds and relationships of care, to maintain and preserve what we have in common, namely life.

Therefore, we possess a highly powerful critical tool, not so we can question the ethical conclusions on AI, but rather the premises on which they are based, rendering them in fact, obsolete, or even completely null. Indeed, we can absolutely question any current ethical perspective on AI insofar as it is based on an artificial—or even completely fantasized—conception of the human being. Thus, we can affirm that any ethic that omits care activities in its reasoning is missing the target because it is failing to consider the fundamental activity of human life, the one that enables it to exist.

Considering that this moral reasoning does not recognize something specific to human nature as being part of a community and its care needs, how can we believe that such reasoning is able to meet our civilizational and technological requirements? Following this critique of the ontological premises of the dominant ethics, the next section will question the form of their moral reasoning and offer ways to emerge from an *ethical saturation in AI*.

EMERGING FROM AN ETHICAL SATURATION: FROM THE PRINCIPLE TO THE EVERYDAY

For some years now, we have observed what we call ethical saturation in the field of AI. We can define ethical saturation as the effect of an increased redundancy of certain AI-related concepts, judgements and proposals, thereby reflecting the same vision of the world and its fundamental values. This saturation is even more obvious when we look at the number of normative frameworks developed to structure AI, which are, for the most part, similar in form and content (Voarino, 2019, p. 170). There appears to be a form of consensus regarding the principles to apply in the field of AI (Jobin et al., 2019). This ethical saturation can be explained in two ways.

First, the normative frameworks are inspired by reasoning that is specific to the theories of justice. Above all, a justice perspective aims to identify abstract principles that are deemed superior, because they *transcend* any form of material specificity or particularity. What is just or good is validated based on adherence to universal principles, separate from ordinary life. From a justice perspective, this set of rules enables the world to function properly.

In addition, theories of justice use a process that makes it possible to free oneself from the political conditions of existence. This is called a "moral point of view." This moral point of view is used to see or view situations by someone who is capable of thinking beyond the world and adopting an outside point of view in order to understand precisely the right or wrong thing to do.

However, the performativity of these principles in the world, i.e. the way in which they unfold in the field of existence, remains to be created, for the subjects of this world. While deliberation may be an interesting tool for thinking about the process that goes from abstraction of the principles of justice to their social and cultural materialization (Noiseau et al., 2021), this form of moral functioning is incomplete, considering the virtues of the care model. There are important aspects of care ethics that help to address these shortcomings.

First, as shown throughout her work and specifically in a series of interviews with women who had an abortion, Gilligan (1982) states that acting fairly does not only consist in following a set of abstract rules, but rather to record one's voice differently in a life context marked by relationships and emotions: "Women see moral dilemmas as a matter of responsibilities and concern for the welfare (care) of others, not as a matter of rights and rules" (Gilligan, 2019, p. 117). In other words, there is a reversal of the morality paradigm: It is no longer a question of going from the top down, from the principle towards real action, but of responsibly moving back and forth between oneself and the other. To put it another way, it's about consistently asking the following questions: How can I take care of others as well as of myself? Am I sustaining and pursuing life? Have I been accountable to others? How can I maintain relationships? Thus, the ethics of care becomes another normative ideal type distinct from the theories of justice (Clement, 1996; Perreault, 2015). This ideal type focuses on the context of the subject's life rather than on a disembodied and abstract non-place; it favours the connections between members of an ecosystem instead of an independent posture experienced as separateness from the world; and finally, it recognizes the morality of action criterion as the ability to maintain human relationships instead of guaranteeing a formal and legal principle of equality.

Second, the ethics of care enables us to place the moral theory creation in the material world. Tronto (1993, p. 62) argues that "morality is always contextual and historicized, even when it claims to be universal." In other words, it is a question of affirming the profoundly political character of qualifying what is fair and which behavior is considered good in the world. Indeed, Tronto recognizes the inability

of philosophers to respond to the current challenges we face[78] by the fact that these theories are voluntarily disengaged from the conditions that made it possible to sustain life in this world: "Ironically, it is precisely the strength of universal moral theory, its detachment from the world, that makes it inadequate to solve the kinds of moral problems that now present themselves." (Tronto, 1993, p. 152-153).

Finally, the ethics of care enables a disruption of epistemological conceptions in morality since it does not originate from an abstract moral point of view. Rather, it is foremost inscribed within a subject situated in the world, a conception largely inspired by feminist epistemologies (Harding, 1993). This would be a total reversal of the perspective, meaning that instead of being extracted from the world in which the subject exists, it takes root entirely in this world in order to acknowledge the place it occupies in all representations and symbols that make up the world. This is what Perreault (2015) calls the situated process of subjectivation. The aim of this process is to recognize oneself in the social space, as a marker of sexual and gender differentiation: "The difference between care and justice is not just a moral or linguistic difference; it also implies a difference in the sensory experience of gendered or sexualized individuals, situated in a symbolic territory that distinguishes them from each other" (Perrault, 2015, p. 46). This experiential distinction conveys namely a way of being with others and with oneself. To see and recognize oneself as being dependent on others means to accept behaving differently with others, because there is an awareness of the ties that bind us. In its aspiration for autonomy and independence, the justice perspective also contributes to denying what psychically constitutes the subject in the world: "In this respect, the sexual difference specific to the patriarchal system would require negation of the fundamental bonds that bring subjects together in the common space. Contributing to the negation of others, the separation that underpins this negation is likewise articulated in the negation of self" (Perreault, 2015, p. 47). Using a formal and abstract approach to morality would therefore have much more significant spiritual and philosophical consequences than one might imagine as by denying their rootedness to the world, the individual also rejects the foundational characteristic of their existence, i.e. the fact of being a living subject.

So, what conclusions can we draw from our ethical analysis of AI? On one hand, it would seem that the frameworks relating to the development or deployment of AI are not very constructive as they are based on a fragmented and incomplete conception of the human being. On the other, as we have just demonstrated from a moral perspective, these frameworks use the justice theory that does not take into account the real and concrete factors that enable an ecosystem to follow its obvious creed, namely to live. Conducting ethical analyses of AI from a moral point of view and a justice perspective cannot explain the global nature of the issues AI encompasses. This is because the moral point of view would incur adopting a *single* form of moral reasoning, that of justice, which, as we recall, is based on an erroneous conception of the human being. All of the normative frameworks such as principles, rules, charters and recommendations on AI suddenly become unfinished, even incomplete, because they do not include the ins and outs of a life context, in which the players are engaged in networks of specific relationships—including human beings, animals and living territories. Consequently, these frameworks establish norms that ultimately do not correspond to anything real. Moreover, the use of a single form of moral reasoning by decision-makers and intellectuals, whether it is done willingly or through ignorance, has major repercussions on the world. Indeed, framing the development of AI using abstract and universal principles of justice is essentially using the ethics frameworks that have specifically led

78. The list of current issues that we face is very long. To name but one, and not the least, there is the ecological state of the planet. See the latest WWF report entitled, "Living Planet Report 2016. Risk and resilience in a new era" (WWF, 2016). We can draw a link between philosophies that promote a liberal conception of the individual and the repercussions of human activities on the planet. By separating the individual from his life context (including human and non-human ecosystems), by setting man up as a species superior to the other species on Earth, the human being has been internally incapable of assessing the consequences of these actions because he is convinced of being the only truly valuable being on the planet.

us to the generalized crisis in which we find ourselves today. How can we believe that to save ourselves from the misdeeds of a technology, we should use the ethical tools that have, as yet, been unable to prevent them? We will discuss this further in the next section.

Moreover, with regard to the assumed moral determination integrated in AI, this notion is profoundly disrupted. Indeed, consider the idea of an ethical AI, a virtuous robot or fair algorithms. All of these proposals stem from the justice theory perspective and, therefore, from its unfounded philosophical beliefs. Among the many works addressing the possibility of inserting a morality of action criterion into algorithms, ethics of care perspectives are set aside. Is this lapse due to ignorance of alternative and feminist ethics? Or on the contrary, is there a structural interest in not analyzing AI from an ethics standpoint that calls entirely into question our ontological and normative preconceptions? The idea of a virtuous (Gibert, 2020), good or just robot would make no sense from a care perspective. Indeed, from a care standpoint, goodness or justice cannot be recognized according to a descending principle because it exists immanently in a life context marked by interdependent relationships with the living as a whole. If we wanted AI to be fair from a care standpoint, it would be necessary to integrate a provision aimed at acting in such a way that it maintains a relational network in order to maintain life, which would require great emotional and relational intelligence from this AI. While it may be possible to integrate a care provision because it relates to interiority and therefore to consciousness, the problem lies with the practical dimension of care. Could an AI take care of people if caring was work? A practice? An action? Moreover, although we want an AI to be ethical from a care standpoint, how could it do this? Being subject to programming, it is necessarily decontextualized and removed from the place it inhabits. The challenges relating to temporal consciousness and to care as a material practice seem difficult to overcome in this day and age. It would appear that we are at somewhat of an impasse: An ethical AI *only* seems possible in a context where justice theory is valued and not the ethics of care.

Therefore, it would not be far-fetched to affirm that AI has a gender, the existence of which would reveal our patriarchal history (Nurock, 2019). By establishing itself as neutral and impartial, AI would, on the contrary, be inspired by and pursue the moral point of view that we have widely questioned. Nurock identifies a significant pitfall, namely the potential artificialization of ethics through the insertion of historically inherited domination structures that support morally and politically unjustified dynamics in a technology that is supposedly neutral or impartial (Nurock, 2019). After having shown that AI ethics originates from a falsified ontological preconception and that it uses disconnected moral forms, we will see that the approach used to structure AI goes against the need for responsibility and attention represented by the ethics of care.

BEYOND A RISK FRAMEWORK: RESPONSIBILITY AND ATTENTION

One could formulate the hypothesis that the ethics of AI originates in the misdeeds of its object. In other words, AI ethics has arisen because we have identified implications of AI that have a potential or actual negative effect on reality and the individuals who constitute this reality. To put it another way, AI ethics is relevant insofar as it is a tool used for the structure and normative management of the implications deemed harmful to the world. If AI did not have deplorable effects, there would be no ethics of AI as it would simply be deemed good for humanity. Doubt as to the acceptability of its use has led to the development of AI ethics.

Notably, most citizen deliberations on the uses of AI—For example, the citizen-led processes of the Montréal Declaration for a Responsible Development of AI (2018) and the Open Dialogue on AI Ethics (2020) —focus on evaluating what are referred to as the “ethical issues of AI,” in other words, its potential or actual risks on the proper functioning of a society. In this sense, the integration of AI into the

realm of human existence entails an acceptance of its risks. The ethical structure in which we live would have accepted the idea of a risk society, i.e. a space for human interaction where there would necessarily be hazards or dangers[79].

In her book entitled *Le risque ou le care?*, Tronto (2012) questions two conceptions of the management of human affairs: risk and care. The risk society, theorized namely by author Ulrich Beck, would accept, and even claim the idea that we would no longer be able to contain the unexpected effects of certain actions, in other words, we would live in a society "out of control" (Beck, 1998). This risk society would emanate from modernity, resulting from a collective discharge of responsibility in the face of industrial and technological development. Tronto largely questions this social conception. The author believes this vision of the world is linked to a gendered experience as the masculine has largely been associated with protection in the face of danger: "Risk society creates the image of a 'risky' world, which induces an understanding of the social world as dangerous and linked to the human duty of protection and management. Thus, risk society operates in a metaphorically masculine universe." (Tronto, 2012, p. 25). It is because the world is risky that we must manage it, protecting and governing individuals. However, what Tronto tells us is that, conversely, a care society would primarily place the responsibility of care on others: "(...) care supposes that individuals become autonomous and capable of acting on their own through a complex process of growth, of development, through which they are interdependent and transformed throughout their life" (Tronto, 2012, p. 33). Risk management would be groundless in a care society as risks would be contained and prevented by the fact that needs are taken care of collectively and responsibly from the outset: "The care society presupposes that people live in a world where they constantly deal with vulnerability and need—sometimes also experiencing joy" (Tronto, 2012, p. 46). Consequently, the intention of AI ethics, which seeks to frame its responsible development, would start from a risk-based social conception.

The ethics of care invites us to pay special attention to the world (Garrau, 2014). Here, paying attention means listening and hearing the voices of those who are at the heart of a particular ecosystem in order to determine whether care is adequate or not (Garrau, 2014). Assessing the opinions—on the use of AI in this case—of the main interested parties, through any device, is absolutely necessary to determine whether the care provided to meet the need is adapted and appropriate to the situation and to the points of view of the actors involved. In other words, the determination of social and political needs must include the people who carry out our society's essential activities; otherwise we will continue to perpetuate structural injustices: "If there is no deliberation or if it takes place without taking into account the voices of care providers, the inequalities that structure care relations in contemporary liberal societies are bound to be repeated" (Garrau, 2014, p. 66). In light of this, the question is: Were care providers consulted before developing certain AI applications? To date, no specific consultation has been carried out with so-called front-line people, i.e. people working in the care professions (we are referring to the fields of education, housekeeping, agri-food, nursing science, etc.).

Finally, this chapter presents some proposals for the potential development of an ethical AI, understood here to mean care-based. One of the proposals, which was developed by Nurock et al. (2021), is to rethink the structure of ethics from its conception, commonly referred to as ethics by design, by including care as one of the key criteria. The process involves answering four questions, making it possible to create ethical care-based AI from the design stage. These concerns are: 1) What do we care

79. This type of life, which creates risks and dangers, is all the more obvious when it helps to reinforce structures of oppression and violence against women. I am referring here to sexual robotics and its symbolic, material and political effects. See my master's thesis research (2018).

about? 2) What or who are we doing it for? 3) Are we providing care? 4) Are we caring with others?[80]. These questions, if answered in good faith and in a responsible manner, i.e. by being actively involved in the process, enable us to determine and recognize the profoundly ethical nature of this use of AI in a world where all beings are vulnerable, interdependent, and therefore need care above all to live, or even to survive.

Moreover, it is important to note and recognize that there are certain developments in AI that only present themselves as serving major common causes. This is called "AI for humanity," "AI for the common good" or "AI for the planet." While these types of AI are deployed with the aim of a better community life, it remains to be confirmed whether they adequately and relevantly answer the questions posed by the ethics of care by design, as presented above.

CONCLUSION

In her book entitled *Courage Calls to Courage Everywhere*, Winterson (2018) wondered if AI was the worst thing that could happen to women. We asked ourselves another question: Could AI be in line with a feminist, democratic and inclusive ethic, i.e. an ethic of care? To answer this, from an ethics of care standpoint, we began by questioning the ontological foundations on which the main ethics theories base their moral judgement of AI. Then, using the criticisms raised by ethics of care philosophers, we adopted an ethical perspective by debunking the form of moral reasoning used in the dominant ethics, namely the moral point of view specific to the theory of justice. Finally, we showed that the ethics of AI becomes relevant in a risk society and not in a care-based society. In this sense, the ethics of care possesses a very powerful critical and revolutionary quality to question the ethics of AI and its conclusions on the world. The goal of this chapter was to emerge from ethical saturation in AI to broaden our view of the different forms of life that we can formulate together. From this new ethical perspective, we must determine whether AI remains relevant or not, from an ethics of care standpoint.

The ethics of care represents an extraordinary opportunity for the earth, for living beings and for humans to see themselves differently, to establish new relationships and to build a just and equitable world with a fairly simple goal, namely the preservation of life on Earth. It remains to be seen what place AI should or could occupy in this new world.

80. These four questions are followed by equally relevant reformulations: 1) *What is important to us in the development of AI?* 2) *Have we attended to the most vulnerable?* 3) *Have we taken care to safeguard users' choices and integrated their requirements, rights, needs, etc. in the system?* 4) *How do we govern AI democratically and remain mindful of the transformations that AI is capable of bringing about in our democratic institutions and in the public arena?* (Nurock et al., 2021)

REFERENCES

Algora Lab. 2020. *Le dialogue inclusif sur l'éthique de l'IA. Contribution à la recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle*. University of Montreal et Mila and Quebec Institute of Artificial Intelligence. https://opendialogueonai.com (Accessed 11 March 2022.)

Beck, U. 1998. Le conflit des deux modernités et la question de la disparition des solidarités: liens personnels, liens collectifs. In: *Lien social et politiques*, RIAC, Vol. 39, p. 15-25.

Brugère, F. 2008. *Le sexe de la sollicitude*, Paris, Seuil.

––––. 2011. *L'éthique du « care »*. Paris: Presses Universitaires de France.

Boden, M.A. 2005. *The Philosophy of Artificial Intelligence*. Oxford, Oxford University Press.

Clement, G. 1996. *Care, Autonomy, and Justice. Feminism and The Ethic of Care*. Boulder, Westview Point.

Crawford, K. 2021. *Atlas of AI*. New Haven,Yale University Press.

Debourdeau, A. 2013. *Les grands textes fondateurs de l'écologie*. Paris, Flammarion.

Dilhac, M., Christophe, A. and Voarino, N. 2018. *Rapport de la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*. Montreal, University of Montreal.

Fisher, B. and Tronto, J. 1991. Towards a feminist theory of care. In: Abel, B. and Nelson, M. (eds.). *Circles of Care: Work and Identity in Women's Lives*. New York, State University of New York Press.

Garrau, M. 2014. *Care et attention*, Paris, Presses universitaires de France.

Gibert, M. 2020. *Faire la morale aux robots. Une introduction à l'éthique des algorithmes*. Montreal, Atelier 10.

GIEC. 2021. *Climate Change 2021. The Physical Science Basis*. Groupe d'experts intergouvemental sur l'évolution du climat. https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Full_Report.pdf.

Gilligan, C. 1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, Harvard University Press.

––––. 2010. Une voix différente. Un regard prospectif à partir du passé. In: Nurock, V. (dir.) *Carol Gilligan et l'éthique du care*. Paris, Presses universitaires de France.

––––. 2011. Une voix différente. Un regard prospectif à partir du passé. In: Laugier, S. and Paperman, P. (eds.), *Le souci des autres. Éthique et politique du care*. Paris, Éditions de l'École des hautes études en sciences sociales.

––––. 2019. *Une voix différente. La morale a-t-elle un sexe?* (French version translated by Annick Kwiatek). Paris, Flammarion.

Gilligan, C., Hochschild, A. and Tronto, J. 2013. *Contre l'indifférence des privilégiés: à quoi sert le care?* Paris, Payot.

Goodfellow, I., Bengio, Y. and Courville, A. 2016. *Deep Learning*. Cambridge, MIT Press.

Hache, E. 2016. *Reclaim. Recueil de textes écoféministes*. Paris, Éditions Camourakis.

Hamrouni, N. 2015. *Le care invisible: genre, vulnérabilité et domination*. Doctorate thesis. University of Montreal and Catholic University of Louvain.

Harding, S. 1993. Rethinking Standpoint Epistemology: What is "Strong Objectivity"? In: Alcoff, L. and Potter, E. (eds.). *Feminist Epistemologies*. New York, Routledge.

Jobin, A., Ienca, M. and Vayena. E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol.1, pp. 389-399.

Laugier, S. 2012. *Tous* vulnérables*? Le care, les animaux, l'environnement*. Paris, Payot.

Noiseau, P. Lanteigne, C., Flores Echaiz, L., Gomez Salazar, F.G., Mai, V., Dilhac, M-A and Mörch, C-M. 2021. Le dialogue inclusif sur l'éthique de l'IA: délibération en ligne citoyenne et internationale pour l'UNESCO. In: *Communication Technologies et développement.*

Noiseau, P. 2019. *Les enjeux éthiques de la robotique sexuelle: une perspective critique féministe*. Master's dissertation. University of Montreal.

Nurock, V. 2019. L'intelligence artificielle a-t-elle un genre? Perspectives philosophiques sur l'artificialisation de l'éthique, du social et du politique. *Cités*, Vol. 80, pp. 61-74.

Nurock, V., Chatila, R. and Parizeau, M-H. 2021. What does "Ethical by Design" Mean? In: Braunschweig, B. and Ghallab, M. (eds.), *Reflections on Artifical Intelligence for Humanity*, Vol. 12600. Springer International Publishing.

Paperman, P. 2011. Les gens vulnérables n'ont rien d'exceptionnel. In: Paperman, P. and Laugier, S. (eds.) *Le souci des autres. Éthique et politique du care*. Paris, Éditions de l'École des hautes études en sciences sociales.

––––. 2021. D'une voix discordante: désentimentaliser le *care*, démoraliser l'éthique. In: Molinier, P., Paperman, P. and Laugier, S. (eds.) *Qu'est-ce que le care? Souci des autres, sensibilité, responsabilité*. Paris, Payot.

Perreault, J. 2015. Renégocier la « voix différente »: retour sur l'œuvre de Gilligan. In: Bourgault, S. and Perreault, J. (eds.) *Le care: éthique féministe actuelle*. Montreal, Les éditions du remue-ménage.

Servigne, P. and Stevens, R. 2015. *Comment tout peut s'effondrer*. Paris, Seuil.

Starwak. 2019. *Quel monde voulons-nous?* (French version translated by Isabelle Stengers). Paris, Éditions Cambourakis.

Tronto, J. 1993. *Moral Boundaries. A Political Argument for an Ethic of Care*. New York, Routledge.

––––. 2009. *Un monde vulnérable. Pour une politique du care.* (French version translated by Hervé Maury). Paris, Éditions La découverte.

––––. 2012. *Le risque ou le care?* (French version translated by Fabienne Brugère). Paris, Presses Universitaires de France.

Voarino, N. 2019. *Systèmes d'intelligence artificielle et santé: les enjeux d'une innovation responsable*. Doctorate thesis, University of Montreal.

WWF. 2016. *Planète vivante 2016. Risque et résilience dans l'Anthropocène*.

Artificial Intelligence (AI) has an increasingly profound impact on our societies. As scientific and technological developments accelerate at an unprecedented rate, it is crucial that we also promote a comprehensive and inclusive dialogue on how to oversee and guide their advancement. In this context, Mila and UNESCO have joined forces to compile a publication of 18 selected submissions from a global open call for proposals launched in 2021, featuring the perspectives of academics, civil society, and innovators to help shift the conversation on AI from what we do know and foresee to what we do not comprehend yet, the missing links in AI Governance.

With this publication, Mila and UNESCO aim to provide policymakers, innovators, academics, and civil society with fruitful perspectives to help us face the immense task we are presented with: shaping the development of AI so that no one is left behind. This means working towards AI systems that are human-centered, inclusive, ethical, sustainable, as well as upholding human rights and the rule of law.

9 789231 005794