Editors
LUCA BELLI and WALTER B. GASPAR

AI FROM THE GLOBAL MAJORITY

Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition

IGF Internet Governance Forum
FGV DIREITO RIO

AI from the Global Majority

Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition

Edition produced by FGV Direito Rio
Praia de Botafogo, 190 | 13th floor
Rio de Janeiro | RJ | Brasil | Zip code: 22.250-900
55 (21) 3799-5445
www.fgv.br/direitorio

AI from the Global Majority

Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition

Edited by *Luca Belli* and *Walter B. Gaspar*

FGV Direito Rio Edition

Printed in Brazil.
1 edition finalized in November 2024.
This book is in the Legal Deposit Division of the National Library.

Coordination: FGV Direito Rio
Book cover: Tangente Design
Layout: Tangente Design

CONTENTS

ABOUT THE AUTHORS

Luca Belli, PhD, is Professor of Digital Governance and Regulation at Fundação Getulio Vargas (FGV) Law School, Rio de Janeiro, where he directs the Center for Technology and Society (CTS-FGV) and the CyberBRICS project. Luca is also editor of the International Data Privacy Law Journal, published by Oxford University Press, and Director of the Computers Privacy and Data Protection conference Latin-America (CPDP LatAm). He is currently member of the Brazilian Presidency's National Committee on Cybersecurity, Board Member of the Global Digital Inclusion Partnership and member of the Steering Committee of the Forum for Information & Democracy. He is author of more than 80 publications on law and technology and his works have been quoted by numerous media outlets, including The Economist, The New York Times, Financial Times, Forbes, Le Monde, BBC, China Today, The Beijing Review, The Hill, O Globo, Folha de São Paulo, El Pais, and La Stampa. Luca holds a PhD in Public Law from Université Paris Panthéon-Assas and can be found on LinkedIn and on Twitter as @1lucabelli

Walter B. Gaspar. Lawyer with a master's in public health from UERJ (2017), studying the interface between innovation, intellectual property and access to medicines policies in Brazil. Grantee of the Fundación Botín Programme for the Public Interest in Latin America (2013). Researcher in the Fiocruz and Shuttleworth Foundation project on intellectual property and access to medicines (2017). National Coordinator of the NGO Universities Allied for Essential Medicines (2013-2016). Certified Graphic Designer by the Istituto Europeo di Design (2018). Currently, researcher in the CyberBRICS and Data Regulations projects at FGV's Center for Technology and Society and Ph.D. candidate at the Public Policies, Strategies and Development Programme at the Economics Institute of the Federal University of Rio de Janeiro (UFRJ). Substitute Advisor to the Municipal Council for the Protection of Personal Data and Privacy of the city of Rio de Janeiro.

Alice Rangel Teixeira is a PhD candidate in Philosophy at the Universitat Autònoma de Barcelona (UAB). Her research focuses on the ethics of artificial intelligence from a decolonial feminist

perspective. She integrates the Politics and Ethics of Public Health research project and the Humanistics Group on Science and Technology at the UAB. She is a member of the IGF Data and Artificial Intelligence Coalition (DAIG), and the Artificial Intelligence and Data Visitation of the Research Data Alliance (RDA).

Amrita Sengupta. Amrita is a Research and Programme Lead at the Centre for Internet and Society, India. A trained sociologist, Amrita's research interests and work lie in the areas of artificial intelligence, trust and online harms, platform accountability, gender and technology, and sustainability and tech. In the past, Amrita has worked in managing and implementing large scale people practices, diversity and inclusion in the workplace, as well as in conducting and leading long-form research on impacts of tech on businesses and society, with both quantitative and qualitative methodologies.

Andrea Bauling holds a Master of Laws in Private Law from the University of Pretoria and a Master of Education in Open Distance Learning from the University of South Africa. Since 2012, she has been a lecturer in the Department of Jurisprudence in the School of Law at the University of South Africa. She acts as treasurer for the Southern African Society of Legal Historians. Her current research interests include legal education, open distance learning, digital visual literacies, and data privacy in higher education.

Anuti Shah is an Undergraduate Student at the University of Virginia studying Computer Science and Environmental Thought and Practice with a Minor in Data Science and a Research Assistant on the project "Data Justice and Climate Resilience in the Global Automotive Industry." She is a previous Technology and Data Products Intern at National Geographic Society and a Data Analysis Intern at the United States Geological Survey, particularly interested in examining the environmental impacts that come with an increasingly digital world.

Avantika Tewari, a Ph.D. candidate at Jawaharlal Nehru University in New Delhi and a former senior research associate at IT for Change, is currently a Visiting Fellow at the Centre for the Study of Developing Societies. She has co-authored the book "Psychoanalysis and ChatGPT" with Slavoj Žižek and has edited a volume of essays titled, "Feminist Perspectives on Social Media Governance." Additionally,

Avantika is the co-convenor of a Global Working Group on Feminist Digital Justice. She also played a key role in coordinating the collaboration between Friedrich Ebert Stiftung (FES) — Berlin and IT for Change's project, "Feminist Visions of the Digital," including making a submission at the UN's 66th Commission on the Status of Women Forum, themed "Feminist Futures in the Digitalizing World of Work," emphasizing the vision of "The Deal We Always Wanted".

Bárbara Lazarotto, LSTS, VUB. Barbara joined LSTS in February 2022 as a PhD student. She works in the H2020 LeADS — Legality Attentive Data Scientists project under a Marie-Sklodowska-Curie ITN grant. Her research focuses on Public-Private Data sharing.

Chinasa T. Okolo. Dr. Okolo is a Fellow in the Center for Technology Innovation at The Brookings Institution and a recent Computer Science Ph.D. graduate from Cornell University. Her research focuses on AI governance in the Global South, datafication and algorithmic marginalization, and the socioeconomic impact of data work. Dr. Okolo has been recognized as one of the world's most influential people in AI by *TIME* and advises numerous multilateral institutions, national governments, corporations, and nonprofits. Her research has been covered widely in media outlets and published at top-tier venues in human-computer interaction and sociotechnical computing.

Dennis Ramphomane, LLB, Advocate of the South African High Court, Member of the National Bar Council of South Africa.

Catherine Bielick, Dr. is an infectious diseases physician and a data scientist researching bias in healthcare AI, especially related to people with HIV and underrepresented groups. She is passionate about healthcare justice, medical ethics, and practices clinically at Beth Israel Deaconess Hospital.

Ekaterina Martynova, Ph.D. candidate, Lecturer at the School of International Law of Higher School of Economics in Moscow, head of the research group "From Science Fiction to Legal Science: State Responsibility in Cyberspace", executive editor of the "HSE University Journal of International Law".

Elise Racine is an AI expert, human rights advocate, and doctoral candidate at the University of Oxford, where she is also a Clarendon

Scholar and recipient of the Baillie Gifford-Institute for Ethics in AI Scholarship. Affiliated with the Institute for Ethics in AI and the Ethox Centre, her research examines the socio-ethical implications of AI, particularly in global health and international development. This includes algorithmic harms (bias, colonialism), global power dynamics, reparative and participatory accountability mechanisms, and transformative AI developments across Asia, Africa, and the Middle East. Elise brings interdisciplinary experience spanning the academic, non-profit, public health, development, and private sectors. She has worked with organizations such as Kiva, BlueDot Impact, Reclaim Childhood, the Edgewood Center for Children and Families, and the Global Infectious Disease Ethics (GLIDE) Wellcome Open Research Gateway. She holds an MPA with a concentration in Digitalization and Big Data (Hertie School), MSc in Health and International Development (London School of Economics), BA in Sociology (Stanford University), and Certificate in International Migration Studies (Georgetown University).

Faizo Elmi holds a Bachelor of Arts in History and is completing a Bachelor of Education in Secondary Education. She is passionate about the intersection between technology, policy, and education. Her research explores how AI can address educational challenges faced by the global majority, making learning more accessible and tailored to diverse needs. Faizo's work emphasizes cultural sensitivity and equitable access to technology, advocating for a thoughtful approach to integrating AI in varied educational settings. Her commitment to global education reflects a dedication to innovative solutions that support inclusive and adaptive learning.

Guangyu Qiao-Franco, Dr., is an Assistant Professor of International Relations at Radboud University and a Senior Researcher of a European Research Council-funded project AutoNorms that investigates international regulations surrounding military AI. Her recent publications from this project include a special issue with Global Society that features various interdisciplinary studies on algorithmic warfare (co-edited with Prof. Ingvild Bode), and several journal articles and reports on the evolution of Chinese AI policy. She is currently working on articles on dual-use technologies-related export controls and China's policymaking on AI and arms control.

Hellina Hailu Nigatu is a PhD student in Computer Science at the University of California at Berkeley. Her research is at the intersection of NLP and HCI, particularly focusing on African communities and languages. She got her BSc in Electrical Engineering from Addis Ababa Univerity in Ethiopia and her MSc in CS from UC Berkeley. Hellina is also an active member of Maskhane. She is a SIGHPC Computational and Data Science fellow and a 2024 FAccT DEI Scholar.

Isha Suri is a Research Lead at the Centre for Internet and Society, India where she manages the research verticals on telecommunications, digital competition, and internet governance. She co-authored the Market Study on the Telecom Sector in India (2021), commissioned by the Competition Commission of India and is a member of several research bodies including the Data Governance Network, India, the India Internet Governance Forum (IIGF) Thematic Committee. She has also worked as a consultant with the Ministry of Electronics and Information Technology, India and the National Internet Exchange of India on ICANN-related research and a multistakeholder engagement assistance program.

Jess Reia is an Assistant Professor of Data Science at the University of Virginia, a faculty lead at the Digital Technology for Democracy Lab (UVA Karsh Institute) and a Non-Resident Fellow at the Center for Democracy & Technology. In 2024, Reia joined Fudan University's Institute for Global Public Policy as a visiting scholar. They work primarily on topics of data justice, urban governance, and technology policy transnationally. Before joining UVA, they were appointed Mellon Fellow at McGill University (2019-2021) and worked at the Center for Technology & Society at FGV Law School in Rio de Janeiro from 2011 to 2019.

Julio Gabriel Mercado is a researcher and practitioner, focused on the intersection of open judicial data and their uses for driving innovation and digital transformation within the judiciary. He works with institutions and key stakeholders to implement data publication processes, with a focus on their quality and reuse in value creation. A highlight of his career was contributing to the development of Argentina's first open judicial data portal at the Ministry of Justice. He is a professor of open justice at the Pontifical Catholic University of Argentina (UCA) and leads the Open Data Group of the International

Open Justice Network (RIJA), where he advocates for greater judicial data transparency, accessibility and reuse.

Laura Middeldorp, MSc, Scientist, Unit Defense Safety and Security, The Dutch Applied Sciences Institute, TNO. Laura Middeldorp works as a scientist at the Netherlands Organisation for Applied Scientific Research (TNO) in the department of Intelligence and Decision Support. Her research focuses on Intelligence, Information Operations, Land Operations and Artificial Intelligence (AI). She obtained a Master of Science degree in Applied Mathematics at Delft University of Technology, where she specialised in statistics and uncertainty quantification.

Leo Celi, Dr. Celi is a prominent researcher, thought leader, critical care physician, and teacher. He works internationally to lay foundation for open data, community development, and educational outreach.

Liisa Janssens, LLM MA. Lead Scientist of an Interdisciplinary Team working on the nexus of the Rule of Law and Technology at the Unit Defense, Safety & Security, TNO. She holds dual master's degrees in Law and Philosophy from Leiden Law School and Leiden University. In different engineering teams, she has the lead for formulating new questions on how to responsibly navigate AI development processes. Her expert roles include serving as an external expert on NATO's Data AI Review Board, the European Defence Agency, and as a Reserve Member of the European Group on Ethics in Science and New Technologies. Janssens is recognized for her work on AI applications in military and law enforcement contexts, with two significant reports: one for NATO on AI applications in counter-unmanned aircraft systems, focusing on aligning AI with the Rule of Law in military contexts, and another for the Dutch National Police on technical AI requirements that comply with the Rule of Law. Her position paper, "Responsible AI and the Rule of Law," showcases her extensive experience in integrating legal and technical perspectives.

Liu Shaoyu is a Researcher at Guanghua Law School of Zhejiang University. He qualifies to supervise masters and doctors. He holds a bachelor's degree in management from the East China University of Political Science and Law, a master's degree in law from China University of Political Science and Law, a master's degree in law

from the University of Hamburg, Germany, a doctorate in law jointly trained by China University of Political Science and Law and the University of Munich, and a postdoctoral fellow in law from the Institute of Law, Chinese Academy of Social Sciences. His fields of interest include administrative law and digital law.

Liu Zijing is a PhD candidate at Guanghua Law School of Zhejiang University. Her research interests are digital government, digitization of administrative discretion, and Artificial intelligence law.

Mahmoud Javadi is a doctoral researcher at the Centre for Security, Diplomacy and Strategy (CSDS) in Vrije Universiteit Brussel (VUB). At CSDS he is involved in the European Research Council-funded project "Competition in the Digital Era: Geopolitics and Technology in the 21st Century" (CODE). Prior to this, Mahmoud was an AI Governance Researcher at Erasmus University Rotterdam (EUR) in the Netherlands, contributing to the EU-funded research consortium "Reignite Multilateralism via Technology" (REMIT). He also has experience with the Carnegie Endowment for International Peace, where his research focused on EU external relations.

Masego Morige, BCom Law, LLB, Candidate Attorney (Awaiting Admission).

Matheus Alles. Professor of Law at the Lutheran University of Brazil (ULBRA). Master 's degree in Law from the Department of Economic and Labor Law at the Federal University of Rio Grande do Sul (UFRGS). Specialist in Labor Law and International Relations from UFRGS. Member of the research group on Law and Fraternity at the same institution. Co-responsible professor for the research group on Justice and Consensual Means of Conflict Resolution at ULBRA.

Mbali Nzimande, BA (Law), Paralegal, Snail Attorneys.

Nils Brinker, Cyber Security Expert at intcube, Berlin.

Pablo Trigo Kramcsák, LSTS, VUB. Pablo joined LSTS in January 2021 as a PhD student, funded by the "Becas Chile scholarship in digital transformation and technological revolution", awarded by the Chilean National Research and Development Agency. His PhD research focuses on legitimate interest as an appropriate lawful basis for processing Artificial Intelligence training datasets, addressing the

difficulties involved, risks, and potential impacts on data subjects' rights and freedoms.

Rachel Leach is an Undergraduate Student at the University of Virginia studying Sociology, Government, and Data Science and a Research Assistant on the project "Data Justice and Climate Resilience in the Global Automotive Industry." She is particularly interested in how the AI industry works to shape data privacy and environmental policy and research, and possible alternatives to this.

Richard Ngamita is an experienced researcher specializing in influence operations, disinformation, and digital security across the Global South. With over 10 years of experience, he is the founder of Thraets but previously led research at major tech companies including Twitter, Facebook, and Google. Ngamita's work focuses on understanding digital threat behaviors, analyzing social media's impact on democratic processes, and advocating for ethical technology development. He has collaborated with civil society organizations, contributed to training on investigative journalism and digital security, and frequently speaks on data, security, and disinformation topics. Ngamita has a strong background in open-source intelligence, data analysis, and research methods.

Richard Skalt, Cybersecurity Advocacy Manager at TÜV SÜD, Munich.

Rocco Saverino, LSTS, VUB. Rocco joined LSTS in July 2022 as a PhD student. He is part of the ALTEP-DP project and focuses his research on the critical role of Data Protection Authorities in enforcing the regulations on data protection and considering their impact in the context of AI (Act).

Rodrigo Gameiro, Dr. Gameiro is a physician and a lawyer working with MIT Critical Data. His main interests include regulation science and characterizing machine learning bias in healthcare.

Ronald Musizvingoza, PhD, is a researcher at the United Nations University Institute in Macau, a UN think tank dedicated to exploring the impact of digital technology on Sustainable Development Goals (SDGs). His work focuses on generating and synthesizing evidence for research and policy, particularly in the areas of AI ethics, digital poverty, gender, and data. Ronald's research emphasizes power,

privilege, equity, and sustainability in digital technology through a feminist decolonial lens.

Saskia Lensink, PhD, Scientist, Unit ICT, The Dutch Applied Sciences Institute, TNO. Saskia Lensink is a consultant at the Netherlands Organisation for Applied Scientific Research (TNO) specializing in language and speech technologies, with a PhD in Linguistics from Leiden University. As the Product Manager of GPT-NL, she leads cross-functional teams in developing a Dutch Large Language Model, managing data acquisition and curation, algorithmic modeling, system architecture, and end-user engagement across sectors such as safety and security, healthcare, and government. Saskia also leads use case development for the HORIZON TrustLLM project, driving the creation of sovereign European LLMs with a focus on Germanic languages. In her role as co-lead of the Data & AI Technologies Task Force at the Big Data Value Association (BDVA), she provides expert guidance and recommendations to empower industry in leveraging AI and data technologies responsibly and sustainably, while advocating for trusted and secure data practices. A frequent speaker at events, Saskia combines technical expertise, stakeholder management, and a results-driven approach to shape the strategic vision for LLMs and generative AI in Europe.

Shiva Kanwar is a Fellow at the Indian Council for Research on International Economic Relations (ICRIER) where she works on contemporary digital policy issues. She is a lawyer by training with over seven years of experience in the digital policy space and national & international multistakeholder processes in interdisciplinary contexts. Her areas of professional experience include Intellectual Property Rights, emerging technologies, and digital equity. She has also worked with the Ministry of Electronics and Information Technology, Government of India, as a consultant. She holds a Bachelor's degree in Law and an LL.M. in Intellectual Property Rights.

Shweta Mohandas. Shweta is a Researcher at the Centre for Internet and Society, India. Her areas of work include Artificial Intelligence, Privacy, and Intellectual Property Rights and India's policies surrounding them. She is currently researching on developments in health data from digitisation to use deployment of AI in healthcare.

Sizwe Snail Ka Mtuze, LLB — UP, LLM — SA, Attorney of the South African High Court as well as Adjunct Professor, Nelson Mandela University and Visiting Professor, CTS-FGV FGV University.

Ying Lin is a PhD candidate at the Vrije Universiteit Brussels (VUB). He is a doctoral researcher at the Law, Science, Technology and Society (LSTS) research group and a Cyber and Data Security Lab (CDSL) member. He also practices as a lawyer in China. His academic interests span data governance, cross-border data transfers, AI regulation, and the international interoperability of legal digital regimes.

Yonah Welker. Former Tech Envoy, Ministry of AI Advisor, human-capacity-centered AI policy and transatlantic safety dialogue, oversee and evaluate ai, tech, IoT, deeptech inno programs. Former founder — Hardwaretech (2005), big data and language-related technologies in public / administration applications (2014). Curator of AI Summits and hackathons (over 240 technologies, OECD — 120 technologies). Visiting lecturer — MIT, organizer of many scientific panels, groups and programs. Author — MOOCs and open programs (AI, ethics, policy). Tech and business mentor — EU programs, MIT, Insead, Masschallenge. Contributed to research, development and adoption frameworks, including reports and frameworks — OECD, UNESCO, UNDP/UN (AI in education, health, public service and administration, digital infrastructure and policy), supported ontologies and taxonomies of assistive and accessibility AI, public and learning systems.

Zeerak Talat, **Dr.**, is a Chancellor's Fellow in Responsible ML and AI at the Centre for Technomoral Futures at the Edinburgh Futures Institute and the Institute for Language, Cognition, and Computation at Edinburgh University. They work on the intersection between machine learning, science and technology studies, and media studies. Their research seeks to examine how machine learning systems interact with our societies and the downstream effects of introducing machine learning to our society.

PART 1

LOCAL APPROACHES TO GLOBAL PROBLEMS

1 AI from the global majority: What are we debating and why?

Luca Belli and Walter Britto Gaspar

Abstract

The first Annual Report of the UN IGF Data and Artificial Intelligence Governance (DAIG) Coalition, released at IGF 2023, focused on "The Quest for AI Sovereignty, Transparency, and Accountability." Building on this outcome, the DAIG Coalition initiated a multistakeholder effort to discuss "AI from the Global Majority," aiming to provide insights for IGF 2024. This paper provides an introduction to this volume, which was compiled from an open Call for Essays, with the aim to analyse AI initiatives from the perspectives of populations in Africa, Asia, Latin America, and the Middle East. These regions, often underrepresented in AI governance discussions, share a history of colonial exploitation and face ongoing neo-colonial and digital colonialism dynamics, which are becoming particularly evident as regards their adoption and regulation of AI as well as their capacity to contribute to global AI fora. The essays emphasise the need for inclusive and equitable representation in global AI dialogues. They explore the impact of AI on civil, political, economic, and social rights, addressing issues like surveillance, labour displacement, and environmental degradation. The volume advocates for AI systems designed with diverse data sets and inclusive practices to mitigate biases and promote fairness. Importantly, it outlines not only problems faced by the global majority, but also relevant solutions emerging from such countries.

Keywords: AI, Artificial Intelligence, Global Majority; Global South; Internet Governance Forum; IGF.

Introduction

This volume presents the results of the 2024 works undertaken by the Data and Artificial Intelligence Governance (DAIG) Coalition[1]

1 For further information about the DAIG Coalition of the UN Internet Governance Forum, see https://intgovforum.org/en/content/dynamic-coalition-on-data-and-artificial-intelligence-governance-dc-daig.

established under the auspices of the United Nations Internet Governance forum (IGF). The Coalition is a multistakeholder group aimed at fostering discussion of existing approaches to data and AI governance, promoting analysis of good and bad practices to identify what solutions should be replicated and which ones should be avoided by stakeholders to achieve a sustainable an effective data and AI governance.

To do so the DAIG Coalition aims at promoting studies and multistakeholder efforts to collect and discuss evidence, critically analyse existing and proposed regulatory and institutional arrangements, and suggest policy updates in AI governance. Importantly, the DAIG will act as a hub to connect global UN IGF discussions with regional and local initiatives, with a particular focus on Global South debates.

After having successfully released at the IGF 2023 the first Annual Report of the Coalition, featuring analyses from 34 authors dedicated to "The Quest for AI Sovereignty, Transparency and Accountability" (Belli, L. and Gaspar, W.B., 2023), the DAIG Coalition has promoted a multistakeholder effort aimed at discussing "AI from the Global Majority", to provide valuable inputs that could feed into IGF 2024 discussions and beyond. Authors of this volume responded to an open Call for Essays, shared over the DAIG mailing list, with the aim to collect valuable insight analysing AI initiatives from the perspectives of global majority populations. Indeed, most DAIG members felt that such perspectives are often underrepresented in discussions of data and AI governance, as noted in a dedicated multistakeholder workshop organised by DAIG Coalition members during the Computers Privacy and Data Protection Conference Latin America (CPDP LatAm) 2024.[2]

The term "global majority" refers to the populations of an ample range of highly heterogenous countries from Africa, Asia, Latin America, and the Middle East, which together make up most of the world's population. This concept challenges the traditional Eurocentric perspective, highlighting the importance of recognising and valuing

2 See the report of the CPDP LatAm 2024 session on "AI from the Global Majority: Meeting of the UN IGF Coalition on Data and AI Governance" available at https://cpdp.lat/en/.

the diverse range of cultures, histories, and contributions from the abovementioned regions.

Importantly, despite their heterogeneity, almost all countries and populations from the global majority share the feature of being former colonies of global north countries, which implemented extractive practices, concentrating resources and violently subjugating local populations for multiple centuries, thus considerably contribution to the enormous inequalities that still characterise these countries. Importantly, this volume starts form the assumption that some of these exploitative practices merely evolved into neo-colonial dynamics, while other new forms of digital colonialism have emerged over the past decades (Quijano, 2000). As a result, global south countries find themselves in a very thorny situation, trying to shape their AI approaches while being in a situation of clear dependency and lack of essential transparency and accountability tools, which are necessary to elaborate solid strategies, policies and regulations.[3]

By focusing on the global majority, we emphasise the need for understanding the point of views, the needs and the perspective of this global majority of countries, in a perspective of inclusivity and equitable representation in global dialogues and decision-making processes. This shift in perspective is crucial for addressing global challenges in a way that is fair and just for all.

1.1 AI and Human Rights in the Context of the Global Majority

Artificial Intelligence (AI) significantly impacts civil and political rights, especially for the global majority. Surveillance technologies powered by AI can infringe on multiple fundamental rights, especially privacy, data protection, and freedom of expression. In many countries, these technologies are used to monitor and suppress dissent, disproportionately affecting marginalised communities. For instance, facial recognition systems, often less accurate for people of color, can lead to wrongful arrests and increased surveillance of minority groups. Additionally, AI in law enforcement and judicial

3 Such consideration is corroborated by the general lack of AI sovereignty in most countries, as highlighted in the 2023 Outcome of the DAIG Coalition. See supra n. (2).

systems can perpetuate existing biases, resulting in unfair treatment and discrimination. In this context, it is essential to develop and implement AI technologies that respect and protect civil and political rights, ensuring that they do not become tools of oppression.

Furthermore, AI's influence extends to economic and social rights, where it can both create opportunities and exacerbate inequalities. Automation and AI-driven technologies have the potential to transform industries and create new jobs. However, they also pose a risk of labour displacement, particularly in regions where economies are heavily reliant on low-skilled labour. For example, the introduction of AI in manufacturing and agriculture can lead to job losses for workers who lack the skills to transition to new roles.

Access to AI technologies and the benefits they bring is often uneven, further widening the gap between the global majority and more developed nations. To address these challenges, it is crucial to invest in education and training programs that equip workers with the skills needed for the AI-driven economy.

As an instance, the rise of AI and automation has significant implications for labour markets, particularly in the global majority. While AI can enhance productivity and create new job opportunities, it can also lead to labour exploitation. Workers in low-wage jobs may face increased pressure and job insecurity as companies adopt AI technologies to cut costs. For instance, gig economy workers may be subjected to algorithmic management practices that prioritize efficiency over worker well-being. It is important to develop policies and regulations that protect workers' rights and ensure fair labour practices in the AI-driven economy.

As this volume illustrates, AI systems can perpetuate exclusion and discrimination if not designed and implemented with inclusivity in mind. Biases in data and algorithms can lead to discriminatory outcomes, affecting access to services, employment, and justice. For instance, biased hiring algorithms can disadvantage candidates from certain backgrounds, while biased credit scoring systems can limit access to financial services for marginalised communities. The lack of representation of the global majority in AI development exacerbates these issues, as the perspectives and needs of these

populations are often overlooked. It is essential to ensure that AI systems are developed with diverse data sets and inclusive practices to mitigate these risks.

The underrepresentation of the global majority populations, interests, and perspectives in AI research and development means that their unique challenges and contexts are not adequately addressed. This lack of diversity in the tech industry leads to the creation of AI systems that do not serve the needs of all users equally, reinforcing existing inequalities. For example, language models trained primarily on English data may not perform well for speakers of other languages, limiting their accessibility and usefulness. Increasing the representation of the global majority in AI development is crucial for creating technologies that are equitable and inclusive.

Furthermore, global majority populations are likely to be the ones suffering g the most from the environmental impact of AI infrastructure, which is increasingly recognised as a critical issue. The energy consumption of AI systems, particularly those requiring large-scale data processing and storage, contributes to carbon emissions and environmental degradation. This impact is felt disproportionately in regions already vulnerable to climate change, many of which are part of the global majority. For example, data centres in developing countries may strain local energy resources and contribute to pollution. To mitigate these effects, it is essential to develop energy-efficient AI technologies and promote sustainable practices in the tech industry.

The following section provides an overview of how the issues are explored in this volume, illustrating the complexities that the global majority is facing but also the solutions that are emerging from these countries.

1.2 How is this volume addressing these issues?

This volume examines the transformative impact of Artificial Intelligence (AI) on contemporary societies. Papers are organised around five thematic axes that provide a structured exploration of those impacts and proposed frameworks and solutions to existing issues, with contributions from diverse regional and global

perspectives focused on some of the key challenges which are particularly relevant for the Global South.

The first section, **Local Approaches to Global Problems**, delves into how nations tailor AI-driven solutions to address unique domestic challenges while engaging with broader global trends. Luca Belli's paper, "AI Meets Cybersecurity: A Brazilian Perspective", evaluates the role of AI in cybersecurity both as defensive and an offensive tool. Belli situates the discussion in Brazil and advocates for integrated multistakeholder governance frameworks through the creation of a "Brazilian Cybersecurity and Digital Transformation System". Subsequently, Sizwe Snail and colleagues offer a constructive critique of South Africa's AI governance landscape in "The Law on Artificial Intelligence in South Africa". Their analysis of the South African Draft AI Strategy (SADAIS) and the National AI Policy Framework highlights the urgency of aligning national policies with evolving regional and global legal standards — an effort that, as the authors demonstrate, is still lacking in the country.

Additionally, Zijing Liu et al. provide an empirical study of China's judicial innovations in "Building Smart Courts Through Large Legal Language Models". Their work reflects on the current landscape of AI-enabled legal decision-making in Chinese courts, drawing lessons from regional variations in smart court implementation. Finally, Nils Brinker and Richard Skalt, in "Fox Guarding the Chickens", expose inherent biases in the EU AI Act's risk management obligations, emphasizing the critical need for impartial mechanisms to address third-party risks. This points to relevant blind spots in the European model of AI regulation — an important point to be considered by Global South countries amidst the so-called "Brussels effect" on various regulatory fronts, indicating that, while learning from the achievements and shortcomings of the European experience is useful, an original and context-adequate regulation is crucial for these countries.

In the second section, **The Emergence of Regional Solutions**, the focus shifts to regional strategies that navigate AI's complexities within distinct socio-political contexts. Pablo Trigo Kramcsák et al. analyse the "Incipient Latin American Approach to AI Governance", showcasing how Latin American countries have been establishing

their own AI frameworks influenced by EU regulatory principles. These are early-stage efforts and face significant challenges, especially in promoting regulatory coordination between existing and new authorities. Chinasa T. Okolo's "RICE Governance Framework" proposes a cohesive strategy for African nations, emphasizing the need to reform governance measures, integrate regulatory efforts, improve regional cooperation, and boost enforcement. Andrea Bauling explores the legal challenges of AI in education in "AIED and Student Data Privacy in Africa", highlighting the need to craft Africa-centric policies focused on AIED that protect student data while fostering technological innovation. Ekaterina Martynova contributes a commentary on the Council of Europe Framework Convention on AI and Human Rights, which provides, through its soft regulatory approach, common ground and a first step toward international regulatory approaches — a model that could inspire the BRICS. Finally, Yonah Welker introduces a human-capacity-centred AI policy emphasizing disability inclusion and emphasizes the need for disability representation in policy geared toward AI.

The third section, **Global Majority Facing AI**, centres on equity and justice for the Global Majority, addressing exploitation and systemic inequalities perpetuated by AI. Elise Racine's "Reparative Algorithmic Impact Assessments" outlines a justice-oriented framework for mitigating the harms of AI-powered systems through an approach that combines culturally sensitive participatory methods and a reparative praxis and decolonial, Intersectional principles. Alice Rangel Teixeira challenges the ethical foundations of mainstream AI principles in "AI Ethics for the Global Majority", proposing decolonial feminist bioethics as an alternative approach focused on power relations, relational autonomy, shared responsibility, empirical evidence, and local contexts.

This theme continues with analyses of content moderation harms and how they disproportionately affect Global Majority communities, with Zeerak Talat and Hellina Hailu's "Exploitation all the way down: Calling out the root cause of bad online experiences for users of the 'majority world'"; and the socio-political implications of false information, compounded by generative AI technologies, as well as

the way forward in Isha Shuri and Shiva Kanwar's "Countering False Information: Policy Responses for the Global Majority in the Age of AI".

Richard Ngamita's contribution, "Addressing the Challenges of AI Content Detection in the Global South", explores the limitations of existing AI systems in detecting manipulated media, particularly cheap fakes, and advocates for the development of AI models trained on local data, alongside inclusive content moderation policies, to safeguard civic participation and democracy in the Global South. This axis is closed by Guangyu Qiao-Franco and Mahmoud Javadi's "Bridging the gap between the North and South in the governance of dual-use artificial intelligence technologies", on the implications of dual-use AI technologies, highlighting the critical need for equitable global AI governance.

In **Social Challenges of AI**, the fourth axis, the discussion broadens to include labour, education, and environmental sustainability. A case study from MIT Critical Data by Catherine Bielick, Rodrigo Gameiro and Leo Celi underscores the necessity of inclusive AI development practices and how to achieve them, while Avantika Tewari critiques various aspects of the relations between platforms and users and how conceptualizing data subjects as prosumers reinforces issues related to participation and labour and to the effective control over personal data.

Papers on healthcare and education explore how AI can address, but also exacerbate, existing inequities. Amrita Sengupta and Shweta Mohandas, in "Cost or Benefit? The Impact of AI on the Work of Medical Practitioners", analyse the integration of AI into healthcare practices, focusing on its current use and impact on medical workflows in India. Through primary research, the authors highlight both the potential benefits and the challenges for medical professionals. Faizo Elmi's "Reimagining Education: Potential Solutions for Nomads", explores how AI technologies, such as adaptive learning platforms and virtual classrooms, can address the educational challenges faced by nomadic populations, highlighting the need to overcome barriers like technological infrastructure and cultural adaptation to ensure equitable and effective implementation.

Finally, Jess Reia, Rachel Leach, and Anuti Shah, in "The Need for Transnational Perspectives on the Social, Legal and Environmental

Impact of Artificial Intelligence", argue for integrating environmental justice into AI regulatory frameworks. By examining cases in the US and Brazil, they highlight the geopolitical and ecological costs of AI development and propose incorporating hidden costs especially affecting marginalised and global majority communities into the transnational regulatory ecosystem.

The final section, **Foresighted Solutions for Present Problems**, offers innovative approaches to pressing AI-related challenges. Matheus Alles examines the ontological shifts at the intersection of law and data science, advocating for a reflexive legal rationality. This would allow for effective and ethical integration of data science in the legal field, through a process of critical assessment of the legal community and adaptation to new forms of knowledge and rationality.

Julio Gabriel Mercado, in "People-Centered Justice AI: Data Dimensions for Embracing a Responsible Digital Transformation", discusses how effective digital transformation of justice must go beyond mere technological adoption, incorporating Open Justice principles such as transparency, accountability, and public participation. Liisa Janssens, Saskia Lensink, and Laura Middeldorp, in "Fostering AI Research and Development: Towards a Trustworthy LLM", discuss compliance challenges and ethical considerations in the development of Large Language Models (LLMs) through a scenario-based analysis, focusing on the implications of including or omitting an opt-out option for personal data removal. Finally, Ronald Musizvingoza examines the potential of using synthetic data to create representative datasets that reflect diverse gender experiences while addressing the risks of bias and misuse. Collectively, these contributions emphasize the importance of aligning technological advancements with ethical imperatives and human-centred design principles.

By weaving together these thematic axes, this volume provides a comprehensive understanding of AI's transformative role in society. It not only highlights the potential benefits of AI but also critically engages with its risks, especially in the face of Global Majority communities and countries, advocating for a context-specific and balanced perspective for inclusive, equitable, and sustainable AI governance.

1.3 References

Belli, L, Curzi, Y. & Gaspar, W. B. (2023). AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience, Computer Law & Security Review 48, 105767, https://doi.org/10.1016/j.clsr.2022.105767.

Belli, L. and Gaspar, W.B. (Eds.) The Quest for AI Sovereignty, Transparency and Accountability. Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition. FGV. (2023). https://hdl.handle.net/10438/34295.

Quijano, Anibal (2000) 'Coloniality of Power, Eurocentrism, and Latin America', Nepantla: Views from South, 1(3), pp. 533– 580; Everisto Benyera. (2021). The Fourth Industrial Revolution and the Recolonisation of Africa: The Coloniality of Data. Routledge.

2 AI Meets Cybersecurity: A Brazilian Perspective on Information Security and AI Challenges

Luca Belli

Abstract

Artificial Intelligence (AI) has transformed the cybersecurity landscape over the past decade, leading to an increase in the frequency, impact, and sophistication of cyberattacks. While AI can be leveraged by organisations to enhance their cyber defences, detecting cyberthreats and improving decisions about how to react, it can also be exploited by cybercriminals to launch targeted attacks at an unprecedented speed and scale, bypassing traditional detection measures. This paper starts by exploring the distinction between defensive AI and offensive AI in the context of cybersecurity. Subsequently, it focusses on the Brazilian context to explore how the country is dealing with the emerging threats and opportunities presented by the intersection of AI and Cybersecurity. Lastly, it puts forward some concise recommendations for policymakers advocating for multistakeholder cooperation to be embedded in the future Brazilian Cybersecurity Strategy and Brazilian Cybersecurity Agency, to cope with the increasing complex intersection between cybersecurity and AI. Ideally, such recommendations could be integrated in the proposals for a new strategy and agency that will be issued by the new National Cybersecurity Committee (known as "CNCiber") a multistakeholder advisory body recently established by the Brazilian Presidency.

Keywords: AI, Artificial Intelligence, Cybersecurity, Information Security, Brazil.

Introduction

Artificial Intelligence (AI) has transformed the cybersecurity landscape over the past decade, leading to an increase in the frequency, impact, and sophistication of cyberattacks. While AI can be leveraged by organisations to enhance their cyber defences, detecting cyberthreats

and improving decisions about how to react, it can also be exploited by cybercriminals to launch targeted attacks at an unprecedented speed and scale, bypassing traditional detection measures.

Indeed, the increasing use of AI systems in a wide range of processes in various safety-critical sectors — such as health, justice (Salomão, 2022), autonomous vehicle-management, etc. — creates numerous new, and sometimes unpredictable, risks and can open new avenues in attack methods and techniques.(Belli et al., 2023; ENISA, 2020) Such risks may be maximised when AI is deployed for automated decision making, directly affecting both individuals and organisations, thus leading legislators around the world, including in Brazil, to consider appropriate risk regulations aimed at framing AI systems.

This paper adopts the definition of AI system offered by article 4 of the latest version of Bill 2338/2023, which is largely based on the definitions offered by the EU AI Act and the OECD (Russell et al., 2023), and therefore highly unlikely to be altered. The definition proposed by the Brazilian bill reads as follows:

> *"artificial intelligence (AI) system: a machine-based system that, with different degrees of autonomy and for explicit or implicit purposes, infers, from a set of data or information it receives, how to generate results, in particular, prediction, content, recommendation or decision that can influence the virtual, physical or real environment."*[4]

Importantly, the dual nature of AI allows to utilise such technology to both strengthen and undermine cybersecurity. However, while both the aforementioned Brazilian AI Bill and other leading examples of AI regulation such as the EU AI Act consider cybersecurity a key concern for the development and deployment of AI systems, neither offers clear guidance on how to concretely assess risks and implement regulation cybersecurity aspects of IA.

4 Bill No. 2,338, of 2023 on the development, promotion, ethical and responsible use of artificial intelligence based on the centrality of the human person. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.

In this respect this paper argues that considerable work is needed to support the implementation of existing and proposed frameworks, particularly through the adoption of technical standards able to specify and give meaning to highly vague formulations, that are typically adopted by AI regulatory frameworks to define cybersecurity risk management provisions.

First, this paper explores the distinction between defensive AI and offensive AI in the context of cybersecurity. Second, it focusses on the Brazilian context to explore how the country is dealing with the emerging threats and opportunities presented by the intersection of AI and Cybersecurity and what type of provisions are dedicated to the issue in the proposed AI Bill. Lastly, it puts forward some concise recommendations for policymakers.

2.1 AI and Cybersecurity: A complicated relationship

The relationship between AI and cybersecurity is based on how the former is used to impact the latter and vice versa, and the resulting defensive, offensive, or adversarial capabilities (Belli et al., 2023). While there is already a conspicuous body of research on the technical aspects of AI and cybersecurity, it is surprising that remarkably scarce research exists on the interactions of AI and cybersecurity from a regulatory and governance angle. This essay aims at understanding what is at stake when we adopt this latter angle and policy issues should be considered as priorities.

To do so, we should initially distinguish between defensive AI and offensive AI. Defensive AI usually leverages machine learning and other AI techniques to enhance the cybersecurity and resilience of computer systems, networks, and data bases, and to protect individuals, shielding them against cyber threats (Geluvaraj, 2019). In this perspective, AI systems can increase the effectiveness of security controls aimed at protecting specific assets, for instance through automated malware analysis, active firewalls, automated cyber threat intelligence operations, etc. (Belli et al., 2023).

In contrast, offensive AI, also known as AI-powered cyberattacks, involves the use of AI to launch malicious activities, enhancing vulnerability detection and exploitation, developing new cyberattacks

types and strategies or automating the exploitation of existing vulnerabilities. Lastly, we should mention the that adversarial AI is a subcategory of offensive AI and refers to the manipulation of AI systems to cause them to make incorrect predictions. This can be achieved by tampering with the input data or poisoning the training data used to develop the AI system (Malatji, 2024).

Importantly, this paper adopts the definition of cybersecurity provided by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), which is noteworthy for being a rare example of consensual definition at the international level, according to which:

> *"Cybersecurity is the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and user's assets. Organization and user's assets include connected computing devices, personnel, infrastructure, applications, services, telecommunications systems, and the totality of transmitted and/or stored information in the cyber environment. Cybersecurity strives to ensure the attainment and maintenance of the security properties of the organization and user's assets against relevant security risks in the cyber environment."* (ITU-T., 2009)

The amplitude provided by the above definition tellingly illustrates the complexity of cybersecurity and the necessity of adopting a collaborative and coordinated multistakeholder approach to the issue, as no single stakeholder can guarantee cybersecurity in a vacuum.

2.2 A paradigm shift

The integration of AI capabilities has constituted a watershed moment in the development of cyber threats, significantly augmenting the efficacy, scope, scale, and precision of malicious cyber operations. This evolution marks a paradigm shift in the cybersecurity landscape,

fundamentally altering in multiple ways the nature of both offensive and defensive strategies.

First, the democratisation and increased sophistication of AI tools enables cybercriminals to automate and refine their attacks, making them more effective, callable, dynamic, and difficult to detect. Machine learning algorithms, for instance, can analyse vast amounts of data to identify vulnerabilities in systems and networks, enabling attackers to exploit these weaknesses with greater precision. Automated phishing campaigns can be tailored to individual targets based on data harvested from social media and other sources.

This personalisation increases the likelihood of success, as the messages appear more convincing and relevant to the recipient. Critically, concern about AI-enhanced malicious attacks now represents the top emerging risk according to the latest version of the periodic Gartner study dedicated to risk monitoring, due to "the relative ease of use and quality of AI-assisted tools, such as voice and image generation, increase the ability to carry out malicious attacks with wide-ranging consequences" (Gartner, n.d.).

Second, AI is likely to expand the scope of cyberthreats by allowing attackers to increase the scale of their operations with minimal human intervention. As an instance, AI-powered botnets can be used to operate massive Distributed Denial-of-Service (DDoS) attacks, able to overwhelm electronic networks. Ransomware attacks are also becoming more sophisticated and widespread due to AI support, leading to the consolidation of Ransomware-as-a-Service (RaaS) as a thriving industry with global range. In this context AI is sensibly lowering barriers to entry for attackers, increasing ease and availability of ransomware, via AI-driven malware capable of quickly and autonomously spread across networks, encrypt data, and demand ransoms, leading to high cost of recovery and downtime (Hassan, 2023).

Third, AI systems can substantially increase attackers' ability to analyse complex datasets and recognise patterns, thus allowing to execute highly targeted and precise attacks. For example, AI can be used to identify high-value targets within organisations and tailor attacks to their specific roles and responsibilities. AI can also allow cybercriminals to create realistic audio and video impersonations,

that can be considered as “deepfakes”, which can be used in social engineering attacks to manipulate individuals into divulging sensitive information or authorising fraudulent transactions (MIT Technology Review Insights, 2021). It is now memorable the case of an elaborate deepfake scam, where a finance worker at a multinational firm was duped into paying USD 25 million to fraudsters who had lured him into a fake emergency call (Cen, Magramo, 2024).

Fourth, the increasing sophistication of deepfakes can be used to orchestrate disinformation campaigns for both financial and political purposes. These technologies pose a novel cybersecurity threat to of democratic processes by enabling malicious actors to undermine information integrity at an unprecedented scale. The current democratisation of AI implies much greater and easier access to AI systems that until just few years ago were only accessible to researchers and highly specialised companies or governmental actors. This process leads to an enormous expansion of the attack surface, both in terms of potential perpetrators and in terms of potential vulnerabilities and attack strategies that can be used (Pupillo, 2021).

Importantly, AI-driven cyberattacks have acquired a dynamic nature, being able to adapt to changing defensive measures, making detection and mitigation more challenging. By using machine learning capabilities, attackers can alter malicious software in real time to avoid detection by traditional antivirus systems. For instance, AI-enhanced polymorphic or metamorphic malware is able to mutate its features or automatically “recoding” itself when it propagates to evade pattern matching detection systems that are traditionally deployed as security solutions. Furthermore, AI systems can be used to quickly identify and exploit zero-day vulnerabilities before patches can be developed and deployed (ENISA, 2020).

Crucially, defenders are also increasingly employing AI-based systems to detect cyber threats and vulnerabilities and rapidly respond e.g., leveraging AI to identify software bugs and self-patch them. However, within a sort of cybersecurity arms race, attackers are also leveraging AI to outmanoeuvre these defences. This situation where both sides continuously refine their techniques, defensive AI systems must evolve rapidly to detect new attack patterns and anomalies, while policy and governance framework must be crafted

to mitigate risks and facilitate communication, collaboration and coordination amongst cybersecurity stakeholders.

2.3 Understanding the Brazilian Context

Despite relevant advancements in recent years, the regulation of AI and cybersecurity in Brazil is highly fragmented, limited and poorly implemented. Due to the adoption of multiple sectoral regulations dedicated to cybersecurity, Brazil has climbed several rankings,[5] but the regulatory oversight and cybersecurity implementation remains patchy, being the responsibility of many different and uncoordinated entities, including sectoral regulators, private and public Computer Security Incident Response Teams, and the military (Belli et al., 2023).

While Brazil does not have a general cybersecurity law, the top institution responsible for cybersecurity governance and policy proposal is the Institutional Security Cabinet (GSI in its Portuguese acronym) of the Brazilian Presidency. However, the GSI remit is limited to the federal administration thus limiting enormously the scope of its reach. Importantly, in December 2023, Brazil adopted a new National Cybersecurity Policy and established a new multistakeholder National Cybersecurity Committee (Brazilian Presidency, 2023) (known as "CNCiber"), of which the author of this paper has been appointed a member (Brazilian Presidency, 2024; CyberBRICS, 2024). Amongst the tasks of CNCiber is the elaboration of a proposal for a new national cybersecurity strategy and a new body for cybersecurity governance and regulation.

Indeed, one of the reasons of the fragmented Brazilian approach to cybersecurity is the lack of a unique institution responsible for coordinating the various dimensions of it. At the same time, the previous Brazilian National Cybersecurity Strategy expired in December 2023. Hence, at the moment of this writing, Brazil does not have an actionable cybersecurity strategy allowing the country to organically tackle the multiple — and mounting —

5 Most notably, in 2020 Brazil jumped up 53 positions, from 71st to 18th, in the Global Cybersecurity Index (GCI) elaborated by the International Telecommunications Union. In the Americas region, Brazil reached the 3rd position, surpassed only by the USA and Canada. The 2024 edition of the GCI considers Brazil as a "Tier 1 — Role-modelling" country. https://www.itu.int/epublications/publication/global-cybersecurity-index-2024.

cyberthreats it faces and assess the ways in which AI technologies are impacting such threats.

Furthermore, only limited AI regulation exist, falling primarily under the purview of the Brazilian Data Protection Authority, ANPD in its Portuguese acronym. In this context, the Brazilian National Congress is currently considering regulating AI with a dedicated framework, which would include cybersecurity obligations related to AI systems. While numerous AI bills are under consideration, Bill 2338/2023 seems to be the most complete and well-structured, being the result of multiple years of hearings and multistakeholder consultations. However, at the time of this writing, the Brazilian Congress has not adopted the Bill yet.

2.3.1 Information security?

Information security is an essential dimension, common to both AI and cybersecurity. In Brazil, the National Data Protection Authority (ANPD) is tasked with enforcing the Brazilian General Data Protection Law (LGPD) and ensuring that organisations comply with data protection obligations, including regarding the implementation of data security obligations. Data security is a fundamental principle set by the LGPD, aimed at ensuring that data is protected against unauthorised access, loss, alteration, damage, or destruction. Importantly, the LGPD explicitly establishes a security-by-design obligation for data controllers and processors, who need to implement security measures that the data subject "can expect", to demonstrate that personal data processing activities are regularly undertaken (Article 44).

According to Article 46 of the LGPD, "The processing agents must adopt security, technical, and administrative measures capable of protecting personal data from unauthorised access and from accidental or unlawful situations of destruction, loss, alteration, communication, or any form of inappropriate or unlawful processing."[6] In particular, indent 2 of this article highlights that information security measures "must be observed from the design phase of the product or service until its execution." Additionally, Article 49 specifies that "The systems used for processing personal data must

6 The Brazilian General Data Protection Law (LGPD) — Unofficial English Version. CyberBRICS. (2020). https://cyberbrics.info/brazilian-general-data-protection-law-lgpd-unofficial-english-version/.

be structured to meet security requirements, best practices, and governance standards, as well as the general principles provided for in this Law and other regulatory standards."[7]

To comply with the LGPD, processing agents are supposed to implement solid information security solutions. Such measures are suggested by ANPD in Orientation Guide (ANPD, 2021) and include administrative measures, such as i) the definition of an information security policy; ii) awareness raising and capacity building; iii) and contract management; as well as the establishment of technical measures, such i) the establishment of access controls to ensure that only authorised individuals have access to personal data; ii) the use of security measures such as encryption to protect personal data during storage and transmission; iii) backup and recovery to ensure data availability in case of loss or damage; iv) and vulnerability monitoring and detection, to promptly identify and respond to data security breaches.

In practice, however, data security compliance is poor at best, given the total absence of ANPD oversight as regards this matter, in the first four years since its inception, despite the enormous and growing amount of information security incidents in Brazil. Indeed, the tropical giant ranks second globally for cyberattacks (Nakamura, 2024), which have exploded in number and sophistication due to the adoption of AI systems, making them more complex and difficult to detect, as exposed previously.

The ANPD is the body responsible for overseeing and regulating the implementation of LGPD, including regarding data security. However, so far, the ANPD has not regulated data security, despite having the possibility to do so, having simply adopted the above-mentioned Orientation Guide and a recent Regulation on the Communication of Information Security Incidents (ANPD, 2024).

However, focusing on the communication of cybersecurity accidents rather than on overseeing the implementation of the existing information security obligation seems rather counterproductive. It is rather absurd to invest resources in overseeing the communication

7 Ibid.

of the tragedy rather than on the implementation of the norms that would avoid or at least mitigate the tragedy itself. Furthermore, despite having a clear mandate to enforce the LGPD provisions on data security, no single sanction has been adopted so far for lack of compliance with such norms, in a country where the number of cybersecurity incidents is raising exponentially and only 2023 registered 103 billion cyberattacks (Belli et al. 2023).

A more interesting approach has been recently adopted by worth noting that the Ordinance SGD/MGI No. 852 of 28 March 2023 established the Privacy and Information Security Program (PPSI)[8] dedicated to enhance cybersecurity of the Brazilian public administration. Data governance in the Brazilian public sector is rather heterogeneous with most public administrations still having very basic cybersecurity governance despite the enormous digitalisation that Brazilian public services undertook since the Covid19 pandemic (Belli et al., 2024). The PPSI program is therefore a welcome initiative, characterised by a set of projects and adaptation processes aimed at increasing cybersecurity maturity, resilience, effectiveness, collaboration, and intelligence.

The LGPD and PPSI should be considered as two essential information security pillars, but not sufficient on their own. In this context, it is essential that the future Cybersecurity Strategy, to be proposed by the National Cybersecurity Council, specify information security criteria for categories of sensitive information that are not personal.[9] Furthermore, it seems desirable that the future Brazilian Cybersecurity Agency establish cooperation agreements, and ideally a coordination mechanism, with the ANPD as well as other sectoral regulators with mandate to ensure cybersecurity in their specific sectors, in order to enhance much needed coordination.

Indeed, the sole existing coordinating body for information security is the Information Security Management Committee, another GSI body representing of multiple governmental entities, with very limited

8 The ordinance was issued by the Secretariat of Digital Government of the Ministry of Management and Innovation in Public Services. Further information on the program and can be found at https://www.gov.br/governodigital/pt-br/privacidade-e-seguranca/programa-de-privacidade-e-seguranca-da-informacao-ppsi.

9 Such specification is utilised e.g., by the Chinese Cybersecurity Law and Data Security Law, which prescribe the adoption of specific measures to protect "important" or "core" data whose security is essential for the well-functioning of national critical infrastructure. See Belli L. (2021).

impact. In its current configuration, this Committee can promote joint regulatory actions and the development and implementation of coordinated policies. However, over the past years, this Committee has been incapable of establishing any concrete initiatives, having promoted no multistakeholder interaction, or proposed not even a single educational, capacity building or compliance-promotion effort.

2.3.2 An "appropriate" way of regulating AI?

It is important to emphasise that both cybersecurity and AI are quintessentially multidimensional matters. Information security is only one for the many dimensions that compose them and, for each dimension them, different regulations, regulators, and regulated entities may already exist.

The success of both cybersecurity an AI governance depend on having a good understanding of how the different component of digital and AI technologies interact, how they are utilised, and what can be the vulnerabilities in their use and deployment (Safitra, 2023).

Sound management of information and infrastructure, good stakeholder coordination, and solid capacity-building are therefore essential. However, as stressed, at the Brazilian level each dimension or component of both AI and cybersecurity is regulated by multiple entities with limited or no coordination at all.

As mentioned in the introduction, Brazil is in the process of elaborating a new AI framework. However, several critiques can be raised as regards both the way in which the framework proposes to regulates cybersecurity aspects of AI and the way in which it proposes to foster coordination amongst sectoral regulators.

In its article 2, the Bill 2338/2023 usefully states that the guarantee of information security and cybersecurity is one of the fundamental principles of AI regulation. However, it subsequently includes a considerable amount of vaguely worded cybersecurity provisions. These include the obligations to adopt "**appropriate** information security measures along the entire AI system lifecycle" (article 17); "perform test to assess the **appropriate** levels of reliability, consistent performance, safety" of AI systems (article 18); or conceive and develop AI systems to achieve "**appropriate** levels of performance

predictability, interpretability, correctability, security and cybersecurity assessed through **appropriate** methods" (article 32) (emphasis added).

Appropriate and adequate, along with reasonable, are every lawyer's favourite adjectives, as they can mean literally anything. These flexibility clauses are very welcome to create an agile regulation that does not stifle innovation. However, without an effective mechanism to specify these qualifiers through technical standards[10] or administrative regulation, flexibility turns into legal uncertainty. The opposite of what regulation should bring.

The specification of these elements will require considerable technical skills and is key for the functioning of the AI framework. It is not a coincidence that the European AI Act delegates the specification of such technical, yet vital issues, to standardisation bodies[11], a solution that has raised concerns from human rights advocates (Ada Lovelace Institute, 2023), but is completely understandable considering the level of technicality that the standardisation of such issues require.

To solve the implementation issue, the Bill proposes to establish an AI Governance and Regulation System, where all sectoral regulators should come together under the leadership of the ANPD. The idea of a coordination system is promising, but the Bill fails to define how it will function in practice. Particularly, it seems a risky gamble to entrust the leadership of the system to the ANPD, considering that is a severely overstretched organ that barely manages to cope with fulfilling its current mission.

Although AI regulation needs to deal with much more than data related-risks, it is understandable that the ANPD is looked to as the leader of such a system. However, to think that ANPD, in its current structure, can effectively lead a new system of such relevance seems

10 The ISO 27000 family of standard is particularly relevant in this regard. For a general overview of existing and under development relevant standards on cybersecurity and AI, see ENISA (2023).

11 According to recital 61 of the proposed AI Act "Standardisation should play a key role to provide technical solutions to providers to ensure compliance with this Regulation. Compliance with harmonised standards as defined in Regulation (EU) No 1025/2012 of the European Parliament and of the Council should be a means for providers to demonstrate conformity with the requirements of this Regulation. However, the Commission could adopt common technical specifications in areas where no harmonised standards exist or where they are insufficient." In this respect, in December 2022, the EU Commission adopted the "Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence." See https://ec.europa.eu/docsroom/documents/52376?locale=en.

overly optimistic. The structure of the Authority should be substantially reformed to have even a minimal chance to successfully coordinate the new AI system.

2.4 Conclusions

As exposed, the relationship between AI and cybersecurity unleashes significant and transformative developments. While it has empowered malicious actors to conduct more effective, far-reaching, and precise attacks, it has also underscored the importance of proactive and adaptive cybersecurity strategies. Indeed, the integration of AI into cyber offensive and defensive capabilities demands a fundamental shift in cybersecurity strategies.

In this context, fostering collaboration between government entities, private sector organisations, and research institutions, becomes essential for Brazil — or any other state — to address the challenges posed by AI in the cybersecurity domain. The adoption of a multistakeholder approach is essential to understand the cyberthreats scenario, develop effective regulations, standards, and governance mechanisms. Indeed, these elements are key to implement robust cybersecurity measures, and promote innovation in defensive AI technologies to safeguard the nation's critical infrastructure and protect its citizens from AI-driven cyberattacks.

However, given the considerations presented in the preceding sections, the current Brazilian institutional arrangement does not seem be fit to provide an effective governance system able to cope with existing cyberthreats, despite the relevant advancements of the country over the most recent years. It seems particularly important that a multistakeholder approach is enshrined in the future strategic and institutional approach adopted by Brazil, not only to increase the quality of policymaking and support it with well-crafted standardisation but, chiefly, to increase the inter-stakeholder coordination and implementation of cybersecurity measures.

Concretely, multistakeholder cooperation should be designed through the development of a "Brazilian Cybersecurity and Digital Transformation System", aimed at facilitating communication, cooperation and — ideally — coordination amongst all governmental entities with these issues.

This system should be moulded on the successful experiences of the Brazilian and National Consumer Protection System and the Brazilian Military Cyberdefence System (National Consumer Protection System, n.d.; Military Cyber Defense System, 2020). Ideally such system should be couple with a National Cybersecurity Network facilitating the participation all stakeholders and both the System and the Network should be headed by a much-needed National Cybersecurity Agency, able to act as a focal point for cybersecurity governance and regulation (Belli et al., 2023).

Hopefully, the aforementioned recommendations will be enshrined in the upcoming proposals on these matters to be issued by the Brazilian Presidency' CNCiber.

2.5 References

Ada Lovelace Institute. Inclusive AI governance. Civil society participation in standards development. Discussion paper. (March 2023).

ANPD. Orientation Guide on Information Security for Small-Scale Data Processing Agents. (October 2021). https://www.gov.br/anpd/pt-br/documentos-e-publicacoes/guia-vf.pdf.

ANPD. Resolution No. 15/2024. Approves the Security Incident Communication Regulation. (26 April 2024). https://www.in.gov.br/en/web/dou/-/resolucao-cd/anpd-n-15-de-24-de-abril-de-2024-556243024.

Belli et al. Cibersegurança: uma visão sistêmica rumo a uma proposta de Marco Regulatório para um Brasil digitalmente soberano. FGV. (2023). https://hdl.handle.net/10438/33784.

Belli L. "Cybersecurity Policymaking in the BRICS Countries: From Addressing National Priorities to Seeking International Cooperation" (2021) The African Journal of Information and Communication (AJIC), (28). doi:10.23962/10539/32208.

Belli, L. et al. Governança de dados no setor público: dados abertos, proteção de dados pessoais e segurança da informação para uma transformação digital sustentável. Lumen Juris. (May 2024). https://hdl.handle.net/10438/35341.

Brazilian Presidency. GSI Ordonnance 6/2024/ https://www.in.gov.br/en/web/dou/-/portaria-n-6-de-9-de-fevereiro-de-2024-542752145.

Brazilian Presidency. Presidential Decree 11.856/2023 https://www.in.gov.br/en/web/dou/-/decreto-n-11.856-de-26-de-dezembro-de-2023-533845289#wrapper.

Cen, H.; and Magramo, K. Finance worker pays out $25 million after video call with deepfake chief financial officer. CNN. (4 February 2024). https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

CyberBRICS. Professor Luca Belli appointed member of the new Brazilian Cybersecurity Committee. (16 February 2024). https://cyberbrics.info/professor-luca-belli-appointed-member-of-the-new-brazilian-cybersecurity-committee/.

ENISA. AI Cybersecurity Challenges Threat Landscape for Artificial Intelligence. (2020). https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

ENISA. Cybersecurity and AI Standardisation. (March 2023). https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation.

Gartner. 2Q24 Emerging Risks Report. https://www.gartner.com/en/documents/5529395.

Geluvaraj, B.; P. M Satwik; and T.A. Ashok Kumar. The future of cybersecurity: Major role of artificial intelligence, machine learning, and deep learnin in cyberspace. International Conference on Computer Networks and Communication Technologies: ICCNCT 2018. Springer. (2019).

Hassan, S.M., Wasim, J.: Study of artificial intelligence in cyber security and the emerging threat of ai-driven cyberattacks and challenges. J. Aeronaut. Mater. 43(1), 1557–1570. (2023).

ITU-T. (2009). Recommendation X.1205 (04/08): Overview of cybersecurity. Approved in 2008-04-18. https://www.itu.int/rec/T-REC-X.1205-200804-I.

Malatji, M.; A. Tolah. Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. AI and Ethics (2024).

Military Cyber Defense System (SMDC). Ordinance No 3.781/GM-MD. (17 November 2020). https://www.in.gov.br/web/dou/-/portaria-n-3.781/gm-md-de-17-de-novembro-de-2020-289248860.

MIT Technology Review Insights. Preparing for AI-enabled cyberattacks. (8 April 2021). https://www.technologyreview.com/2021/04/08/1021696/preparing-for-ai-enabled-cyberattacks/.

Nakamura J. Brasil é vice-campeão em ataques cibernéticos, com 1.379 golpes por minuto, aponta estudo. CNN Brasil. (30 October 2024).

National Consumer Protection System (Sistema Nacional de Defesa do Consumidor — SNDC) https://www.consumidor.gov.br/pages/conteudo/publico/6 ;.

Pupillo, L. et al. Artificial Intelligence and Cybersecurity Technology, Governance and Policy Challenges Final Report of a CEPS Task Force. Centre for European Policy Studies (CEPS) Brussels. (May 2021). https://www.ceps.eu/wp-content/uploads/2021/05/CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf.

Russell, S.; Perset, K.; Grobelnik, M. Updates to the OECD's definition of an AI system explained.OECD.AI Policy Observatory. (29 November 2023). https://oecd.ai/en/wonk/ai-system-definition-update.

Safitra, M.F.; Lubis, M.; Fakhrurroja, H. Counterattacking cyber threats: a framework for the future of cybersecurity. Sustainability. (2023). https://doi.org/10.3390/su151813369.

Salomão, L. F. Artificial Intelligence: technology applied to conflict management within the Brazilian judiciary. FGV. (2022). https://hdl.handle.net/10438/33954 ; L. Belli et al. Courting AI: How Brazilian Courts are Using and Regulating AI. in A. Limante and M. Zalnieriute. Cambridge Handbook of Courts and AI. (forthcoming).

3 The law on artificial intelligence (AI) in South Africa in the evolving African legal landscape

Sizwe Snail Ka Mtuze, Masego Morige and Mbali Nzimande

Abstract

This research article will look at the status of AI Laws and Policies in South Africa, the South African Draft AI Strategy (SADAIS) and critique thereof, South Africa's National Artificial Intelligence Policy Framework (NAIPF) as well as African initiatives to regulate Artificial Intelligence as it evolves. The article concludes with a thought on how South Africa is dealing with Artificial Intelligence and the scramble to publish Policy Frameworks to govern it.

Keywords: African Intelligence (AI), South Africa, AI Strategy, AI Policy, African Framework on AI.

Introduction

The regulation of Artificial Intelligence in South Africa has been a contentious and ongoing issue discussed by various African Scholars. (Adeyoju: 2018:1, Ncube, et al:2023 & Snail & Morige:2024). This research article will look at the status of AI Laws and Policies in South Africa, the South African Draft AI strategy and critique thereof, South Africa's National Artificial Intelligence Policy Framework as well as African initiatives to regulate Artificial Intelligence. South Africa does not have a formal AI Policy and it is because of this that mention has been made to the various pieces of legislation. These pieces of legislation are currently being used to govern AI in the absence of the policy. This article will then conclude with a thought on how South Africa is navigating the evolution of AI.

3.1 The Status of AI In South African Law

AI has had an impact on a variety of industries in the modern world and the legal profession is no exception. The 4th Industrial

Revolution has had a negligible effect on the legal profession and this is as a result of the immunity that it enjoys when compared to other professions. This immunity is protected by professional rules, guidelines and ethics. (Adeyoju: 2018: 2-3) However, it seems that this immunity will not be functional for much longer because a majority of the protections are being eroded as the laws on AI evolve (Adeyoju: 2018: 2-3). AI has percolated into the profession and the need has arisen for there be legislation that will regulate and guide its ethical use. AI has emphatically made its presence known and there has been an escalation in the need for it to be increased rapidly.

The issue we face in South Africa is that there is currently very little existing legislation, regulatory mechanisms or policies that will do this. (Adams: 2021:13 & Brand: 2022:142). For those who argue that there is such existing legislation, they have neglected to note that it may only be applied in a general sense and that its relevance to AI is limited. What is lacking is specific regulatory frameworks and policies governing how we use AI in our country. (Adeyoju: 2018:2-3).

3.1.1 PC4IR Report

Following the development of AI and the lack of legislation designated for AI-related matters, the South African President in 2019, initiated the Presidential Commission on Fourth Industrial Revolution Commission[12] (which then issued the Presidential Commission on Fourth Industrial Revolution Commission (PC4IR Report)[13] which came up with eight key recommendations, including the establishment of an artificial intelligence (AI) institute and the review and amendment (or creation) of policy and legislation. The PC4IR Report put forward key points which are the pillars of AI's development in South Africa. (SAAIP: 2024:21). According to the South African government setting up of the AI Institute is a summary of all of the intended actions of the government in ensuring the smooth transition of AI into both

12 Department of Telecommunications and Postal Services, Terms of Reference for the Presidential Commission on the Fourth Industrial Revolution GN 209 in GG 42388 of 2019-04-09 (https://www.gov.za/sites/default/files/gcis_document/201904/42388gen209.pdf).

13 Commission on the Fourth Industrial Revolution, Summary Report & Recommendations GN 591 in GG 43834 of 2020-10-23 (https://www.gov.za/sites/default/files/gcis_document/202010/43834gen591.pdf)..

the public and private sector and how it will enhance the already existing skills and research.

3.2 South Africa's Draft AI Strategy & Critique

The South African government has put together a discussion document named the South Africa's Draft AI Strategy (SADAIS) (SAAIP: 2024:25). It discusses its priorities, intentions and objectives for the adoption of AI into South Africa's various sectors and to bring about the envisioned economic advancement (SAAIP:2024:3). The discussion document is divided into 3 (three) sections and each section touches on a different aspect. The plan by the government is to facilitate a better use of AI in the future through a variety of measures. These measures are the,

> *"creation of policy and regulatory experiments; set of positive goals for what South African society require from AI; building an understanding of the AI technological possibilities; management of negative AI impacts on society and industry and providing certainty to society on this rapidly evolving AI technology through flexibility and accommodation of skills, software, innovations and applications". (SAAIP: 2024:8).*

In order for the SADAIS to succeed, there are 8 (eight) pillars on which it will rely. These pillars are envisaged to ensure that all sectors are accounted for in this transitory period. The most important pillar is the one that speaks to the need for there to be separate legislation which will highlight that AI is important and that its field of technology is equally important (SAAIP: 2024:8). Another equally important pillar is one that touches on the belief that South Africa truly has the potential to be valuable and bring about positive change (SAAIP: 2024:22).

With regard to the actual adoption of AI, the PC4IR report states the terms and conditions of the approach which must be taken (SAAIP:2024:15). The approach must be one that is inclusive, integrated, adaptive and mindful of the socio-economic impact (SAAIP:2024:24). Those who find themselves tasked with regulating should concentrate on how they are going to overcome the hurdle

of the lack of exploration of AI in laws and regulation (SAAIP: 2024:24). As a result, those tasked with making policy should focus on building trust among people in AI-driven systems. This trust can be built through the development of intelligible frameworks and clear attributions of accountability.

The SADAIS consists of 4 (four) phases and they span from the year 2023 to 2026. Phase 0 (zero) is scheduled for 2023 and the plan is that strategy formulation takes place and strategies are developed (SAAIP: 2024:29). Phase 1 (one) is scheduled for 2024 and the plan is to activate initiatives, test them and assess their effectiveness and efficiency (SAAIP:2024:23).

Phase 2 (two) is scheduled for 2025 and the plan is to expand execution through the activation of more institutes in order to achieve strategic objectives. Phase 3 (three) is scheduled for 2026 and is the finale where all the existing initiatives are accelerated to a national level (SAAIP: 2024:23). Technology can be used as a tool of choice and it will have an impact on two sectors, namely the social and economic sectors. It aims to achieve the following 4 (four) outcomes: AI Predictive maintenance abilities, AI Logistics optimization and Automated services, AI Diagnostic abilities and AI Analytical abilities (SAAIP:2024:23).

AI also has benefits that will benefit the country and accelerate the transition into the new normal of an AI-driven modern country using the four-phase (strategy formulation, activate initiatives, expand execution and accelerate execution) plan (SAAIP:2024:35).

3.2.1 Critique: The Good

The SADAIS was published by the Department of Communications and Digital Technologies (DCDT) in April of 2024 at a time when Africa is undergoing a comprehensive review of AI policies and laws. The purpose of the document was to commence talks and strategizing between the public and private sector. (SAAIP: 2024:35). These talks were initiated with the aim of facilitating AI innovation, government-led AI initiatives, regulatory frameworks and principles and ultimately, the development of a national AI policy (Bhagattjee:2024). It was a working paper and it was a step in the

right direction with regard to the regulation of AI. (Bhagattjee:2024) If it happens that it is adopted as a White paper, it will prove useful as it is well compiled and would play a key role as a regulatory and governance tool. The SADAIS provided key proposals and insights into the government's approach. One of the key proposals was to ensure that any ethical considerations relating to AI are addressed appropriately under the legal framework to guard against any potential harm as an important component of the SADAIS is that it considers that the future use of AI could cause harm to humans and raise ethical concerns (Bhagattjee:2024).

As a result, this requires regulation on aspects such as the social risk of loss of employment, dangerous outcomes which would come with increased criminal behaviour, the risks that come with robotic or autonomous devices that are AI-centric and the risks posed by the potential detriment humanity faces from AI. The SADAIS advocated for AI literacy as it is needed by South Africa and it can be provided through education and training and by investing in technology start-ups (Bhagattjee:2024). The hope is that this Discussion Document is reworked and that it is then published with input from key stakeholders from both the private and public sector as well as the AI Expert Advisory Council and any other relevant AI bodies (Bhagattjee:2024).

When the document was launched, the Minister of DCDT alluded to the type of approach that government would take in its approach to regulating AI as it is not set out clearly in the document (Bhagattjee:2024). If one were to look for the most positive take-away from this document, it is that it takes into account how different jurisdictions around the world regulate AI and implement it using effective mechanisms to foster and encourage AI use and development. This is done whilst also striking a balance between risk-management, assessment of harms and a consideration of major and minor ethical risks as they are equally important (Bhagattjee:2024).

3.2.2 Critique: The Bad and Ugly

The SADAIS was a 53-page long document and it has a disclaimer which states that it is a discussion document. (Pierce:2024). Pierce is in agreement with the disclaimer and further voices that it lacks

clear deliverables and that is complicated and not up to standard. The Plan is lengthy, it contains a high volume of jargon and has a number of unfinished thoughts (Pierce:2024). In November of 2022, the Artificial Intelligence Institute of South Africa was set up and things seem to be progressing slowly as even its website has not yet been updated since March 2023 (Pierce:2024).

The issue that the slow progression poses is that the rollout of the AI plan is centred around this Institute and this could seriously delay things. Another critique is the unrealistic timetable that has been set for the adoption of AI (Pierce:2024). Other countries have given themselves much more time to adopt it whereas South Africa has set themselves 12-month deadlines to achieve the impossible (Pierce:2024). Pierce makes mention of Rwanda's National AI Policy and in comparison to South Africa's, he is of the opinion that it is a far more practical document that is uncluttered and has very little room for misinterpretation.

Therefore, Pierce recommendation was that the entire SADAIS is reworked and that it is released timeously as South Africa runs the risk of getting left behind while the rest of the world passes AI Acts and publishes practical policies (Pierce:2024). Seth Thorne has also given his views on the document and he shares sentiments which are similar to those of Pierce. Thorne shares an important view with Pierce which is that the draft touches on Data Sovereignty which will be managing the data that will be needed for the AI training however, it fails to address the challenges that will come about in securing the computing power necessary for the achievement of its objectives (Thorne:2024).

3.3 South Africa: National Artificial Intelligence Policy Framework

The South African National Artificial Intelligence Policy Framework (NAIPF) was drafted in August of 2024 and it can be considered to be the first step in the actual development of a National AI Policy. It follows the Draft AI Strategy Document (SAAIP: 2024:1). The NAIPF is intended to serve as the foundational basis for creating AI regulations and potentially an AI Act in South Africa, and guide the development of robust regulatory mechanisms that ensure that AI

applications are safe, ethical and in the public interest. (Rosenburg & Madondo: 2024:1). The rationale for the development of an AI policy document in South Africa was that it is imperative that there is a set of guidelines to ensure the responsible and ethical use of AI across all societal sectors. The rapid advancement of AI technologies offers opportunities for an enhanced quality of life, economic advancement and an improvement in public services (Rosenburg & Madondo:2024:3).

However, these opportunities are at risk of never being actualized due to the risks that are posed by letting AI develop without any policy to regulate it. Thus, having an AI policy will provide the foundation for AI to be regulated and for there to eventually be an AI Act (Rosenburg & Madondo:2024:3).

The NAIPF acknowledges global trends in AI governance and the need to harmonise with international standards, pushing South Africa to develop its own AI policies. It seeks to align with international norms and standards to ensure ethical and effective AI deployment. (Rosenburg & Madondo: 2024:5). The NAIPF has 12 (twelve) fundamental components which can be characterized as the support behind the implementation of the goals and objectives of the National AI Policy (Rosenburg & Madondo:2024:5). Talent and capacity development is one of the components and its aim is to ensure that South Africa has a robust AI talent pool. Digital infrastructure is also one and its aim is to create an environment which will foster AI innovation. Research, development and innovation is component number three and its aim is to be the driving force behind AI innovation and ensure the advancement of technological capabilities. Component number four is Public Sector Implementation and its aim is to use AI to enhance the efficiency of our government (SANAIPF:9).

Component number five is Ethical AI Guideline Development and they are there to ensure that the use of AI is ethical and responsible. Component six requires Privacy and Data Protection has the aim of safeguarding the personal information of all people who will be bound by the National AI policy (SANAIPF:9). Safety and Security is component number seven and its aim is to protect citizens and ensure that cybersecurity protocols are safeguarded by AI systems.

Transparency and Explainability have the aim of building trust amongst members of the public in component number eight. If the National AI Policy provides clear and understandable information on AI, it will be easier for the public to understand the AI systems (SANAIPF:10). In order to ensure that AI is deployed equitably there must be Fairness and Mitigating Bias as per component nine (SANAIPF:10).

The importance of the Mitigating Bias is to ensure that it identifies any bias that might be present in the AI systems. Component number ten is Human Control of Technology and it is there to ensure that the AI systems have human oversight and to ensure the prioritisation of a human-cantered approach within the systems. Professional Responsibility is component number eleven and it creates a code of conduct for AI professionals and ensures the upholding of ethics as per component eleven. The final component is the Promotion of Cultural and Human Values and its aim is to ensure that the development of AI is aligned with societal values and promotes environmental sustainability and human well-being (SANAIPF:11). The abovementioned key pillars are crucial to the meaningful contribution of AI technologies to important sectors such as healthcare and education (SANAIPF:12).

The NAIPF outlines key pillars such as robust Data Governance Frameworks, Infrastructure Enhancement, and Significant Investments in Research and Innovation, which the DCDT believes are crucial components to create an enabling environment where AI technologies can thrive and contribute meaningfully to sectors such as healthcare, education and public administration (Rosenburg & Madondo:2024:6) Overall, the NAIFP seeks to lay the groundwork for South Africa to emerge as a leader in AI innovation while addressing challenges and opportunities in a holistic and sustainable manner (Rosenburg & Madondo:2024:6).

3.4 Overview of Regulation of Artificial Intelligence in Africa

AI is slowly making it to the meeting agendas of organisations globally including across Africa. Such that, there is an important piece of African International law namely the African Union Convention

on Cyber Security and Personal Data Protection[14]. It has limited AI regulatory properties and spearheads matters of data protection, cybercrime and cyber security in the African continent. Article 9 of the Convention regulates data processing and this is inclusive of the automated processing of personal information by AI. Article 14.5 of same confers rights on all data subjects that they may not be affected by legal effects that significantly affect them solely based on automated data processing (Orji et al:2024:172).

In the AU Digital Strategy Information for Africa for 2020-2030[15] a proposition was made in Kenya and makes extensive references to the governing of AI. The proposition is that there be a continent-wide digital governance African Peer Review Mechanism on AI use. It would be applicable to member states and it would prescribe rules on AI with a basis on solidarity and to ensure that Africa is cooperative with forthcoming digital infrastructure (Ncube et al:2023:69). One of the first African countries to establish a national policy and institutional framework to govern AI is Mauritius. Its national AI strategy was established in November of 2018 and its aim is to address any ethical concerns surrounding the development and use of AI as well as to promote capacity building (Orji et al:2024:69). As far as Africa is concerned as of March 2024. 9 (nine) out of the 54 (fifty four) states had established AI policy frameworks and only 2 (two) had already established institutional framework (Orji et al:2024:171).

The Continental AI Strategy calls for unified national approaches among AU Member States to navigate the complexities of AI-driven change, aiming to strengthen regional and global cooperation and position Africa as a leader in inclusive and responsible AI development.[16] The AU AI Strategy contains a key action point and that is the building of a AI knowledge base speaking on AI use cases and the monitoring of the implementation of the recommendations of the Strategy (Alayande & Adams:2024:3).

14 African Union Convention on Cyber Security and Personal Data Protection (2014) (https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection).

15 The Digital Transformation Strategy for Africa (2020-2030) (https://au.int/en/documents/20200518/digital-transformation-strategy-africa-2020-2030).

16 Ibid.

3.5 Conclusion

What we can conclude from all that has been said is that South Africa does not seem adequately prepared to deal with the multifaceted and evolving AI. There is an evident lack of regulatory policies, undefined laws, critical judgements having been handed down and 53-page discussion documents which do not have clear directives and implementation procedures. It also seems that the South African government has seem the deficiencies in the SADAIS hence it has rushed within months to develop the NAIPF which has been received less critically that the previous SADAIS.

3.6 References

Adams, N., 2024 Parker v Forsyth no lessons for using ai for legal-research. Available at: https://www.michalsons.com/blog/parker-v-forsyth-no-lessons-for-using-ai-for-legal-research/66884.

Adeyoju, A. (2018) *Artificial Intelligence and the Future of Law in South Africa*.

Alayande, A., Adams, R., (2024) "Africa Now Has a Continental AI Strategy: What Next?" August.

Bhagattjee, P. and Stephens, S (2024) The AI National Policy: South Africa's initial step to establish an AI policy and regulatory framework. Available at: https://www.werkmans.com/legal-updates-and-opinions/the-ai-national-policy-south-africas-initial-step-to-establish-an-ai-policy-and-regulatory-framework/.

Brand (2022) Responsible Artificial Intelligence in Government: Development of a Legal Framework for South Africa "14(1) in JeDEM.

Pierce, L. South Africa's Draft AI Plan: Not Good Enough (2024) https://www.linkedin.com/pulse/south-africas-draft-national-ai-plan-good-enough-lucien-pierce-gp5cc.

Rosenburg, W and Madondo, N (2024) The National AI Policy Framework: A step closer to aligning with international trends (https://www.werksmans.com/legal-updates-and-opinions/the-national-ai-policy-framework-a-step-closer-to-aligning-with-international-trends/).

Snail Ka Mtuze, S., Morige, M. (2024) Towards Drafting Artificial Intelligence (AI) Legislation In South Africa in Obiter.

Thorne, S. (2024) South Africa's proposed AI plan needs a rework: experts (https://businesstech.co.za/news/government/768147/south-africas-proposed-ai-plan-needs-a-rework-experts/.).

Adams,N (2021) South African Company Law in the Fourth Industrial Revolution: Does Artificial Intelligence Create a Need for Legal Reform? (LLM thesis, Wits University).

Marwala, T. and Mpedi, L.G., 2024. Artificial Intelligence and the Law. (Palgrave).

Ncube,C., Oriakhogba,D. Rutenberg, Schonwetter, T. (2023) Artificial Intelligence and the Law in Africa (Lexis Nexis).

Orji, U. Regionalising the Governance of AI in Andersen, L.H., Broeders, D. and Csernatoni, R., (2024). "Emerging and disruptive digital technologies: National, regional, and global perspectives".

African Union (2024) Continental Artificial Intelligence Strategy (https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy).

African Union Continental Artificial Intelligence Strategy (https://au.int/sites/default/files/documents/44004-doc-EN_Continental_AI_Strategy_July_2024.pdf).

Commission on the Fourth Industrial Revolution, Summary Report & Recommendations GN 591 in GG 43834 of 2020-10-23 (https://www.gov.za/sites/default/ files/gcis_document/202010/43834gen591.pdf) Department of Telecommunications and Postal Services, Terms of Reference for the Presidential Commission on the Fourth Industrial Revolution GN 209 in GG 42388 of 2019-04-09 (https://www.gov.za/sites/default/files/gcis_document/201904/42388gen209.pdf).

DCDT, South Africa National Artificial Intelligence Policy Framework (https://www.policyvault.africa/policy/south-africa-national-artificial-intelligence-ai-policy-framework-2024/)Electronic Communication and Transactions Act 25 of 2002.

Smart Africa (2021) Artificial Intelligence in Africa (https://smartafrica.org/knowledge/artificial-intelligence-for-africa/).

South Africa's Artificial Intelligence (AI) Planning: Adoption of AI By The Government (https://www.dcdt.gov.za/images/phocadownload/AI_Government_Summit/National_AI_Government_Summit_Discussion_Document.pdf).

The Digital Transformation Strategy for Africa (2020-2030) (https://au.int/en/documents/20200518/digital-transformation-strategy-africa-2020-2030).

4 Building Smart Courts Trough Large Legal Language Models? Experience from China

Zijing Liu, Shaoyu Liu and Yin Lin

Abstract

Artificial intelligence is already all around us and has been applied to almost every aspect of society and business. One of the most striking innovations in the application of AI has been the introduction of large legal language models in judicial decision-making. There has been growing interest in the use of AI in legal systems worldwide in recent years, particularly in the role of judges. The main question addressed in this Article is that what should be the potential legitimacy, weaknesses, and limitations of large legal language models in judicial scenery. To address it, it takes the Chinese smart court construction as an example and studies 133 cases of smart courts from 2017 to 2024. It summarizes four patterns from Shanghai city, Zhejiang province, Jiangsu province, and Shenzhen city. Based on this, this article analyzes the achievements and shortcomings of the application of large language models in China's smart courts.

Keywords: Artificial intelligence, large legal language models, smart courts, Judicial decision-making, empirical study.

Introduction

Technological innovations such as big data, cloud computing, and artificial intelligence have created a worldwide digital revolution regarded as 'the Fourth Industrial Revolution' which impacts everyone's life (Klaus Schwab, 2016). Already, artificial intelligence is all around us and has been applied to almost every aspect of society and business, from assigning credit scores to assessing the criminal risk of people (Antunes H. S. et al., 2024, p.281). One of the most striking innovations in the application of AI in the justice system in recent years has been the introduction of large legal language models in judicial decision-making and other assistance jobs (Bin Wei, 2024).

In recent years, there has been growing interest in the use of AI in legal systems, particularly in the role of judges (Ulenaers, 2020). In 2023, a Colombian judge used the AI chatbot ChatGPT in preparing a ruling in a children's medical rights case by asking the chatbot whether an autistic child's medical insurance should cover the cost of related therapies (Luke Taylor, 2023). Later this year, an intellectual property law Judge in England used ChatGPT to assist judicial decision-making, such as summarizing information on the law in a particular field (Gareth Corfield, 2023). Compared to Colombia and England, East Asian countries such as India and China use the large language model in the judiciary sector more aggressively. The Indian Supreme Court has set up an AI Committee with a focus on the translation of legal documents; process automation; increasing administrative effectiveness; automating forecasting, prediction, and filing; scheduling of cases; and early case resolution using chatbots (Gandhi & Talwar, 2023). The Chinese government is even more ambitious. It issued a '*New Generation Artificial Intelligence Development Plan*' in July 2017, advocating to establishment of an AI-powered 'Smart Court'.[17] In Oct 2024, Zhang Jun, president of the Supreme People's Court, stressed the need to explore the use of artificial intelligence technology to empower the judiciary and promote the deep integration of artificial intelligence and judicial work (Zhang Jun 2024).

The implementation of large legal language models in judgment is controversial. Despite the potential advantages of robot judges, it raises significant concerns, such as the concerns of algorithm bias, non-transparency, inaccuracy, weak interpretability, hallucinations, lack of human empathy, data privacy, and security problems (Magnus Kristoffersson, 2024). Consequently, scholars around the world have highlighted that the application of generative AI, particularly the large legal language model, cannot be a substitute for human judges (Parikh et al., 2023). The main question addressed in this Article is, thus, what should be the potential legitimacy, weaknesses, and limitations of large legal language models in judicial scenery. To address it, it takes the Chinese smart court construction as an example and studies 133 cases of smart courts from 2017 to 2024. It summarizes four patterns from

17 State of Council of People's Republic China, *Notice of the State Council on Issuing a New Generation Artificial Intelligence Development Plan*, 2017, https://www.gov.cn/zhengce/zhengceku/2017-07/20/content_5211996.htm.

Shanghai city, Zhejiang province, Jiangsu province, and Shenzhen city. Based on this, this article analyzes the achievements and shortcomings of the application of large legal language models in China's smart courts.

4.1 Methods

4.1.1 Legal empirical analysis

Data analysis

This article collects 133 case samples of smart court construction from 55 regions in China from 2017 to 2024. It analyses specific cases utilizing the large legal language model and discusses the experiences, achievements, and shortcomings of China's foundation model construction. The primary textual source is '*the China Court Informatization Development Evaluation Report*' conducted by the Chinese Academy of Social Sciences. This report is published annually since 2017, making it the most authoritative and systematic public resource for the construction of smart courts in China (Tian He, 2024).

Face-to-face interviews

In addition, this article also uses questionnaires and interviews to conduct in-depth interviews with judges who use AI large language models to understand the current status and potential problems of the smart court.

Normative analysis

Normative analysis method is a unique method of jurisprudence. It mainly focuses on the legality of law, the operation effect of law, the substantive content of law, and examines the constituent elements of law in an all-round way.

4.2 Discussion

4.2.1 Background: the motivation of China's smart court construction

Through the holistic approach of 'Smart Court' construction, China has significantly advanced the application of the foundation model in the field of adjudication, which is closely related to the functional

requirements and inherent challenges faced by courts: Firstly, the contradiction between the increasing caseload and limited judicial personnel has intensified, leading to inefficiencies in the judicial process. In 2015, to address the issue of 'difficulty in filing cases,' Chinese courts initiated reforms to the case registration system, resulting in a surge of disputes entering the courts and placing immense pressure on their adjudication capacity. In 2022, the average number of cases concluded per judge in grassroots courts reached 274, with some exceeding 400, yet the number of judges nationwide has not increased significantly over the past decade, further exacerbating the case-to-judge imbalance. Secondly, judicial fairness needs improvement. Despite hierarchical trial supervision mechanisms such as second-instance final judgments and retrials, inadequate case quality inspection mechanisms have led to repeated instances of inconsistent judgments for similar cases and misjudgements. Thirdly, judicial credibility remains insufficient, with 'visible justice' not fully achieved, and public trust in the judiciary requires further enhancement (Jia Yu, 2024).

4.2.2 Four patterns: utilization of large legal language models in Chinese smart court

The construction of smart courts encompasses a comprehensive process that integrates informatization, datafication, and intelligence across various stages such as case filing, trial, supervision, and management. Among these, the trial phase prominently showcases the application and technological characteristics of the foundation model, specifically including (Wei Bin, 2022):

- Similar Case Recommendation. It involves retrieving and presenting similar cases and their corresponding judgments based on the current case being heard.
- Legal Judgment Prediction. This entails extracting key information from judgments, categorizing it, and utilizing text classification techniques to forecast the outcome of the current case, including charges, applicable laws, and sentences.
- Automated Generation of Legal Documents. It involves constructing a knowledge graph of the case based on trial data and legal knowledge and then employing machine learning

and natural language generation techniques to automate the generation and proofreading of legal documents.

Currently, the development of foundation models is primarily driven by local pilot projects, resulting in four distinct patterns as follows.

4.2.2.1 Shanghai Model

In 2017, the Shanghai Higher People's Court introduced the 'Intelligent Assistant System for Criminal Cases' (206 System), which leverages AI for evidence analysis, unifying evidence standards, formulating evidence rules, and constructing evidence models. This aims to achieve the judicial goals of uniform law application and prevention of miscarriages of justice. The 206 System employs new AI technologies such as optical character recognition, natural language processing, intelligent speech recognition, element extraction, and machine learning to provide guidance for case handlers in collecting and fixing evidence, enabling judgment, verification, control, and supervision of evidence. Through this system, flaws and contradictions in evidence can be promptly identified and flagged for case handlers, thereby preventing miscarriages of justice (Cui Yadong, 2020).

4.2.2.2 Zhejiang Model

Led by the Zhejiang Higher People's Court, the Full-process Intelligent Trial System (FITS) 'Xiaozhi' was developed, capable of tasks like legal information extraction, evidence classification, question generation, dialogue summarization, judgment prediction, and judgment document generation (Yu Shujun, 2019). The system first extracts elements from legal texts to assist judges in effectively identifying the essence of cases. It then verifies the consistency of all evidence to demonstrate its validity. Additionally, it features an automatic questioning robot that assists judges in posing questions during trials, both procedural and factual. The system can also summarize points of contention during court debates under a multi-task learning framework, generating real-time trial records automatically. Lastly, it proposes a natural language generation method based on attention and counterfactual reasoning to produce court judgments. Currently, the system can assist judges in handling specific types of simple cases such as financial loan contracts, private lending, motor vehicle accidents, theft, and divorce, enhancing case handling efficiency.

4.2.2.3 Suzhou Model

In Jiangsu, the Suzhou Intermediate People's Court has also piloted a generative AI-assisted case-handling system. Building upon electronic case file data and legal knowledge data accumulated from previous paperless case-handling initiatives, the court has integrated the 'General AI foundation model' technology to create a specialized prophecy model tailored for courts, boasting multi-modal document comprehension, legal semantic cognition, and natural language interaction capabilities. This AI-powered system can accurately identify and present factual elements required by judges within electronic case files, including their original sources. Its built-in annotation and element backfilling functions facilitate judges in reviewing case files, retrieving evidence, and organizing facts. Furthermore, the system can mimic judicial thinking to organize language and generate relevant legal documents, with accuracy rates exceeding 95% for party information and 'fact finding' sections, and around 70% completeness for reference 'judgments' (Suzhou Intermediate People's Court, 2023).

4.2.2.4 Shenzhen Model

On June 28, 2024, the Shenzhen Intermediate People's Court launched an AI-assisted trial system that supports judges throughout 28 critical nodes and 57 auxiliary nodes, from case filing to closure. During case review, the system enables precise data tracing and comparison, facilitating refined and user-friendly information processing. During trials, it assists in real-time in evidence verification and logical review, enhancing trial quality and efficiency. During judgment, it matches similar cases, applicable laws, and authoritative viewpoints to ensure uniformity in judgment standards. Additionally, it innovatively employs a foundation model tree-structured prompt engineering component to manage judgment standards. Lastly, the system incorporates a self-learning and feedback mechanism, dynamically optimizing based on judges' usage and actual judgment outcomes (Guangdong Higher People's Court, 2024).

4.2.3 Achievements and Problems

Although China has made certain achievements in the development of law and artificial intelligence, there are still numerous issues, mainly

manifested in three aspects: technical issues, issues of justice, and institutional issues.

4.2.3.1 Technical Concerns

The large legal language model has exposed problems such as weak interpretability and the generation of false content due to 'hallucinations' in the judicial field. Firstly, the foundation model's use of neural network algorithms leads to the 'black box' problem in algorithmic decision-making, thereby rendering the process and results of legal predictions lacking in transparency and interpretability (Wei Bin, 2024 b). Secondly, foundation models still suffer from data 'hallucinations' that may compromise the accuracy of results. Thirdly, judicial artificial intelligence systems are still primarily expert systems, and constructing expert graphs requires extensive manual annotation and organization, which might paradoxically increase judges' workload. Currently, China's foundational model is still in its infancy, with notable flaws that prevent it from replacing legal professionals and relegating it to an auxiliary role. In tasks such as legal prediction, the foundation model still struggles to handle the core work of legal professionals, including legal reasoning, legal argumentation, judicial proof, legal interpretation, and judgment of complex cases.

4.2.3.2 Justice and Ethics Concerns

The existence of trial assistance systems can easily make judges susceptible to flawed preconceived judgments, leading to psychological anchoring effects and potentially even being 'monitored' and 'hijacked' by artificial intelligence, thereby affecting judges' discretion. Moreover, if paperless, visualized, and integrated artificial intelligence systems are fully implemented in courts of different levels in the future, they might undermine judicial independence.

4.2.3.3 Institutional concerns

Judicial artificial intelligence is significantly constrained by local fiscal capacity and regional economic development levels, resulting in significant disparities in development between regions. Taking the construction of the Guangzhou Internet Court as an example, just the first phase requires a budget of 15.59 million yuan (Zhou Xiang, 2021). Additionally, the development of judicial artificial

intelligence systems cannot be achieved without the support of technology companies. Regions like Beijing, Hangzhou, Shenzhen, and Shanghai enjoy distinct geographical advantages, which will lead to a fragmented market landscape.

4.3 Conclusion

This article investigated the use of large legal language model in China's smart court construction as it exists today regarding their skill to solve legal problems. The conclusion based in this is that large legal language model such as ChatGPT or other AI chatbot can be used as robot judges in the judicial decision-making, and that the future is already coming. The large legal language model is used not only in the judicial decision-making, but also in the public legal service, which needs a further discussion (Dai Xin, 2024). Besides, the 'robot lawyer' as well as 'robot judge' calls for more empirical studies.

4.4 References

Schwab, K. (2016, Jan). *The Fourth Industrial Revolution: What It Means and How to Respond*, World Economic Forum, Website. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

Antunes, H. S. et al. (2024). *Multidisciplinary Perspective on Artificial Intelligence and the Law*. Springer. https://doi.org/10.1007/978-3-031-41264-6.

Wei, B. (2024). Judicial Application and Specification of Large Language Model, *Oriental Law*, vol.5, pp.57-73. DOI:10.19404/j.cnki.dffx.2024.05.012.

Ulenaers, J. (2020). The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge? *Asian Journal of Law and Economics*, 11(2). p. 2, https:// doi.org/10.1515/ajle-2020-0008.

Taylor, L. (2023, Feb. 2), *Colombian Judge Uses ChatGPT in Ruling on Child's Medical Rights Case*, CBS News. https://www.cbsnews.com/news/colombian-judge-uses-chatgpt-in-ruling-on-childs-medical-rights-case/.

Corfield, G. (2023, Sept. 14). *British judge uses 'jolly useful' ChatGPT to write ruling*. Telegraph.https://www.telegraph.co.uk/business/2023/09/14/british-judge-uses-jolly-useful-chatgpt-to-write-ruling/.

Gandhi, P. & Talwar, V. (2023). Artificial Intelligence and ChatGPT in the Legal Context, *Indian Journal of Medical Sciences*, vol. 75. pp. 1–2. doi:10.25259/IJMS_34_2023.

Kristoffersson, M. (2024). The Concept of Robot Judges Using Generative Artificial Intelligence and the Rule of Law. In: Rigmor Argren (ed.), *Rule of Law in a Transitional Spectrum*, pp. 369-388.

Parikh PM, Shah DM, Parikh KP. (2023). Garcia JJ. ChatGPT and a controversial medicolegal milestone. *Indian Journal of Medical Sciences*, 75(1).

Tian He.(2024). China 'Court Informatization Development Report (No.8, 2024), *Social Sciences Academic Press*.

Jia Y, (2024), On Digital Courts, *Chinese Journal of Law*, vol.46, No.4, pp.3-20.

Wei B, (2022). Legal Argumentation Analysis of the Interpretability Challenge in Judicial Artificial Intelligence. *Legal System and Social Development*, vol. 30, No.4, pp.76-92.

Cui Yadong. (2020). Application and Governance of Artificial Intelligence, *Reform of Public Administration*, vol.42, No.6.

Yu Shujun (2019, Sept. 26), *Zhejiang Court first 'Phoenix Financial smart court'*, http://zjnews.china.com.cn/yuanchuan/2019-09-24/189935.html.

Suzhou Intermediate People's Court. (2023, Nov 23). The 'Generative AI-Assisted Case Handling System' Approved for Provincial Court Pilot, *Jiangsu Legal Daily*, http://www.zjrmfy.suzhou.gov.cn/fypage/toContentPage/xwzx/82a07a488c3068a6018c37980cca0012.

Guangdong Higher People's Court.(2024, June 7). Shenzhen Intermediate People's Court's AI-Assisted Judicial System Officially Launched, https://news.southcn.com/node_d16fadb650/3eb7350386.shtml.

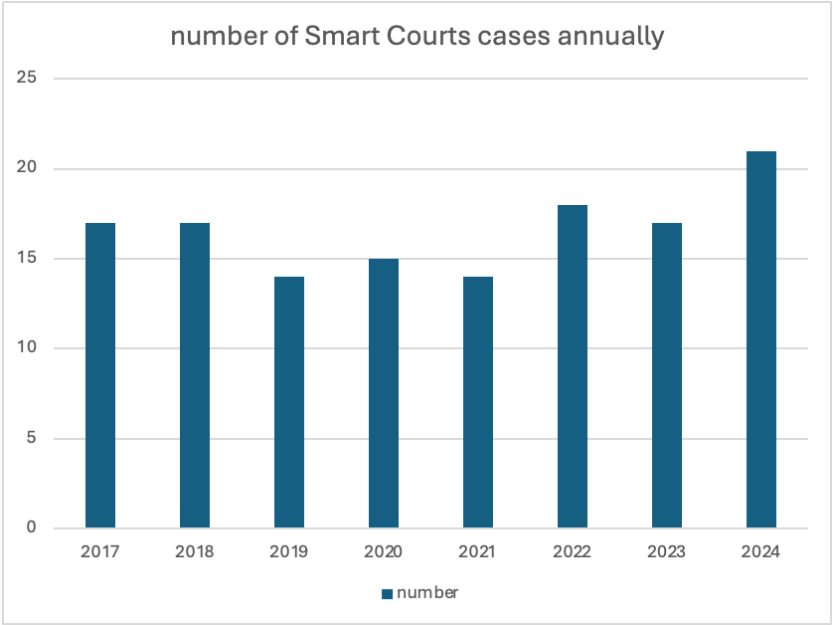
Wei B., (2024 b). Legal Argumentation Analysis of the Interpretability Challenge in Judicial Artificial Intelligence. *Legal System and Social Development*, vol. 30, no.4, pp.76-92.

Zhou X., (2021) The Formation Mechanism and Future Development Trends of Smart Courts. *Journal of Xi'an Jiaotong University (Social Sciences)*, vol. 4, no.3, pp.131-140.

Dai X., Who Wants a Robo-Lawyer Now?: On AI Chatbots in China's Public Legal Services Sector, *Yale Journal of Law & Technology*. Volume 26, Issue 3.pp.528-559.

Figure 1 Number of smart court cases annually

Note. Chinese Academy of Social Sciences makes reports concerning the smart court annually. The above table is the number of smart court samples each year from 2017 to 2024.



5 Fox Guarding the chickens — Bias in Risk Management Obligations for high-risk AI Systems under the EU AI Act

Nils Brinker and Richard Skalt

Abstract

In the context of the regulation of so-called high-risk AI applications by the EU AI Act, the obligation to conduct risk management plays a decisive role. In theory, manufacturers and operators of these systems must already mitigate the risks posed by their systems during the development phase. However, this paper argues that there is a fundamental bias on the part of manufacturers and operators, which threatens to result in third-party risks in particular not being adequately taken into account. It is also shown that the concretization mechanisms of the relatively abstractly formulated AIA play a critical role in ensuring that third-party risks receive appropriate attention.

Keywords: EU AI-Act, Risk Management, Principal Agent Relationship, Principle based Regulation.

Introduction

The European AI Act[18] has taken on the task of creating a regulation for AI as a technology that is still in development. For the category of "high-risk" AI systems in particular, a product safety regulation has been chosen that permits the development and marketing of such systems provided that certain requirements are met. A key aspect here is the implementation of risk management, which in theory should reduce the risks of these systems to an acceptable level. Since this risk management must be carried out by the manufacturer of the AI systems, this paper will address the question of whether such a methodology adequately takes into account the risks to third parties, i.e. to persons who are not the manufacturers, users or operators of these systems.

18 Regulation (EU) 2024/1689 (Artificial Intelligence Act) further referred to as AIA.

5.1 Discussion

5.1.1 The EU's Approach to AI Governance

The AI Act fundamentally follows a risk-based approach in multiple respects. On one hand, it categorizes various AI applications and imposes different levels of regulatory requirements depending on the category (Floridi et al., 2022; von Welser, 2024, p. 484 ff.).

A key regulatory focus of the AI Act is the formulation of obligations for so-called high-risk systems (Chapter 2 AIA). The operation of these systems is not inherently prohibited, but they must comply with a series of regulatory requirements and function as a product safety regulation (Rohrßen, 2024). Consequently, it is fair to assume that these requirements will have the most impact on the design of future systems available on the market.

The fundamental part of those obligations is the risk management laid down in Art. 9 AIA. Such risk management is regulated in Art. 9 and can be roughly summarized as a continuous, iterative process that identifies, evaluates and, where possible, mitigates risks. For market approval, the risks must be reduced to an "acceptable level". In this context, "known or reasonably foreseeable risks" to "health, safety or fundamental rights" that may arise from the use of the product for its intended purpose or from "foreseeable misuse" must be taken into account (von Welser, 2024).

In theory, risks to all stakeholders affected by an AI system must be considered. This includes not only the manufacturers, operators, and users of an AI system but also groups who are indirectly impacted by the system without having direct influence on its use. While risks to third parties are therefore theoretically acknowledged, the practical implementation of risk management under the AI Act may fall short in effectively addressing these risks, as it is discussed in the following sections.

5.1.2 Subjectivity in Risk Management

Risk management is not a precise, mathematical process that produces a deterministic outcome. There is always a certain degree of subjectivity on the part of the actor conducting the risk

management. This subjectivity influences both the identification of risks — whether they are even considered in the first place — as well as the evaluation of the likelihood of occurrence and the expected damage. This subjectivity particularly affects intangible risks, which are difficult to quantify using discrete categories such as numbers. (Ramnarine, 2015).

The responsibility for conducting risk management falls on the manufacturers, operators, or economic intermediaries laid down in Chapter 3 AIA. This makes sense to a certain extent, as these parties are capable of making concrete changes to an AI system and thus can operationally mitigate risks (Brinker, 2024). However, these actors naturally have vested interests in the design or functionality of the AI system, especially economic interests in bringing an AI system to market or using it in a particular form, which biases the risk management process.

5.1.3 Lack of specificity in risk management requirements

The AI Act is fundamentally designed as a "principle-based" regulation (Schuett, Anderljung, Carlier, Koessler, & Garfinkel, 2024). This high level of abstraction is also evident in the requirements for risk management. No specific methodological guidelines are provided, only eclectic requirements that a risk management process must fulfill.

The lack of specificity becomes critical, however, especially with regard to the types of risks that must be considered. Art. 9 (2) AIA refers to "known or reasonably foreseeable risks" to "safety, health, or fundamental rights." While the obligation to consider fundamental rights is, of course, commendable, there is a danger that this broad category will become a mere compliance checkbox to tick during the risk management process conducted by economic actors. Fundamental rights are universally valid, but due to the high level of abstraction, it is difficult (or nearly impossible) to derive concrete risk scenarios that must be considered for practical risk management. This lack of specificity means that there is a risk that the selection of risks taken into account will remain eclectic. If the manufacturer has no self-interest, there is a danger that third party risks are "forgotten". Yet even if there are no bad intentions involved, the manufacturers and operators are biased by their own subjective

perspective. It's in the nature of third-party risks, that they are not as obvious for others as they are to the parties directly involved.

Additionally, there may be a lack of methodological expertise on the part of manufacturers or operators of high-risk systems in identifying and evaluating risks to fundamental rights or third parties. While fundamental rights must be universally respected, the methods for identifying or weighing potential infringements are not trivial and are not universally mastered (Janssen, Seng Ah Lee, & Singh, 2022). Given that AI system manufacturers tend to be experts in technical domains, it is likely that they lack the necessary methodological tools or only have rudimentary knowledge of them.

It should be noted that a lack of methodological understanding is no excuse for failing to comply with legal requirements. In case of doubt, an entity is obligated to acquire the necessary methodological knowledge. However, in light of the inherent subjectivity of the risk management process, this is another factor that makes it unlikely that risk management will consistently produce the highest-quality outcomes. Instead, it will likely be conducted at the edge of what is just acceptable.

5.1.4 Principal-Agent Relationship in Risk Management

In essence, it is not the risk owner who decides how a risk affecting him is to be considered, evaluated and, in case of doubt, mitigated, but an actor with a certain vested interest. Since risk management likewise does not lead to an exact result, the actor who carries out the risk management can at least partially influence the result in the direction he desires. Risk management is thus not to be seen as a balancing of interests of all stakeholders involved, but as a means of ensuring minimum standards.

This relationship between legislators and AI manufacturers and operators can be understood through the lens of the principal-agent theory (e.g. Ross, 1973). In this framework, the legislators act as the principals, setting out the requirements and goals (such as safety and protection of fundamental rights, consideration of risks for third parties), while the manufacturers and operators are the agents tasked with implementing these requirements through risk management

processes. In principle, this arrangement functions effectively as long as the supervisory authorities, representing the principal, are diligent in their oversight duties (Hussein & Menon, 2003).

However, in practice, challenges arise when the agent's interests diverge from the principal's goals, particularly if the agents are primarily motivated by meeting only the "necessary minimum" requirements. This can lead to a "race to the bottom," where agents do just enough to comply with the law without fully embracing the spirit of the regulation. Such minimal compliance is difficult to counteract once it becomes the norm, even with subsequent legal adjudications or adjustments to the regulatory framework.

5.1.5 Concretization gone wrong

Although the AIA is "principle-oriented", it contains its own tools for concretizing its abstractly formulated requirements. In addition, a fundamental concretization can develop in practice through the application of law in court rulings, the action of the supervisory authority, or through generally developing conventions (such as public or private standards) (Schuett et al., 2024, p. 33 ff.). While the mechanisms mentioned in the previous sentence are rather indirect in nature, the mechanisms of the AI Act aim at a direct concretization. Accordingly, harmonized standards pursuant to Art. 40 AIA or common specifications pursuant to Art. 41 AIA can be adopted by the Commission by means of an implementing act.

However, neither direct nor indirect concretization takes place in a vacuum, but always against a material technical background. If manufacturers and suppliers have a vested interest in taking third-party risks into account at the smallest justifiable level, this minimal principle will also affect the design of their products. However, by defining the technical facts, they are setting the starting point for the discussion of further concretizations.

This applies in particular to the concretization through case law. The aim of case law is not to identify an optimal risk assessment, but merely to determine inadmissible interpretations. Thus, case law may at most shift the minimum threshold upwards.

However, concretizations through standards (including those officially defined by the instruments of the AI Act) are also influenced by existing technical possibilities, especially if they are already widespread.

To illustrate these effects, the evolution of the infamous "cookie consent" in the context of the GDPR can be used as a somewhat comparable scenario. Even the quite clearly formulated requirements for consent — it must, among other things, be given voluntarily, in an informed manner (i.e. the data subject must know exactly what he or she is consenting to) and unambiguously — led to a series of stylistic bloopers in the implementation of the website operators (who had a corresponding self-interest in ensuring that consent is given) (see e.g. Möller, 2022, p. 455). It took several years for what were actually obviously unlawful consent forms, such as the continued use of the website, pre-selected checkboxes, etc., to be addressed by courts (e.g. EuGH ECLI:EU:C:2019:801). And even eight years after the GDPR came into force, there appears to be little sign of effective enforcement. Even if the most obvious cases have been dealt with in court, the courts are still struggling with more subtle means of manipulation, such as dark patterns (Leiser & Santos, 2023) or simply misleading wording of the consent text. It would be naive to assume that the average internet user actually has an informed idea of what consent to "cookies" actually means.

In order to ensure that third-party risks are adequately taken into account when the AI Act is finalized, similar dynamics must be prevented and, above all, the "minimum standard" set by the manufacturers (Wehkamp, 2022) must not be used as the sole starting point for the discussion.

5.2 Conclusion

5.2.1 Make Third party risks known

It has been shown that third-party risks, for systematic reasons that lie primarily in the risk management carried out by the operators or manufacturers, tend to be given less consideration in risk management.

If this is to be avoided, the AI Act will have to be able to concretize the currently relatively abstract requirements, with the market surveillance authorities playing a central role here. This is especially true for formalized concretization processes through standards or common guidelines. When developing these, care should be taken to ensure that all stakeholders are able to contribute their input, and that civil society is not neglected in favour of the technical community. This is the only way to ensure that manufacturers and providers take due account of third-party risks that do not fall within their own sphere of interest. Although it may not be feasible to get them to do more than "work to rule" and always take the minimalist approach to risk management, it is important that third-party risks are also considered as part of the "work to rule" approach.

In this context, civil society actors have the particular role of making third-party risks generally known. Even manufacturers and operators who are willing to take into account all risks for third parties have a bias due to their subjective perspective and can thus overlook third-party risks. In order to have any chance of being considered, affected stakeholder groups or their representatives must therefore publicly draw attention to any "forgotten" negative effects on themselves or others.

5.2.2 Outlook

As explained, risk management for high-risk AI systems is not about balancing the interests of all stakeholders, but about ensuring minimum standards. This does not necessarily speak against the AI Act as a whole, as minimum standards do not necessarily lead to an optimal result for society as a whole, but are initially a step in the right direction. Nevertheless, the mistake must not be made to see the AI Act as a definitive part of AI regulation due to its generic designation, which takes into account all social impacts.

In order to achieve an appropriate consideration of third-party risks, it is particularly important to concretize the currently still very abstract provisions of the AI Act. In this context, it is important for civil society to draw attention to risks and for the authorities to adequately acknowledge them in the context of concretization.

5.3 References

Brinker, N. (2024). Identification and demarcation — A general definition and method to address information technology in European IT security law. *Computer Law & Security Review*, *52*, 105927. https://doi.org/10.1016/j.clsr.2023.105927.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). capAI — A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091.

Hussein, K., & Menon, A. (2003). The principal-agent approach and the study of the European Union: Promise unfulfilled? *Journal of European Public Policy*, *10*(1), 121–139. https://doi.org/10.1080/1350176032000046976.

Janssen, H., Seng Ah Lee, M., & Singh, J. (2022). Practical fundamental rights impact assessments. *International Journal of Law and Information Technology*, *30*(2), 200–232. https://doi.org/10.1093/ijlit/eaac018.

Leiser, M., & Santos, C. (2023). Dark Patterns, Enforcement, and the emerging Digital Design Acquis: Manipulation beneath the Interface. *BILETA Special Issue*, *1*(15). Retrieved from https://ssrn.com/abstract=4431048.

Möller, C. C. (2022). Dark Patterns in Consent-Bannern. *Verbraucher Und Recht*, *37*(12), 449–458.

Ramnarine, E. (2015). Understanding Problems of Subjectivity and Uncertainty in Quality Risk Management. *Journal of Validation Technology*, *21*(4).

Rohrßen, B. (2024). KI & CE — Die KI-VO, das Produktsicherheitsrecht für Künstliche Intelligenz. *Zeitschrift Für Produkt Compliance*, *1*(3), 111–123.

Ross, S. A. (1973). The economic theory of agency: The principal's problem. *The American Economic Review*, *63*(2), 134–139.

Schuett, J., Anderljung, M., Carlier, A., Koessler, L., & Garfinkel, B. (2024). *From Principles to Rules: A Regulatory Approach for Frontier AI* (Version 1). Version 1. arXiv. https://doi.org/10.48550/ARXIV.2407.07300.

von Welser, M. (2024). Die KI-Verordnung — ein Überblick über das weltweit erste Regelwerk für künstliche Intelligenz. *Gewerblicher Rechtschutz Und Urheberrecht in Der Praxis*, *16*(15), 485–488.

Wehkamp, N. (2022). Internalization of Privacy Externalities through Negotiation: Social costs of third-party web-analytic tools and the limits of the legal data protection framework. *Companion Proceedings of the Web Conference 2022*, 525–533. Virtual Event, Lyon France: ACM. https://doi.org/10.1145/3487553.3524631.

PART 2

THE EMERGENCE OF REGIONAL SOLUTIONS

6 The Incipient Latin American Approach to AI Governance: Highlighting Data Governance Issues through Emerging Supervisory Authorities

Pablo Trigo Kramcsák, Bárbara Lazarotto and Rocco Saverino

Abstract

Influenced by global trends, particularly the European Union's (EU) digital regulations, Latin American countries are starting to incorporate artificial intelligence (AI) rules into their data protection frameworks while exploring comprehensive AI laws.

This paper examines the emerging AI regulations in Latin America (LatAm), highlighting diverse approaches in countries such as Brazil and Chile, where the establishment of specialised AI regulatory bodies reflects the region's awareness of the complex issues these technologies present. The analysis emphasises data governance as a key factor in shaping AI oversight. As LatAm refines its approach to AI regulation, the region is well-positioned to contribute to the global discourse on AI governance.

Keywords: AI regulation, personal data protection, supervisory authorities.

Introduction

AI systems have rapidly emerged as a transformative technology. As these models evolve and their applications expand, coherent regulatory responses have become urgent. Around the world, countries are racing to establish AI regulations, often drawing inspiration from landmark legal frameworks like the EU's AI Act.

In LatAm, the journey toward AI governance has begun. However, these efforts remain in the early stages, marked, inter alia, by integrating AI governance into existing data protection frameworks, adopting a risk-based approach (classifying AI systems into different risk categories), creating new supervisory authorities, and emphasising data governance challenges.

This work analyses the region's adaptation to and engagement with global trends in AI regulation, with particular attention to data governance and supervisory authorities. It analyses emerging AI laws and regulatory frameworks in Brazil and Chile to present the region's challenges and opportunities in building an effective AI governance model.

6.1 Global Influence: The European Union's AI Regulatory Framework

The EU AI Act represents a pioneering legal framework, distinguished by its comprehensive, human-centric, and risk-based approach (Kusche, 2024). This Act has significantly shaped global discussions on AI governance. The EU's influence extends beyond its borders, primarily through what is known as the "Brussels Effect" (Bradford, 2020), where its regulations, such as the General Data Protection Regulation (GDPR), have set global standards that other regions often emulate (Greenleaf, 2021).

Like the GDPR, the AI Act is designed with extraterritorial reach (Hacker, 2023), meaning its impact is felt even in countries not part of the EU. This is particularly relevant for LatAm, where countries have historically aligned their data protection laws with the European legal approach.[19]

The AI Act's emphasis on data governance, transparency, and accountability is expected to have a similar influence on regional AI regulations. However, while the AI Act is setting the pace for global AI governance, the extent to which Latin American countries will replicate this model remains uncertain.

This uncertainty is closely linked to the situation in Europe, where each state's approach to relying on existing data protection authorities (DPAs) or establishing new AI authorities does not contribute to a harmonised framework. One of the most challenging aspects of enforcing the AI Act is the role of DPAs alongside AI authorities, particularly considering the potential variance in the structure of competent authorities from country to country. Even during the

19 See, e.g., Gadoni Canaan, 2023.

proposal stage of the AI Act, there was an apparent broadening of the supervisory framework within the GDPR (Chamberlain & Reichel, 2023). Given the close connection between data and AI systems, cases of overlapping and confusion regarding the competency of DPAs or AI authorities are always possible.

6.2 The Rise of AI Authorities in Latin America

Latin American countries are beginning to establish their own AI frameworks, which have been influenced by the EU[20] but tailored to their specific contexts. One of the critical aspects of these emerging frameworks is the creation of supervisory authorities responsible for overseeing AI systems. These efforts are still nascent, and there is considerable variation in how countries perceive these authorities.

For instance, Brazil's AI Law Proposal No. 2338/2023 outlines the creation of a National System of Regulation and Governance of Artificial Intelligence (SIA), which includes a network of authorities such as the Brazilian Data Protection Authority (ANPD), state AI regulators, and other entities responsible for AI certification and self-regulation. This multifaceted approach reflects Brazil's recognition of the complexity of AI governance and the need for a collaborative framework involving multiple stakeholders. More recently, the Brazilian Data Protection Authority (ANPD) issued an opinion on the bill, emphasising that the overlap between Brazil's General Data Protection Law (LGPD) and the AI governance framework could not be overlooked. Therefore, the ANPD should play a leading role in AI governance. After that, the bill was modified, and the ANPD was designated as SIA's coordinating authority.[21]

In parallel, Chile is advancing its AI governance model through Bill No. 16821-19, which proposes establishing an AI Technical Advisory Council to guide the Ministry of Science, Technology, Knowledge, and Innovation. This council will be complemented by

20 The major influence is Spain, with very close links to Latin American countries. Indeed, it is a member of the Ibero-American Data Protection Network and the Permanent Secretary. Furthermore, Spain was the first country to establish an AI authority independent of the existing data protection authorities: the Spanish Artificial Intelligence Supervisory Agency (AESIA).

21 Análise preliminar do Projeto de Lei nº 2338/2023, que dispõe sobre o uso da Inteligência Artificial. Available at https://www.gov.br/anpd/pt-br/assuntos/noticias/analise-preliminar-do-pl-2338_2023-formatado-ascom.pdf.

the Data Protection Agency, which will enforce the AI law once it is established under forthcoming legislation to modernise Chile's data protection framework.

These examples illustrate LatAm's varied approaches to AI governance, where existing data protection authorities are being reconfigured to take on AI oversight or new bodies are being created altogether. However, Latin American countries are exploring a broader spectrum of adaptation, ranging from close emulation of the EU model to more independent strategies.

6.3 Data Governance: A Central Issue in AI Regulation

Data governance is a crucial component of AI regulation, given that AI systems rely heavily on data for their development and deployment. The EU AI Act underscores the importance of data governance, emphasising transparency, accountability, and the protection of fundamental rights. This focus on data is mirrored in the emerging AI regulations in LatAm, where data protection remains a central concern.

In many Latin American countries, AI regulation efforts are closely linked to their data protection approaches, reflecting the influence of the GDPR. The GDPR's significant impact on crucial aspects of AI systems, such as big data processing, profiling, and automated decision-making, should be noted.

For example, Brazil's LGPD, which mimics the GDPR (Erickson, 2019), plays a significant role in the country's AI governance framework. The LGPD's principles of transparency, purpose limitation, adequacy, necessity, prevention, data quality, non-discrimination and accountability are expected to extend to AI systems (Belli et al., 2023), ensuring that they operate within a framework that prioritises protecting personal data.

Chile's approach to AI governance also highlights data protection issues. A yet-to-be-established Data Protection Agency will oversee the proposed AI law. This approach underscores the importance of data governance in AI regulation, as the effectiveness of AI oversight will largely depend on the robustness of the underlying data protection framework.

Nonetheless, effective AI regulation faces steep data governance hurdles (fragmented data protection laws, uneven institutional capacity, and the struggle to balance innovation with fundamental rights). Additionally, integrating AI governance into existing frameworks raises a critical question: Do current data protection authorities have the expertise and resources to oversee AI systems effectively?

6.4 Challenges and Opportunities in Latin American AI Governance

Developing AI governance frameworks in LatAm presents challenges and opportunities. On the one hand, the region can draw on the experiences of other regions, such as the EU.[22] On the other hand, Latin American countries must navigate a complex landscape of political, economic, and social factors that are very different from those of European countries, which may hinder the implementation of such frameworks.

The Latin American social and political environment makes these countries vulnerable to abuse through AI systems. For instance, facial recognition technology in Brazil is often used as a security measure due to the country's high number of crimes. However, this technology often infringes on individuals' human and fundamental rights, a concern that must be carefully addressed in Latin American AI laws.[23]

Therefore, one of the main challenges is the need for coordination among different regulatory bodies. The creation of multiple supervisory authorities, as seen in Brazil's AI Law Proposal, can lead to fragmentation, inefficiency, and potential rights violations if these authorities do not work together effectively. Ensuring that these bodies have clear mandates and mechanisms for coordination and collaboration will be crucial for the success of AI governance in the region.

Another challenge is the need for sufficient resources and expertise. Many Latin American countries face limited institutional capacity, which could weaken the effectiveness of AI regulation. Developing the necessary expertise within supervisory authorities, particularly in the

22 See, e.g., Novelli et al. 2024.

23 See Ramiro & Cruz, 2023..

technical aspects of AI, will be highly relevant for effective oversight. Moreover, securing the financial support to sustain these authorities is a critical challenge, particularly in countries with tight public budgets.

Despite these challenges, Latin American countries have significant opportunities to shape AI governance proactively. Adopting a risk-based approach, like the EU AI Act, allows the development of AI regulations that balance innovation with the protection of fundamental rights. Additionally, integrating AI governance into existing data protection frameworks enables the region to leverage its data protection experience, ensuring AI systems operate transparently and accountably.

LatAm also could contribute to the global discourse on AI governance by developing regulatory models that reflect its unique social, economic, and cultural contexts. While the region may draw inspiration from the EU, it is well-positioned to innovate and develop flexible and structured approaches that address AI's specific challenges and opportunities in the Majority World. For instance, the region's emphasis on social justice and human rights could lead to developing AI regulations that prioritise protecting vulnerable populations and promoting equitable access to AI technologies.

6.5 The Path Forward: Toward a Coherent AI Governance Framework

As Latin American countries continue to develop their AI governance frameworks, several key issues must be addressed to ensure the effective regulation of AI. First, there is a need for greater harmonisation of AI regulations across the region. While the diversity of approaches reflects the different contexts of each country, a more coordinated approach could help address cross-border issues and promote regional collaboration in AI governance. Harmonisation does not necessarily mean uniformity but rather the alignment or convergence of key principles and standards to ensure a consistent approach to regional AI regulation.

Second, data protection authorities' role in AI governance must be clearly defined. It is essential to ensure these authorities are equipped with the necessary expertise and resources to effectively oversee and

regulate the data processing aspects of AI systems. This may require capacity-building initiatives, increased funding, and the development of new regulatory tools and methodologies specific to AI.

Third, there is a need for greater public engagement and transparency in developing new AI governance frameworks. AI regulation should not be a top-down process; instead, it should involve a broad range of stakeholders, including civil society, industry, academia, and the public. Public engagement can help build trust in AI systems and ensure that AI regulations reflect the values and concerns of society. Additionally, transparency in the regulatory process can help ensure that AI governance is accountable and that the decisions made by policymakers and regulatory authorities are open to public scrutiny.

Finally, Latin American countries should consider the potential for regional cooperation in AI governance, being necessary to improve the incentives and conditions that allow collaboration in this area, for example, overcoming the transaction costs associated with AI governance and regulation (Contreras, 2024). Initiatives such as creating a regional AI governance framework could help coordinate regional efforts and promote sharing best practices. Regional cooperation could also enhance the region's ability to engage in the global discourse on AI governance and ensure that Latin American perspectives are represented on the world stage.

6.6 Conclusion

The incipient Latin American approach to AI governance reflects the region's recognition of the importance of regulating AI systems in a manner that aligns with global standards while addressing local needs and contexts. While still in its early stages, this approach is marked by a growing awareness of the critical role that data governance plays in the effective oversight of AI technologies. Drawing from the foundation established through data protection laws, Latin American countries are starting to establish supervisory authorities capable of addressing the unique challenges posed by AI.

However, the region faces significant challenges, including the need for greater coordination among regulatory bodies, developing specialised expertise, and allocating sufficient resources to support

effective oversight. Additionally, integrating AI governance into existing frameworks raises important questions about DPAs' capacity to manage the complexities of AI regulation.

LatAm has substantial potential to shape the AI governance debate. A proactive and regionally coordinated approach would enable the region to contribute significantly to the global regulatory conversation while safeguarding citizens' rights, emphasising principles such as social justice, equity, and human rights.

6.7 References

Belli, L, Curzi, Y. & Gaspar, W. B. (2023). AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience, Computer Law & Security Review 48, 105767, https://doi.org/10.1016/j.clsr.2022.105767.

Bradford, A. (2020). The Brussels Effect: How the European Union Rules the World. New York: Oxford University Press.

Chamberlain, Johanna, & Reichel, Jane. (2023). Supervision of Artificial Intelligence in the EU and the Protection of Privacy. FIU Law Review, 17(2), 267-286.

Contreras, P. (2024). International Convergence and Own Paths: Regulation of Artificial Intelligence in Latin America. Actualidad Jurídica Iberoamericana 21, 468-493.

Erickson, A. (2019). Comparative Analysis of the EU's GDPR and Brazil's LGPD: Enforcement Challenges with the LGPD. Brook. J. Int'l L 44, 859.

Gadoni Canaan, R. (2023). The effects on local innovation arising from replicating the GDPR into the Brazilian General Data Protection Law. Internet Policy Review, 12(1). https://doi.org/10.14763/2023.1.1686.

Greenleaf, G. (2021). The "Brussels Effect" of the EU's "AI Act" on Data Privacy Outside Europe. Privacy Laws & Business International Report 117(1), 3-7.

Hacker, P. (2023). AI Regulation in Europe: From the AI Act to Future Regulatory Challenges. arXiv <https://arxiv.org/abs/2310.04072>.

Kusche, I. (2024). Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. Journal of Risk Research, 1–14. https://doi.org/10.1080/13669877.2024.2350720.

Novelli, C., Hacker, P., Morley, J., Trondal, J. & Floridi, L. (2024). A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities. Available at SSRN: https://ssrn.com/abstract=4817755 or http://dx.doi.org/10.2139/ssrn.4817755.

Ramiro, A. & Cruz, L. (2023). The grey-zones of public-private surveillance: Policy tendencies of facial recognition for public security in Brazilian cities. Internet Policy Review 12 (1), 1-28.

7 The RICE Governance Framework: Enabling Comprehensive Data Governance in Africa

Chinasa T. Okolo, Ph.D.

Abstract

New complexities around data production, refinement, and use have impacted African countries, elevating a need for comprehensive data regulation and enforcement measures. While 38/55 African Union Member States have existing data protections, there is a wide disparity in the robustness of these regulations and in the ability of individual countries to enforce these respective protections. This work introduces the RICE (Reformation, Integration, Cooperation, & Enforcement) Data Governance Framework, which aims to operationalize comprehensive data governance in Africa by outlining best measures for data governance policy reform, integrating revamped policies, increasing continental-wide cooperation in AI governance, and improving enforcement actions against data privacy violations.

Keywords: Data privacy, data governance, policy reform, African development, artificial intelligence.

Introduction

The advent of generative artificial intelligence (AI), increasing adoption of AI tools, and the widespread utilization of data workers have changed narratives around data production and use. While data protections exist in 38 out of 55 African Union (AU) Member States, intensifying algorithmization across Africa could impact users through digital platforms used to access education, healthcare, financial, and social services. Given these new complexities and the emerging AI regulatory environment within the continent, African governments must enact comprehensive data protection regulations and reform existing data governance measures to cover aspects such as data quality, privacy, responsible data sharing, transparency, and data worker labor protections. To address these issues, data workers in Kenya have pursued litigation against Facebook regarding subpar

working conditions and unfair termination (Musanga, 2023), and data workers across the continent have established organizations such as Techworker Community Africa (TCA)[24], the African Content Moderators Union, the Nigerian Content Moderators and Tech Workers Union (NCMTW)[25], and the Kenyan Content Moderators' Union. Along with general subpar working conditions across the continent in fields such as oil production and garment manufacturing, the concerns imposed by data work underscore requirements for sectoral reform of existing labor protections in areas including agriculture, economics, education, and healthcare. African countries also have context-specific challenges that differ significantly from those within the West, highlighting a need to understand how to develop culturally aligned and feasible governance solutions (Okolo, 2023).

By balancing lessons from the recent ratification of the African Union Convention on Cyber Security and Personal Data Protection, maturing regulatory environments like the EU, and advancing research on regional and country-specific needs, African nations can work towards more robust regulation. This paper analyzes data governance measures in Africa, outlines data privacy violations across the continent, and examines regulatory gaps imposed by a lack of comprehensive data governance to outline the sociopolitical infrastructure required to bolster data governance capacity. Additionally, it proposes the RICE (Reformation, Integration, Cooperation, & Enforcement) Data Governance framework, which African national governments (NGs), Regional Economic Communities (RECs), and the African Union can leverage to reform and operationalize existing data protection measures. Ultimately, this framework could inform the development and implementation of context-specific AI regulation that centers data privacy rights.

7.7.1 Data Protection Regulation in Africa

The increasing development and adoption of AI have dramatically shifted practices around data, spurring the development of new industries and revealing new forms of exploitation. This has also introduced gaps within existing data protection regulations that could be further

24 https://techworkercommunityafrica.org/.

25 https://www.linkedin.com/company/nigerian-content-moderators-acmu/.

exploited as AI development increases throughout the continent. While companies have traditionally leveraged consumer data to improve ad targeting and personalized recommendations, companies are now leveraging existing consumer data to train AI tools, which few existing data protection regulations have sufficient coverage for. These new complexities around data production, refinement, and use elevate a need for comprehensive governance and enforcement measures. Approximately 38 out of 55 African Union Member States have enacted formal data protection regulations. Some of these countries include top economies within the continent, such as Egypt, Nigeria, and South Africa, and emerging players like Benin, Equatorial Guinea, and Zimbabwe. 15 out of 38 data protection laws passed by African countries were enacted in the last five years, and 26 were enacted in the last decade. The first data protection law in Africa was enacted by Cabo Verde in 2001, and data protection laws were recently enacted by Malawi in June 2024 and Ethiopia in July 2024. As of October 2024, Namibia, South Sudan, and The Gambia have drafted data protection laws yet to be enacted. Along with country-specific data governance regulations, regional efforts towards data protection include the African Union Convention on Cyber Security and Personal Data Protection (African Union, 2020), the Economic Community of West African States (ECOWAS) (ECOWAS, 2010), the East African Community (EAC) Legal Framework for Cyberlaws (East African Community, 2008), and the Southern African Development Community (SADC) Model Law on Data Protection (International Telecommunication Union, 2013). At the moment, there have been no regional governance measures proposed or enacted by the Arab Maghreb Union (AMU), the Community of Sahel–Saharan States (CEN–SAD), and the Economic Community of Central African States (ECCAS). While African countries have made significant progress in enacting data protection laws, various factors hinder responsible and sustainable data governance throughout the continent. Additionally, the rising adoption of AI tools introduces new gaps within existing data protection regulations that could be further exploited as AI development increases throughout the continent.

7.7.2 Data Privacy Violations in Africa

Existing data regulatory gaps may also contribute to the growing number of data privacy violations experienced across Africa. In March

2023, the Angolan Agência de Protecção de Dados (APD) issued a fine to Africell, an electronic communications operator, who collected personal consumer data without requesting prior authorization from APD (Agência de Protecção de Dados, 2023). In November 2023, the Telecommunications/ICT Regulatory Authority of Côte d'Ivoire (ARTCI) issued a formal warning to YANGO, a local ridesharing application, for unlawfully recording passenger phone conversations (l'ARTCI, 2023). In July 2023, the South African Information Regulator issued a ZAR 5 million (~USD 273,000) fine against the Department of Justice and Constitutional Development for failure to implement adequate security measures to prevent a ransomware attack in 2021 and noncompliance with required consumer notifications regarding the subsequent data breach (Information Regulator South Africa, 2023). One of the continent's most recent data privacy violations involves a data breach of Nigeria's National Identity Management Commission of Nigeria (NIMC) system, which has resulted in millions of data points being available for sale on illicit websites for NGN 100 each, which is about USD 6 cents (Paradigm Initiative, 2024). As of October 2024, it is unclear what action the Nigeria Data Protection Commission has taken against the offenders. Kenya Office of the Data Protection Commissioner (ODPC) issued multiple penalties to 4 companies in 2023, totaling over KES 14 million. These fines included noncompliance with a prior enforcement notice on spam calls, harassment from microlending apps, posting minor images, and using customer photos for marketing. ODPC has also made progress in an ongoing investigation regarding violations by Worldcoin, an American cryptocurrency provider that undertook biometric data collection without government notice (Communications Authority Kenya, 2023). While African data protection agencies have increasingly taken actions toward enforcing data protection laws, there is still little understanding of how effective these measures are, given frequent noncompliance with enforcement notices and little information on fine payments by offenders (Lawyers Hub, 2024).

7.1 Operationalizing Data Governance in Africa

In order to ensure that African countries can effectively protect consumers against improper data practices and enforce corrective action against data privacy violations, African governments across

every AU Member State must enact comprehensive data regulatory measures. While existing continental-wide efforts, such as the African Union Data Policy Framework, which was published in 2022 to guide AU Member States in designing and reviewing data regulations, and the Malabo Convention on Cyber Security and Personal Data Protection, offer valuable templates for African governments to adopt, these frameworks have unfortunately not seen wide adoption. To help address this lack of adoption and potential challenges from data regulatory gaps, a number of proposals have outlined alternative measures, including regional data governance approaches (Osakwe & Adeniran, 2021; Balogun & Adeniran, 2024), community-centered governance models (Olorunju & Adams, 2024), and data governance reformation (Okolo, 2024). This section introduces the RICE Data Governance Framework to provide a high-level overview of actions African Union Member States can leverage to operationalize data governance effectively.

7.1.1 Reformation, Integration, Cooperation, & Enforcement (RICE) Framework

To begin operationalizing the RICE framework, African governments should pursue regional data governance measures, given the lack of existing coordination with and insufficient protections within existing continental measures such as the Malabo Convention (Yilma, 2022; ALT Advisory, 2022). Efforts to pursue regional data governance would ideally be led by existing RECs such as ECOWAS, EAC, SADC, AMU, CEN-SAD, and ECCAS. Such efforts can then enable the 19 African Union Member States without existing data protections to draft and enact comprehensive data governance measures in a reasonable timeframe. Additionally, enacting regional data governance policies can help address existing capacity constraints for AU Member States unable to individually draft and enact data legislation.

In lieu of functional continental frameworks, countries, regional, and continental bodies should focus on (1) **reforming** existing data regulation and implementing sectoral policy reformation, (2) collaborating with Civil Society Organizations (CSOs) and Academic Research Institutions (ARIs) to improve **integration** of reformed policies, (3) increasing regional and continental **cooperation** in

data regulation efforts, and (4) strengthening **enforcement** of reformed data regulation. The RICE Data Governance Framework recommendations apply at the national, regional, and continental levels, and the core tenets of the framework are defined as follows:

Reformation: To address concerns regarding a lack of comprehensive data governance measures, the AU, RECs, and individual African NGs must reform existing data governance measures and engage in sectoral policy reform. These entities must also establish local expert groups and advisory bodies to enhance policy reform.

The AU, RECs, and NGs should review existing data protection measures, and to meet data governance needs, they should subsequently reform sectoral policies in agriculture, economics, education, healthcare, and other areas.

Integration: To increase awareness and local integration of data protection regulation, RECs and NGs will need to improve outreach to organizations under their jurisdiction. RECs and NGs should also fund outreach and research efforts by CSOs and ARIs to improve public engagement with data protection measures.

ARIs and CSOs should also focus on conducting in-depth research that advances understanding of regional and country-specific needs for data regulation and reduces reliance on standards such as the EU General Data Protection Regulation (GDPR).

Cooperation: To address issues regarding a lack of regional cooperation and inconsistencies in data protection regulation, the AU must lead harmonization efforts across AU Member States. To mitigate issues with prior harmonization efforts (Kenyanito & Chima, 2016), the AU should actively consult RECs and NGs in new harmonization efforts.

The AU should also establish a continental-wide network of National Data Protection Authorities and Offices (NDPAs/NDPOs), as previously recommended in prior work (Data Protection Africa, 2023).

Enforcement: To help address concerns regarding a lack of enforcement of data protection measures, the AU must establish a continental data supervisory body. African governments must also establish and leverage data protection offices to enforce enacted data protection regulations.

The AU should inaugurate a Data Protection Supervisory Authority (DPSA) to increase regional enforcement for data privacy violations and should also help NGs establish NDPAs and NDPOs to mitigate regulatory enforcement gaps.

7.2 Considerations

While this data governance operationalizing framework aims to ease the implementation of comprehensive data regulation within African countries, many considerations exist for the ability of all governments across the continent to leverage this framework. Existing issues with infrastructure, electricity access, education, digital skills literacy, skilled AI talent, climate change, armed conflict, social unrest, national security, and socioeconomic growth may deprioritize and sideline efforts toward data governance. In light of these existing challenges, however, governments must focus on developing culturally aligned and feasible data governance solutions to ensure that the data rights of African consumers are preserved and that there are adequate outlets for redress of data protection harms.

Regional data governance led by RECs would ideally take precedence over the AU until a formal continental-wide data protection law is passed. However, efforts will be needed to rectify duplicative membership within the RECs and integrate AU Member States without membership in RECs, like the Sahrawi Arab Democratic Republic, which controls the Western Sahara. Prioritizing regional-led data governance before continental reforms are enacted could help address capacity constraints and harmonization issues between AU Member States. Still, there is no guarantee that countries within RECs will reach alignment on data governance measures.

With the growing number of regional and national efforts toward AI regulation throughout the continent, African governments must also understand the fundamental role of data in training ML models, evaluating AI systems, refining predictive models, and improving AI-enabled services (Data Governance Working Group of the Global Partnership on AI, 2020). Given these essential functions, efforts towards enacting effective data governance can also enable more comprehensive AI governance measures. Thus, African governments should consider comprehensive data governance as a viable pathway

and complement to AI regulation. To bolster AI-related governance overall, it will also be crucial for African governments to invest in efforts to understand the diverse policy challenges associated with data, including privacy, transparency, labor, interoperability, discrimination, cross-border data flows, and intellectual property.

7.3 Conclusion

While the potential of AI is still nascent within Africa, African consumers hold valuable data that is subject to exploitation by both local and international firms alike. Companies are increasingly looking towards African countries to supply them with the necessary data to expand target markets for their AI services. With governments, companies, universities, and other institutions in African countries rapidly adopting AI technologies, there are also concerns that algorithmic harms primarily noted in Western contexts could be exacerbated in ways that disproportionately harm marginalized populations throughout the continent. The limited research examining concrete ethical concerns around data privacy and the lack of extensive efforts toward data protection in Africa is concerning. This work examines data governance measures in Africa, highlighting the regulatory gaps imposed by a lack of comprehensive data governance across Africa that could be further exploited by rising AI adoption. This work presents the RICE Data Governance framework to operationalize comprehensive data governance in African Union Member States to reform and optimize existing data protection measures while bolstering Africa's emerging AI regulatory environment.

7.4 References

African Union. (2020). African Union Convention on Cyber Security and Personal Data Protection. https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection.

Agência de Protecção de Dados (APD). (2023). APD multa AFRICELL em 150 mil dólares norte americanos por violação da Lei de Protecção de Dados Pessoais (LPDP). https://www.apd.ao/ao/noticias/apd-multa-africell-em-150-mil-dolares-norte-americanos-por-violacao-da-lei-de-proteccao-de-dados-pessoais-lpdp/.

ALT Advisory. (2022). The Malabo Roadmap: Approaches to promote data protection and data governance in Africa. Mozilla. https://dataprotection.africa/wp-content/uploads/malabo_roadmap_Sept_2022.pdf.

Balogun, K., & Adeniran, A. (2024). Towards A Sustainable Regional Data Governance Model In Africa. Centre for the Study of African Economies (CSEA).

Communications Authority Kenya. (2023). CA and Data Commissioner Warn Kenyans Over Worldcoin. https://www.ca.go.ke/ca-and-data-commissioner-warn-kenyans-over-worldcoin.

Data Governance Working Group of the Global Partnership on AI (GPAI). (2020). The Role of Data in AI. GPAI. https://gpai.ai/projects/data-governance/role-of-data-in-ai.pdf.

Data Protection Africa. (2023). Africa: AU's Malabo Convention set to enter force after nine years. ALT Advisory. https://dataprotection.africa/malabo-convention-set-to-enter-force/.

East African Community. (2008). Draft EAC Legal Framework for Cyberlaws. http://repository.eac.int/handle/11671/1815.

ECOWAS. (2010). Supplementary Act A/SA.1/01/10 on Personal Data Protection Within ECOWAS. https://www.statewatch.org/media/documents/news/2013/mar/ecowas-dp-act.pdf.

Information Regulator South Africa. (2023). Media Statement Infringement Notice and R5 Million Administrative Fine Issued to The Department of Justice and Constitutional Development for Contravention of Popia. https://inforegulator.org.za/wp-content/uploads/2020/07/MEDIA-STATEMENT-INFRINGEMENT-NOTICE-ISSUED-TO-THE-DEPARTMENT-OF-JUSTICE-AND-CONSTITUTIONAL.pdf.

International Telecommunication Union (ITU). (2013). Data Protection: Southern African Development Community (SADC) Model Law. https://www.itu.int/en/ITU-D/Projects/ITU-EC-ACP/HIPSSA/Documents/FINAL%20DOCUMENTS/FINAL%20DOCS%20ENGLISH/sadc_model_law_data_protection.pdf.

Kenyanito, E. P., & Chima, R. J. S. (2016). Room for improvement: Implementing the African Cyber Security and Data Protection Convention in Sub-Saharan Africa. AccessNow.

l'ARTCI. (2023). Communiqué — l'ARTCI. https://www.artci.ci/index.php/33-actualites/informations/629-probables-enregistrements-des-communications-ou-echanges-a-l-interieur-de-vehicules-utilisateurs-de-l-application-denommee-yango-sans-information-prealable-ou-consentement-des-personnes-concernees.html.

Lawyers Hub. (2024). Africa Privacy Report 2023/2024. https://www.lawyershub.org/digital.

Musanga, M. (2023). Facebook workers in Kenya say Meta hasn't paid them for 6 months amid legal case. openDemocracy. https://www.opendemocracy.net/en/facebook-workers-in-kenya-say-meta-hasnt-paid-them-for-6-months-amid-legal-case/.

Okolo, C.T. (2023). AI in the Global South: Opportunities and challenges towards more inclusive governance. The Brookings Institution.

Okolo, C.T. (2024). Reforming data regulation to advance AI governance in Africa. Foresight Africa 2024. The Brookings Institution. https://www.brookings.edu/articles/reforming-data-regulation-to-advance-ai-governance-in-africa/.

Olorunju, N., & Adams, R. (2024). African data trusts: new tools towards collective data governance?. Information & Communications Technology Law, 33(1), 85-98.

Osakwe, S., & Adeniran, A. (2021). Strengthening Data Governance in Africa. Centre for the Study of African Economies (CSEA).

Paradigm Initiative. (2024). Major Data Breach: Sensitive Government Data of Nigerian Citizens Available Online for Just 100 Naira. https://paradigmhq.org/major-data-breach-sensitive-government-data-of-nigerian-citizens-available-online-for-just-100-naira/.

Yilma, K. (2022). African Union's data policy framework and data protection in Africa. Journal of Data Protection & Privacy, 5(3), 209-215.

8 AIED and student data privacy in Africa: challenges and recommendations for legislators

Andrea Bauling

Abstract

Discussions around artificial intelligence in education (AIED) can no longer focus purely on what is technologically possible and pedagogically sound. Advances must be considered within a framework for lawful and responsible learning analytics and data science practices. Lagging efforts to address a widespread lack of AI-specific legislation may be harming millions of students from the majority world. Facilitating the lawful development and implementation of AIED agents that are suited to the needs of African students requires a homegrown approach. Adopting Africa-focussed solutions and legislation could ensure that the great benefit AIED agents may hold for humanity safely includes Africans and others from the global majority.

Keywords: AI in education (AIED), AI regulation, data monetization, data privacy, global majority interests, higher education.

Introduction

The capabilities and social impact of artificial intelligence (AI)[26] agents are expanding at an unprecedented speed. Efforts to regulate AI are lagging, with potentially dire consequences. The challenges faced in the field of AI in education (AIED), which focusses on the use of AI agents to improve educational outcomes and environments, illustrate the need to bolster regulatory efforts. While there are numerous benefits to the use of AIED, many are concerned about the implications for data privacy and data security. Non-existent, ill-suited, and/or unenforced legislation compounds the problem. The inadequacy of the current South African legislative framework

26 For the purposes of the paper, AI is understood as defined by Popenici and Kerr (2017). Generative AI falls beyond the ambit of this paper. See Bozkurt et al. (2023) on generative AI and education.

demonstrates the potentially corresponding risks threatening many jurisdictions from the majority world. Educators representing the global majority should make their voices heard in spheres where technologies and related policies that affect them are developed. This paper attempts to sketch current and potential future challenges related to data privacy infringements by AIED agents, as well as the legislative steps that could be taken to address these.

8.1 AIED and the Processing of Student Data

It is necessary to define certain key concepts to facilitate a discussion on the benefits and dangers that AIED may hold for the stakeholders of higher education systems. Educational datamining (EDM) involves the development and application of datamining and machine learning approaches to change raw data collected from education systems and databases into usable information extracted from patterns and connections identified in the data (Maphosa & Maphosa, 2021). The field of learning analytics (LA) concerns developing an understanding of an individual student and their performance in a specific learning environment, often hosted on an online learning management system (LMS), by gathering and analysing personal learning data to ultimately improve learning outcomes and optimise the learning environment (Long & Siemens, 2011; Prinsloo & Slade, 2015). The primary objective of EDM and LA is to support developers, educators, and institutions in their decision-making (Maphosa & Maphosa, 2021).

AIHED, a booming subfield of AIED dedicated to higher education, can bolster teaching and learning efficiency. Under the supervision of an educator these systems can facilitate immediate instruction, student supervision, and feedback (Bond et al., 2024; Zawacki-Richter, Marín, Bond & Gouverneur, 2019). Intelligent tutoring systems (ITS) are but one example of an AIED-supported intervention in student learning. ITS can teach content, diagnose strengths and weaknesses in student understanding, curate learning materials, and support peer collaboration (Zawacki-Richter et al., 2019). In some instances, they facilitate a form of computed curriculum that can provide a continuously personalised learning experience in real time, based on a learner's pre-existing knowledge, skills, and rate of progress

(Bernhardt, 2023). Clearly, the pedagogical value of such systems is undeniable, but focussing exclusively on their benefits is shortsighted.

8.2 The Interplay Between Personalised Learning and Data Privacy

The complex dichotomy between safeguarding and sharing student, academic, and institutional data, and the development and implementation of AIED agents epitomise "the personalization privacy paradox" (Xu, Luo, Caroll & Rosson, 2011, p.43). AI agents can collect, process, aggregate, and repurpose vast volumes of data housed in institutional silos to generate meaningful insights (Pelletier et al., 2023). But for AI to effectively do so, it must be trained (Bernhardt, 2023). AI agents mainly process personal information in two ways: this data is incorporated in immense datasets employed to train AI machine-learning systems to develop algorithmic models; and once developed, these algorithmic models are applied to other datasets containing personal information to extrapolate predictions about individuals (Bhagattjee, Govuza & Sebanz, 2020). Within an educational context, the individuals in question are students, educators, and administrators.

AIED must be developed by judiciously curating the initial training data used, which is largely based on data generated through EDM and LA activities (Prinsloo & Kaliisa, 2022). Examples of the highly personalised student data processed by LA and EDM systems include learning capabilities and challenges; assessment results and prior academic performance; interaction traces with online content; demographics; funding data; disability status; and health-related indicators (Li, Sun, Schaub & Brooks, 2022; Slade, Prinsloo & Khalil, 2019). From this, AIED agents can deduce students' capabilities, assumed emotional states, mental strategies, and misconceptions (Holmes et al., 2022). Algorithmic models are already capable of diagnosing mental health disorders (Alkahtani, Aldhyani & Alqarni, 2024) and neurodevelopmental disorders such as attention-deficit hyperactivity disorder (Chen et al., 2023). LMSs fully supported by AI are likely to become the new norm (Pelletier et al., 2023). It is not hard to imagine AI-powered diagnostic tools being incorporated into AIED agents and LMSs in the near future, all in the name of

pedagogical progress. The implications for data privacy could be astronomical. Many students prefer to keep their highly personal data private and rightfully fear (future) discrimination based thereon, but they mostly have very little (if any) control over what data of theirs is being collected, repurposed, stored, and shared (Li et al., 2022; Zawacki-Richter et al., 2019).

8.2.1 Ownership of and Access to Educational Data

Of great concern is the fact that higher education institutions are (inadvertently) gathering masses of data on their students (Slade et al., 2019). The volume of data collected by LMSs alone is almost unfathomable.[27] Each student, educator, and administrator's every click is logged and "[t]here are many unanswered questions about who owns this data, who has access to it, [and] how long it will be kept" (Du Boulay, 2023, p.100).[28] At the emergence of LA, most of the data harvested was anonymised, but this is no longer the case (Slade & Prinsloo, 2014). In the pursuit of improved student performance, the prevalence of EDM and LA is increasing, and data is being processed and aggregated in ways that were not initially anticipated or communicated (Willis, Slade & Prinsloo, 2016). Once modern LMSs are implemented "[s]urveillance is insidious and constant" (McGowan et al., 2024, para. 24). The context within which user consent was given, or not,[29] for the collection of (often seemingly harmless) data, becomes further removed from what it may be used for in future, especially as AI algorithmic capabilities progress.[30] The need to protect the personal data of students gathered by higher education institutions, LMSs, and other third-party service providers is evident.

---

27 In October 2024, the world's largest LMS, Moodle, boasted hosting more than 2,4 billion enrolments from 239 countries, 427 million active users, and 801 million discussion forum posts (Moodle, 2024). Moodle provides an invaluable, openly available platform that learning institutions can use freely and modify to suit their needs. While this approach is laudable, the effect is that the organisation has access to quadrillions of datapoints on users from across the globe. Worryingly, biometric data is collected for features such as facial recognition, used to proctor online assessments.

28 See McGowan, Paris & Reynolds (2024) on the dangers inherent in procuring AIED systems under "software-as-a-service" (SAAS) agreements.

29 Higher education institutions often grant consent to vendors or external service provider on users' behalf, and without their knowledge (McGowan et al., 2024).

30 See Slade & Prinsloo (2014) on this "context collapse".

8.3 The South African Legal Position and Potential Regional Interventions

It is essential to consider how we protect the right to data privacy of students and educators from the global majority whose personal data is being collected by LMSs and other for-profit corporations, mainly situated in the developed world. An evaluation of the woefully deficient regulations currently applicable in South Africa provides valuable insights into the legislative challenges faced, which are likely similar to those of various other majority-world jurisdictions.

South African law is enacted, interpreted, and enforced within a constitutionally supreme framework (ss.1(c) & 2 of the Constitution of the Republic of South Africa, 1996). Section 14 of the Constitution protects the right to privacy and the Constitutional Court has confirmed that "the invasion of an individual's privacy infringes the individual's cognate right to dignity" (*AmaBhungane v Minister of Justice* (2021), para.28). To give effect to the right to data privacy, one aspect of the fundamental right to privacy, Parliament enacted the Protection of Personal Information Act (2013) (POPIA). This act currently regulates automated data processing in the jurisdiction, as no other legislation specifically regulating AI has been adopted. As in many other jurisdictions, POPIA is based on its EU counterpart, the General Data Protection Regulation (2016) (GDPR).

Various global data protection laws like the GDPR and POPIA impose data minimisation and purpose limitation principles that restrict what personal data may be collected and how it is processed. These principles are wholly incompatible with the essence of AI-powered data processing and the training of models capable of such activities (Bhagattjee et al., 2020). Unfortunately, POPIA does not prescribe data protection impact assessments or any other accountability requirements as the GDRP does, which diminishes the potency of the Act's regulatory capabilities (Bronstein, 2022). A further point of concern is that POPIA has, to date, not been enforced in earnest (Musoni & Mtuze, 2023). These challenges have serious implications for achieving the goal with which this law was enacted. Based on these and other concerns, South African legal scholars support the promulgation of AI-specific legislation and argue that this should

encapsulate definitive prescripts on the degree of repurposing of personal information by AI agents that would be considered lawful (Bhagattjee et al., 2020; Mahomed, 2018; Musoni & Mtuze, 2023).

Some jurisdictions have moved beyond merely relying on data privacy legislation. In 2023 the European Parliament passed the EU Artificial Intelligence Act (2021). Crucially, the Act classifies the education sector as a high-risk field in which to apply AI systems (arts.6(2), 8 & 9). Because of the high potential for harm to individuals, the Act requires continuous risk assessment and management of AI agents developed for and implemented in educational settings (arts.8 & 9). This special focus on potential risk is sensible.

As in the EU, member states of the Southern African Development Community (SADC) are cooperating in various regional initiatives to coordinate data protection practices (Thaldar & Malekela, 2024). Since a collaborative regional approach to AI regulation would serve SADC citizens and activities (Gwagwa, Kraemer-Mbula, Rizk, Rutenberg & De Beer 2020), engaging these existing working groups could add significant value to discussions on data sharing and data protection, as relevant to AI development projects. While it is paramount that African solutions are adopted to solve African problems, it may be prudent to use the EU AI Act as springboard for a project of this nature (Gwagwa et al., 2020). African AI legislation will need to embrace the inherent dichotomy at play in regulating AI development: promoting technological progress and access to the immense promise of AI, and protecting the interests of the persons these AI agents aim to serve.

8.4 The Necessity of an African Approach to AI and AIED Regulation

Regulators the world over are attempting to circumvent the potential social harm that AI agents may cause by developing global standards for AI (Karanicolas, 2023; 2024). Karanicolas (2023, pp.266-267) argues that "the world would be better served if the standard-setting processes represented … perspectives from the people of the Majority World". One aspect of the potential social harm in question stems from the bias and (often race-based) discrimination inherent in many algorithms and AI models originating in the developed world (Jiao,

Afroogh, Xu & Phillips, 2024). The race to prevent or rectify harm of one form by diversifying datasets, may inadvertently cause another: the infringement of the right to data privacy of millions. Campbell-Stephens (2021, p.6) explains that "[t]he term 'global majority' invites social cooperation across groups, existentially to address the mutual interests of the majority on planet earth through collective mobilisation." It is crucial that such collective efforts extend to the sphere of AI and ultimately AIED development and regulation. The case of the open university illustrates the necessity in this regard.

Open distance universities can provide higher education at scale and therefore serve the needs of the majority world well. At a conservative estimate, the 10 largest public open universities in the world[31] service almost 20 million students, almost all from the majority world (Bozkurt, 2019; De Vries, 2019; Quayyum & Zawacki-Richter, 2019; Zhang & Li, 2019). The most cost-effective way for open universities to provide higher education to hundreds of thousands, or millions, is to do so online by means of an LMS. Initial agreements with LMS service providers often entail seemingly innocuous terms and conditions, which upon closer inspections could have significant implications for data privacy through the assetisation of higher education and the commodification of student data (Prinsloo & Kaliisa, 2020).

The higher education sector has come to be regarded as "a site of value and ongoing wealth extraction" (Scott & Gray, 2023, p.606). Higher education institutions have both a fiducial and moral duty to consider the paramountcy of safeguarding the data privacy of their students and staff when contracting with external education platforms (Prinsloo & Kaliisa, 2020). This may be especially true for open universities and African higher education institutions, as they are most vulnerable to "datafication" and exploitation by international corporations (Bozkurt, 2019; Prinsloo & Kaliisa, 2020). This raises legitimate "concerns about Africa being re-colonised and its data exported and capitalised" (Prinsloo & Kaliisa, 2020, p.896). It is

31 Indira Gandhi National Open University (India), Open University of China, Anadolu University (Türkiye), Allama Iqbal Open University (Pakistan), Bangladesh Open University, National Open University of Nigeria, Dr. B.R. Ambedkar Open University (India), Payame Noor University (Iran), and University of South Africa (see Jones, 2018; Quayyum & Zawacki-Richter, 2019). The author contends that this list may be incorrect. Reliable, aggregated, and up to date sources are not readily available.

therefore crucial that the global majority initiate collaborative efforts to protect their own data privacy interests. Africa should regulate how and when African data may be shared to safely support the interests of her people in AI-related matters.

8.5 Conclusions and Recommendations

In the absence of AI-specific legislation, privacy laws are the only legal safeguards that apply to the development and implementation of AI. Rigid common-law prescripts on privacy and legislation specifically relevant to the right to data privacy stifle innovation in AI, resulting in an untenable and impractical situation. The overarching philosophy that "[p]rivacy promotes safe learning" (Anwar, 2021, p.772) should guide attempts to balance the ostensibly opposing interests inherent in the threats related to the processing of student data by algorithms and AIED agents and facilitating equitable learning experiences as a result thereof. While promulgating comprehensive jurisdiction-specific AI legislation is both critical and urgent, this approach is most likely not a sufficiently judicious regulatory approach to address the complexities of the use of AI data-processing agents at work within higher education systems. AIED-specific legislation and a domestic approach to AI development and regulation are thus crucial, as is set out below.

8.5.1 AIED-Specific Legislation as Ancillary Regulation

International calls for AIED-specific legislation and policies are mounting (Bond et al., 2024). This unique subfield of AI would be best served by a more nuanced approach to regulation. The urgency of this is illustrated by contrasting how we think about consumer and student surveillance. We acknowledge the potential harm that stems from commercial surveillance practices such as the scraping of publicly available information from the internet (Solove & Hartzog, 2024). These practices seem inherently dangerous, as they hold little or no benefit for data subjects. Yet the data collected from users of LMSs through inherent and insidious surveillance practices is of an even more personal, and thus potentially harmful, nature because it involves *inter alia* records of mental and cognitive (dis)abilities, and potentially health information. Global societies mostly regard

education as a key endeavour that advances humanity,[32] and rightly so. Sadly, universities' lax procurement practices (Scott & Gray, 2023) and legislatures' failure to act has shown that we are more likely to regard infringements on data privacy rights in the name of improved educational outcomes as being for 'the greater good'[33]. Specifically legislating AIED is essential, most importantly because doing so will expose the inherent dangers thereof to all potentially affected persons and institutions, as well as the public. Legislation will convert the moral and ethical obligations to protect users of AIED agents to a legal obligation enforceable by sanctions.

Enacting AIED-specific legislation within a given jurisdiction may take time and such a project could be undertaken as a subsequent, more nuanced phase of AI regulation. Enacting overarching AI-specific legislation at national or federal level is an essential interim measure. Here the EU's approach may provide inspiration, as it highlights the domain of education as one of several in which the implementation of AI agents could potentially engender great harm to individuals.

8.5.2 A Regional Approach to AI(ED) Development and Regulation

There is a growing call for African collaboration in both the development of AI agents and AI-related policies and regulations (AU Specialised Technical Committee, 2019; Musoni & Mtuze, 2023). Modifying thinking around AI to suit local contexts requires local sensitivities: "building trust means taking your people on the journey, so that they can internalise what these ideas mean, bring abstract principles to life in their own language and metaphors, and tell user stories they can inhabit" (Buckingham Shum, 2024). However, a context-specific, multi-disciplinary approach to developing AI(ED) laws for the African region could use the EU AI Act as a point of departure, but not a blueprint. Identifying and then adapting relevant aspects of this act into stipulations that are pertinent to and practically enforceable in Africa could serve as a useful first step. To effectively develop homegrown algorithms and AIED fit

32 One example of this perspective is encapsulated in the vision of the University of South Africa: "towards *the* African university shaping futures in the service of humanity" (Unisa, 2024).

33 See Czerniewicz and Cronin (2023) on the role of education "for good" in society.

for African students and educators, African EDM researchers need access to openly available African data sets (Maphosa & Maphosa, 2020). Collaborative African AI regulation should aim to strike a balance that allows such access, while protecting the interests of data subjects. Such a regulatory project could potentially inspire and influence similar majority-world initiatives.

Maphosa and Maphosa (2020) argue that it is not yet clear how higher education institutions must respond to the legal complexities related to the collection and use of student data. It is, however, clear that these institutions have a fiduciary and moral duty to defend student data privacy (Prinsloo & Slade, 2017).

8.5.3 Final Remarks

Educators, researchers, and actors in the sphere of LA, EDM, and AIED who represent the global majority have a duty to help safeguard students from harm stemming from the infringement of their right to data privacy, the future implications of which cannot be fully known. This is essential, as Africa and other majority-world regions are specifically at risk of gross mass data privacy infringements by profiteers. While pedagogical and technological progress is most certainly desired, the achievement thereof should not steamroller fundamental rights. The weighing up of rights, duties, benefits, and risks should be done with diligent consideration to protect our students from potential harm as best we can. While legislation is never perfect, and the enforcement thereof often complex, legal prescripts are more concrete than ethical or moral guidelines. It is essential to pass general AI-specific, and later AIED-focussed, legislation at national or federal level. Regional cooperation throughout the majority world could significantly bolster these endeavours, ensuring that education continues to safely support individual and global development.

8.6 References

African Union Specialised Technical Committee on Communication and Information Technologies. (2019). *2019 Sharm El Sheikh Declaration*. Retrieved from https://au.int/en/decisions/2019-sharm-el-sheikh-declaration-stc-cict-3.

Alkahtani, H., Aldhyani, T. H. H., & Alqarni, A. A. (2024). Artificial intelligence models to predict disability for mental health disorders. *Journal of Disability Research*, *3*(3), 1-12. doi:10.57197/JDR-2024-0022.

AmaBhungane Centre for Investigative Journalism NPC v Minister of Justice and Correctional Services [2021] ZACC 3. Retrieved from https://www.saflii.org/cgi-bin/disp.pl?file=za/cases/ZACC/2021/3.html&query=amabhungane%20near%202021%20near%20zacc%20near%203.

Anwar, M. (2021). Supporting privacy, trust, and personalization in online learning. *International Journal of Artificial Intelligence in Education*, *31*, 769-783. doi:10.1007/s40593-020-00216-0.

Bernhardt, M. (2023). *AI's increasingly important role in L&D*. The Learning Guild, New York, N.Y. Retrieved from https://www.learningguild.com/publications/183/ais-increasingly-important-role-in-ld/.

Bhagattjee, P., Govuza, A., & Sebanz, L. (2020). Regulating artificial intelligence from a data protection perspective — lessons from the EU. *Without Prejudice*, December 2020, 9-10.

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., … Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(4), 1-41. doi:10.1186/s41239-023-00436-z.

Bozkurt, A. (2019). The historical development and adaptation of open universities in Turkish context: Case of Anadolu University as a giga university. *International Review of Research in Open Distance Learning*, *20*(4), 36-59.

Bozkurt, A., Xiao, J., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., … Jandrić, P. (2023). Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, *18*(1), 1-78. Retrieved from http://www.asianjde.com/ojs/index.php/AsianJDE/article/view/709/394.

Bronstein, V. (2022). Prioritising command-and-control over collaborative governance: The role of the Information Regulator under the Protection of Personal Information Act. *Potchefstroom Electronic Law Journal*, *25*, 1-41. doi:10.17159/1727-3781/2022/v25i0a11661.

Buckingham Shum, S. (2024, January 15). Co-designing AI ethics in education [Blog post]. Retrieved from https://simon.buckinghamshum.net/2024/01/codesigning-ai-ethics-edu/.

Campbell-Stephens, R. M. (2021). *Educational leadership and the global majority*. doi:10.1007/978-3-030-88282-2.

Chen, T., Tachmazidis, I., Batsakis, S., Adamou, M., Papadakis, E., & Antoniou, G. (2023). Diagnosing attention-deficit hyperactivity disorder (ADHD) using artificial intelligence: A clinical study in the UK. *Front Psychiatry*, *14*:1164433, 1-13. doi:10.3389/fpsyt.2023.1164433.

Constitution of the Republic of South Africa, 1996. Retrieved from https://www.saflii.org/content/Constitution-of-the-Republic-of-South-Africa_1996.html.

Czerniewicz, L., & Cronin, C. (Eds.). (2023). *Higher education for good: Teaching and learning futures*. doi:10.11647/obp.0363.27.

De Vries, I. (2019). Open universities and open educational practices: A content analysis of open university websites. *International Review of Research in Open Distance Learning*, *20*(4), 167-178.

Du Boulay, B. (2023). Artificial intelligence in education and ethics. In O. Zawacki-Richter & I. Jung (Eds.), *Handbook of Open Distance and Digital Education* (pp. 93-108). doi:10.1007/978-981-19-2080-6 pp.93-108.

Gwagwa, A., Kraemer-Mbula, E., Rizk, N., Rutenberg, I., & De Beer, J. (2020). Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions. *African Journal of Information and Communication*, *26*, 1-28.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., .. Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*, 504-526. doi:10.1007/s40593-021-00239-1.

Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2024). *Navigating LLM ethics: Advancements, challenges, and future directions*. Retrieved from https://arxiv.org/pdf/2406.18841.

Jones, J. (2018). *7 largest universities in the world*. Retrieved from https://largest.org/misc/universities/.

Karanicolas, M. (2023). Developing AI standards that serve the majority world. In L. Belli & W. B. Gaspar (Eds.), *The quest for ai sovereignty, transparency and accountability: Official outcome of the UN IGF Data and Artificial Intelligence Governance Coalition* (pp. 265-282). Retrieved from https://direitorio.fgv.br/publicacao/quest-ai-sovereignty-transparency-and-accountability.

Karanicolas, M. (2024). Challenging minority rule: Developing AI standards that serve the majority world. *UCLA Law Review DisC*, *71*, 196-213.

Li, W., Sun, K., Schaub, F., & Brooks, C. (2022). Disparities in students' propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education*, *32*, 564-608. doi:10.1007/s40593-021-00254-2.

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review*, *46*(5), 31-40.

Mahomed, S. (2018). Healthcare, artificial intelligence and the Fourth Industrial Revolution: Ethical, social and legal considerations. *South African Journal of Bioethics and Law*, *11*(2), 93-95. doi:10.7196/SAJBL.2018.v11i2.664.

Maphosa, V., & Maphosa, M. (2021). The trajectory of artificial intelligence research in higher education: A bibliometric analysis and visualisation. *2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, *IEEE*, 1-7. doi:10.1109/icabcd51485.2021.9519368.

McGowan, C., Paris, B., & Reynolds, R. (2024). Educational technology and the entrenchment of "Business as usual". *Academe*, *110*(1). Retrieved from https://www.aaup.org/article/educational-technology-and-entrenchment-%E2%80%9Cbusiness-usual%E2%80%9D.

Moodle. (2024). *Statistics*. Retrieved from https://stats.moodle.org/.

Musoni, M. & Mtuze, S. (2023). An Assessment of the Key AI Sovereignty Enablers within the South African Context. In L. Belli & W. B. Gaspar (Eds.), *The Quest for AI Sovereignty, Transparency and Accountability: Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition* (pp. 45-58). Retrieved from https://direitorio.fgv.br/publicacao/quest-ai-sovereignty-transparency-and-accountability.

Pelletier, K., Robert, J., Muscanell, N., McCormack, M., Reeves, J., Arbino, N., ... Zimmern, J. (2023). *EDUCAUSE Horizon Report, Teaching and Learning Edition*. Retrieved from https://library.educause.edu/resources/2023/5/2023-educause-horizon-report-teaching-and-learning-edition.

Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(22), 1-13. doi:10.1186/ s41039- 017- 0062-8.

Prinsloo, P., & Kaliisa, R. (2022). Data privacy on the African continent: Opportunities, challenges and implications for learning analytics. *British Journal of Education Technology*, *53*, 894-913. doi:10.1111/bjet.13226.

Prinsloo, P., & Slade, S. (2015). Student privacy self-management: implications for learning analytics. *Proceedings of the LAK '15 Fifth International Conference on Learning Analytics and Knowledge, ACM*, 83–92. doi:10.1145/2723576.

Prinsloo, P., & Slade, S. (2017). Big Data, Higher Education and Learning Analytics: Beyond Justice, Towards an Ethics of Care. In B.K. Daniels (Ed.), *Big Data and Learning Analytics in Higher Education* (pp. 109-124). Cham, Switzerland: Springer.

Protection of Personal Information Act 4 of 2013. Retrieved from https://www.saflii.org/cgi-bin/disp.pl?file=za/legis/num_act/popia2013380/popia2013380.html&query=protection%20near%20of%20near%20personal%20near%20information%20near%20act.

Quayyum, A., & Zawacki-Richter, O. (2019). The state of open and distance education. In O. Zawacki-Richter & A. Qayyum. (Eds.), *Open and distance education in Asia, Africa and the Middle East* (pp. 125-140). doi: 10.1007/978-981-13-5787-9_14.

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) (Pp. 1-88).

Regulation (EU) 2021/0106 of the European Parliament and of the Council 21 April 2021 on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (AI Act) (Pp. 1-108).

Scott, M., & Gray, B. C. (2023). Who cares about procurement? In L. Czerniewicz & C. Cronin (Eds.). *Higher education for good: Teaching and learning futures*. pp. 603-621. doi:10.11647/obp.0363.27.

Slade, S., & Prinsloo, P. (2014). Student perspectives on the use of their data: Between intrusion, surveillance and care. In *Challenges for research into open & distance learning: Doing things better — Doing better things* (pp. 291–300). Retrieved from https://oro.open.ac.uk/41229/1/BRPA_Slade_Prinsloo.pdf.

Slade, S., Prinsloo, P., & Khalil, M. (2019). Learning analytics at the intersections of student trust, disclosure and benefit. *Proceedings of the LAK '19 Ninth International Conference on Learning Analytics and Knowledge*, ACM, 1-10. doi:10.1145/3303772.3303796.

Solove, D. J., & Hartzog, W. (2024). The Great Scrape: The clash between scraping and privacy. *California Law Review*, *113* (forthcoming 2025). Advance online publication. doi:10.2139/ssrn.4884485.

Thaldar, D., & Malekela, M. (2024). Data protection law: Lessons from Tanzania for South Africa? *South African Journal of Bioethics and Law*, *17*(2), 57-58. Retrieved from https://samajournals.co.za/index.php/sajbl/article/view/2301/1072.

Unisa. (2024). *Who we are*. Retrieved from https://www.unisa.ac.za/sites/corporate/default/About/Who-we-are?lang=01.

Willis, J. E., Slade, S., & Prinsloo, P. (2016). Ethical oversight of student data in learning analytics: A typology derived from a cross-continental, cross-institutional perspective. *Educational Technology Research and Development*, *64*, 881-901. doi:10.1007/s11423-016-9463-4.

Xu, H., Luo, X., Caroll, J. M., & Rosson, M. B. (2011). The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing. *Decision Support Systems*, *51*(1), 42-52. doi:10.1016/j.dss.2010.11.017.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education — Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(39), 1-28. doi:10.1186/s41239-019-0171-0.

Zhang, W., & Li, W. (2019). Transformation from RTVUs to open universities in China: Current state and challenges. *International Review of Research in Open Distance Learning*, *20*(4), 1-20.

9 Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law: a Commentary

Ekaterina Martynova

Abstract

This paper provides a brief commentary on the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, and considers its possible impact on the AI regulation in the third states. It analyses the general characteristics of the Convention: its legal nature, object and purpose, along with specific issues relating to the procedure for the use of remedies, the process of accession, and the mechanisms of implementation of the Convention at the national level. The commentary concludes by highlighting the provisions and approaches of the Convention that could be useful in shaping national AI regulation, as well as common regulatory framework on the BRICS platform. Such provisions include, inter alia, standards of transparency, reliability, risk assessment, accountability and responsibility for negative consequences, as well as remedies for the individuals whose rights have been violated by the use of AI systems.

Keywords: artificial intelligence, human rights, Council of Europe, international treaty.

Introduction

The emerging field of international legal regulation of artificial intelligence (hereinafter — the AI) results in the interaction of various sources, such as private agreements (often, market-driven — Chinen, 2023) made by corporations, regulations set by individual states, the body of international law itself, recommendations of international organizations and civil society, which predefines formulation of the normative framework through a range of approaches, from voluntary agreements to formal regulations. Among them, the

Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (hereinafter — the Convention), adopted on 17 May 2024 by the Committee of Ministers of the Council of Europe, is the first international legally binding treaty. Its purpose is to ensure respect for human rights, the rule of law and legal standards of democracy at all stages of the design, development and application of the AI systems.[34] There are three main objectives that the Convention aims to achieve: firstly, to address the problems in interpreting human rights in the context of AI application; secondly, to embed fundamental human rights principles in relation to AI; and thirdly, to establish international human rights norms on the application of AI to promote international trade (Van Kolfschooten, H.; Shachar, C, 2023). This paper provides a brief commentary on the Convention in the light of the stated objectives of its adoption. It first looks at the general characteristics of the Convention and the content of the obligations of signatory states. It then considers the remedies available to individuals whose human rights are allegedly violated in the context of the use of AI systems. The discussion concludes with a consideration of the possible implications of the Convention for third states and the prospects for the BRICS countries to learn from the experience of developing the Convention.

9.1 Discussion

9.1.1 Legal nature of the Convention, its object and purpose

The development of AI systems is welcomed and encouraged by states because of the vast opportunities that AI offers to improve other technologies, industrial growth and the intensification of trade. However, the use of AI has political, social and economic implications for various social relations, both nationally and internationally, that go beyond the legal regime regulating AI as a technology (Crawford, 2022, pp. 185-186). The adoption of the Convention as an international treaty is intended to establish a legal framework that will respond to the new challenges that the international community

34 The Convention, Article 1(1).

and individual states face in connection with the development of AI, in particular with regard to the functioning of democratic institutions (Nemitz, 2018), the protection of rights and freedoms (Donahoe, E., & Metzger, M. M., 2019), and the overcoming of social inequalities and discrimination arising from the use of such computer programs (Eubanks, V., 2018).

The term 'AI system' is defined in the Convention as "a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments" with indication that "[d]ifferent artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment".[35] Thus, the Convention has adopted a so-called 'broad' approach to defining an AI system based on its self-learning and generative abilities, as opposed to a 'narrow' approach to defining AI based on the ability of a system to solve a specific applied task such as translation services or chatbots.[36]

The definition provided in the Convention does not contain an indication of the possible dual-use (military and civilian) nature of AI systems. Herewith, it is important to note that activities related to national defence are excluded from the scope of the Convention.[37] Thus, the obligations of Parties to the Convention to ensure transparency, accountability and responsibility for possible adverse effects do not apply to the activities related to the development or application of AI systems for military purposes. At the same time, the use of AI in defence as an autonomous lethal weapon system has the potential to seriously affect the geopolitical balance between states, creating new international asymmetries (Johnson, 2019).

35 The Convention, Article 2.

36 Despite the ubiquitous nature of AI discussions lately, there is no consistent 'official' definition of AI. In some cases, the technical descriptions offered by computer scientists are not suitable for legal analysis, for example when AI is defined in terms of an 'algorithm', which in turn requires a separate definition and understanding of the social meaning and legal content. For the review of different approaches to define AI for the purposes of legal studies, refer, e.g., to Lee, J. (2022:6-8). On the 'broad' and 'narrow' approach to defining AI see, e.g., Meltzer, J. P. (2018, December 13). The impact of artificial intelligence on international trade. *Brookings*. Retrieved from https://www.brookings.edu/research/the-impact-of-artificial-intelligence-on-international-trade/#footnote-1.

37 The Convention, Article 3(4).

With regard to the scope and application of the Convention, states Parties to the Convention are expected to take the legislative, administrative or other measures necessary to ensure compliance with the provisions of the Convention by both public authorities and private actors acting on their behalf.[38] The Convention provides an alternative means of regulating private actors not acting on behalf of a state: Parties may extend the principles and obligations set out in the Convention to the private sector (thereby putting it on an equal regulatory footing with the public sector) or take other appropriate measures to manage the risks of the use of AI systems by private actors in a manner consistent with the object and purpose of the Convention.[39] The chosen method of fulfilling the obligation to regulate the private sector shall be communicated at the time of signing or depositing the instrument of ratification, acceptance, approval or accession to the Convention (the chosen method can be subsequently changed). There can be no derogation from or limitation on the application by a Party of its international obligations relating to human rights, democracy and the rule of law.[40] This 'fork in the road' in the methods of regulating the private sector does not seem to be entirely appropriate, as it creates an imbalance in the scope of the obligations of the Parties to the Convention, depending on the option chosen.

9.1.2 The main obligations of the Parties to the Convention

As a general comment, it should be noted that, although the Convention enumerates the obligations of states Parties, certain 'saving clauses' anticipate its framework nature. Thus, as a general principle of regulation, it is stipulated that each Party shall fulfil its obligations under the Convention "in a manner appropriate to its domestic legal system".[41] In the text of the Convention, states' obligations are formulated as 'soft' goals and obligations of conduct rather than obligations of result. On the one hand, this approach ensures flexibility in the application of the Convention and is likely to

38 The Convention, Articles 1(2) and 3(1)(a).

39 The Convention, Article 3(1)(b).

40 Ibid.

41 The Convention, Article 6.

increase the number of the Parties, but on the other hand, it 'blurs' the content of the states' obligations and leaves significant room for interpretation. At the same time, according to Martti Koskenniemi, a possible shift in the balance between normativity and certainty towards normativity will inevitably lead to inconsistency in practice and thus to the politicisation of the relevant regulation (Koskenniemi, 2006). In particular, the obligation of the Parties to take measures aimed at protecting democratic processes in the lifecycle of AI systems, including ensuring the "ability to freely form opinions", as set out in the Convention,[42] may be implemented in significantly different ways by states, depending on the chosen approach to regulating social networks.

The Convention does not establish strict requirements for the adoption of specific measures, hence the provisions of this international treaty are largely non-self-executing. This distinguishes the Convention from the European Union AI Act[43] which applies directly on the territory of all EU Member States and creates very specific positive obligations of the Member States with certain deadlines e.g., to establish rules for penalties and enforcement measures, including warnings and non-monetary actions, that can be applied to operators who violate the Act's regulations;[44] to introduce laws, regulations or administrative provisions, more favourable to workers in terms of protecting their rights in respect of the use of AI systems by employers;[45] and to introduce, in accordance with EU law, restrictive laws on the use of post-remote biometric identification systems.[46]

The Convention establishes an obligation for the Parties to implement "adequate transparency and oversight requirements" for the lifecycle activities of AI systems, including the identification of content generated by such systems, taking into account specific contexts

42 The Convention, Article 5(2).

43 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. PE/24/2024/REV/1 // *OJ L, 2024/1689, 12.7.2024.*

44 Ibid, Recital 168 / 179 and Article 99, 113.

45 Ibid, Recital 23 and Article 2(11).

46 Ibid, Recital 96 and Article 27(10).

and risks.[47] This commitment is linked with the requirement to ensure accountability and responsibility for possible adverse impacts of AI systems on human rights, democracy and the rule of law.[48] Precise scope of relevant standards is defined by the state Parties themselves.

Parties to the Convention shall also take measures to ensure that AI systems respect equality, including gender equality, as well as the prohibition of discrimination and do not violate privacy rights of individuals.[49] At the same time, the relevant articles of the Convention include reservations that such obligations shall be implemented by states taking into account international and national law. It appears that the actual content of these obligations in states of different legal systems may vary significantly, in particular with regard to the gender equality and approaches to the grounds for permissible restrictions on the right to privacy. In general, it can be assumed that the choice and 'calibration' of the instruments laid down in the Convention and their implementation in the national regulation of AI will be determined by the specificities of the political regime of the state Party to the Convention: in particular, the degree of involvement of stakeholders in the process of normative regulation, the effectiveness of institutions that determine the rules of behaviour of participants in the life cycle of AI systems, as well as the role of civil society and organisations for the protection of human rights and freedoms.[50]

9.1.3 Remedies

Parties to the Convention are obliged to ensure that remedies are available to persons whose human rights have been violated in connection with the use of AI systems (again, with the proviso that such measures are taken to the extent that they comply with the requirements of the domestic legal system).[51] As basic procedural safeguards for the protection of human rights when interacting with AI systems, the Convention provides for notification to any persons of their interaction with AI systems, documentation of information

47 The Convention, Article 8.

48 The Convention, Article 9.

49 The Convention, Articles 10, 11.

50 For an overview of national AI policy regimes and their typology by political regime, see: Filgueiras, F. (2022).

51 The Convention, Article 14(1).

about the use of AI systems that potentially violates human rights, and the possibility for interested persons to access such information and lodge a complaint with the competent public authority.[52] However, the precise approach to be taken to answer the question of whether the rights of applicants have been violated is left outside the scope of this international treaty, leaving room for a variety of models. Thus, the Convention merely establishes a general procedural vector by guaranteeing the availability of a remedy.

9.1.4 Exemptions for national security and scientific research purposes

Parties to the Convention are exempt from compliance with their obligations when carrying out activities related to the defence of national security interests, provided that such activities are carried out without violating international law, including international human rights obligations, and with respect for democratic institutions and processes.[53] The motivation for this exception is obvious, but its danger is that states may apply it broadly, without providing an explanation of the reasons for applying the exception, on the grounds that the mere explanation of the reasons poses a threat to national security. The Convention also does not restrict the Parties from conducting research and development activities regarding AI systems, provided that such activities do not involve risks to human rights, democracy or the rule of law.[54]

9.1.5 Oversight mechanisms

In order to ensure the effective implementation of the provisions of the Convention, a monitoring mechanism is established in the form of a Conference of the Parties with advisory powers.[55] In addition, Parties to the Convention undertake to establish their own independent mechanism to oversee compliance with the Convention and assess risks of human rights violations, to take measures to raise public awareness, encourage informed public debate and consultation

52 The Convention, Article 14(2).

53 The Convention, Article 3(2).

54 The Convention, Article 3(3).

55 The Convention, Article 23.

with all stakeholders on the use of AI systems,[56] as well as to send periodic reports on progress in the implementation of the Convention for consideration by the Conference of the Parties.

9.1.6 Signature procedure, possibility of reservations

The Convention is open for signature by the Member States of the Council of Europe, the European Union and the states that participated in its elaboration (the states whose representatives participated in the work of the Committee on Artificial Intelligence: Argentina, Australia, Canada, Costa Rica, the Holy See, Israel, Japan, Mexico, Peru, the Unites States and Uruguay). The signing took place in Vilnius, Lithuania, on 5 September 2024 during the Conference of Ministers of Justice. Andorra, Georgia, Iceland, Norway, the Republic of Moldova, San Marino, the United Kingdom, Israel, the United States of America as well as the European Union have signed the Convention (Council of Europe, 2024, September 13). Once the Convention enters into force (on the first day of the month following the expiry of a period of three months after the date of ratification of the Convention by five signatories, including at least three member states of the Council of Europe), states that did not participate in its drafting will be able to accede to it, provided that such accession is approved by a decision adopted by a majority in accordance with Article 20.d of the Statute of the Council of Europe and by a unanimous vote of the representatives of the parties to the Convention entitled to sit on the Committee of Ministers.[57] This strict procedure for accession to the Convention is not unique to Council of Europe treaties: most of them, including the so-called 'open' treaties, i.e. allowing accession by non-Council of Europe member states, require the unanimous consent of the parties.[58]

9.1.7 Possible consequences of the adoption of the Convention for the third parties

The Convention does not impose any obligations on states not parties to it. Even though the Convention contains a declaratory

56 The Convention, Articles 16, 19, 20.

57 The Convention, Article 31(1).

58 Participation of Non-member States. (2023, October 7). Retrieved from https://www.coe.int/ru/web/conventions/participation-of-non-member-states.

provision on the Parties' endeavour to encourage non-parties to act consistently with its principles,[59] this provision does not create any normative obligations.

To date, none of the BRICS countries have signed the Convention. The accession of the Russian Federation, which ceased to be a member of the Council of Europe in September 2022, to the Convention is unlikely to be an issue in the near future, including due to the unanimous vote required of the representatives of the parties to the Convention entitled to sit on the Committee of Ministers for accession by a non-member state of the Council of Europe. At the same time, the two-year experience of the Committee on Artificial Intelligence in drafting the text of the Convention and some of its provisions may be useful in the formation of national and international normative regulation of activities using AI systems, particularly at the BRICS level. Specifically, the principles of transparency, reliability, risk assessment, accountability and responsibility for negative consequences seem to be the most important foundations for the activities of public authorities and private actors within the lifecycle of AI systems. Guarantees of information to citizens, as provided for in the Convention, may also be perceived by other legal systems for example, in the form of labelling of the content generated by an AI system and information that an interaction with an AI system is taking place (for example, when a consumer receives services via a telephone call). In addition, an analysis of the practice of states Parties to the Convention in providing remedies to citizens whose rights are violated by the use of AI systems may be useful for improving other states' national legislation, especially in areas that are sensitive from the perspective of protecting citizens' rights, such as the use of facial recognition systems.

9.2 Conclusion

As noted above, the framework nature of the Convention has influenced the formulation of the obligations assumed by states. The Convention does not lay down strict requirements for the adoption of specific measures. As a result, the provisions of this

59 The Convention, Article 25(1).

international treaty are largely non-self-executing, and the nature of the obligations set forth in the Convention gives states wide discretion in their implementation. Moreover, the broad discretion of states in implementing this treaty is reflected in the right of states Parties to determine the extent to which the Convention applies to the development and use of AI systems in the private sector. The Convention thus embodies a 'soft' model of international legal regulation in the field of AI. At the same time, the adoption of the rules enshrined in the Convention will certainly be a positive incentive for the development of domestic legislation regulating the development and use of AI systems. Moreover, this soft regulatory approach seems to be a possible first step towards developing a common regulatory framework at the international level among states with less integrated legal systems compared to, for example, the European Union. In this sense, this model could be considered for the development of similar international legal instruments, in particular on the BRICS platform.

9.3 References

Chinen, M. (2023). The international governance of artificial intelligence. Edward Elgar Publishing.

Council of Europe. (2024, September 13). Council of Europe opens first ever global treaty on AI for signature. *Portal*. Retrieved from https://www.coe.int/en/web/portal/-/council-of-europe-opens-first-ever-global-treaty-on-ai-for-signature.

Crawford, K. (2022). Atlas of AI: power, politics, and the planetary costs of artificial intelligence. New Heaven: Yale University Press, 2021.

Donahoe, E., & Metzger, M. M. (2019). Artificial Intelligence and Human Rights. Journal of Democracy, 30(2), 115–126. https://doi.org/10.1353/jod.2019.0029.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. Retrieved from https://openlibrary.org/books/OL26681102M/Automating_Inequality.

Filgueiras, F. (2022). Artificial Intelligence Policy Regimes: Comparing Politics and Policy to National Strategies for Artificial Intelligence. Global Perspectives, 3(1). https://doi.org/10.1525/gp.2022.32362.

Johnson, J. (2019). Artificial intelligence & future warfare: implications for international security. Defense and Security Analysis, 35(2), 147–169. https://doi.org/10.1080/14751798.2019.1600800.

Johnson, J. (2019). Artificial intelligence & future warfare: implications for international security. *Defense and Security Analysis*, *35*(2), 147–169. https://doi.org/10.1080/14751798.2019.1600800.

Koskenniemi, M. (2006). From Apology to Utopia: The Structure of International Legal Argument. Cambridge University Press.

Lee, J. (2022). Artificial intelligence and international law. Singapore: Springer.

Meltzer, J. P. (2018, December 13). The impact of artificial intelligence on international trade. Brookings. Retrieved from https://www.brookings.edu/research/the-impact-of-artificial-intelligence-on-international-trade/#footnote-1.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences, 376(2133), 20180089. https://doi.org/10.1098/rsta.2018.0089.

Van Kolfschooten, H., & Shachar, C. (2023). The Council of Europe's AI Convention (2023–2024): Promises and pitfalls for health protection. Health Policy, 138, 104935. https://doi.org/10.1016/j.healthpol.2023.104935.

10 Human capacity (ability)-centred AI policy: Eurasian and Transatlantic safety dialogue

Yonah Welker

Abstract

The Bletchley Declaration was signed by 28 countries that agreed on a risk-based approach to frontier AI models, including areas of social protection, health, education, labor. It involved African nations, such as Nigeria, Kenya and Rwanda, countries from the Middle East, including Saudi Arabia and the United Arab Emirates; and major Western economies, such as Canada and the US. Emerging AI policies and frameworks make an attempt to categorize AI systems based on risks, related compliance frameworks and explanations. Such mechanisms are aimed at both regulating and facilitating a human-centered approach to AI systems development, connecting stakeholders and broader society. However, existing approaches to understanding high and unacceptable-risk systems still miss disability-specific vocabulary, scenarios and associated risks, categorization of impairments, spectrums, actions and non-actions, and complex understanding of intersectionality behind it. It includes not only the areas of law enforcement, police, biometrical and public security systems, but less covered areas of silos, misuse or manipulation presented by autonomous systems.

Keywords: accessibility, disability, AI, safety, policy, ethics.

Introduction

There is an estimated 1 billion people — 15% of the world — live with disabilities (Disability and Employment, n.d.), according the World Health Organization (WHO). And 80% of those people live in developing countries. Historically, individuals with disabilities were excluded from the workplace, educational system, and sufficient medical support. For instance, around 50-80% of the population with disabilities are not employed full time, 50% of children with disabilities in low- and middle-income countries are still not enrolled in school, public spaces meet only 41.28% to 95% (Syaodih, Aprilesti;

2020) of the expectations of people with disabilities, and only 10% of the population have access to assistive technologies. For cognitive disabilities, the level of discrimination is even higher. The unemployment rate among those with autism may reach (Chen et al., 2015) 85%, dependent on the country; while among people with severe mental health disorders, it can be between (Brouwers, 2020) 68%-83%, and for those with Down's syndrome, 43%.

Along with exclusion, individuals with disabilities are disproportionally affected by unjust law enforcement, violence and brutality. Persons with disabilities were victims of 26% of all nonfatal violent crime. 30-50% of individuals subject to the use of force or killed by police have a disability. People with intellectual disabilities are seven times more likely to be sexually assaulted than members of the general population. About one-third of young children and teenagers with disabilities faced emotional and physical abuse.

As for conflicts and crises, people with disabilities are also recognized as among the most marginalized and at-risk population. An estimated 9.7 million people with disabilities are forcibly[60] displaced as a result of conflict and persecution and are victims of human rights violations and conflict-related violence. As a result, these groups are also more affected by posttraumatic disorders and conditions.

Finally, there is a strong component of intersectionality behind disabilities that may amplify this exclusion and discrimination, including aspects of demography, co-occurring conditions and socioeconomic factors. For instance, individuals with learning disabilities also experience mental health problems, with estimates suggesting that between 25 and 40% ("Learning disability statistics", 2016) fall into this category. Girls are often diagnosed at a much lower rate than boys, with a ratio of 4:1, and may also be misdiagnosed due to different manifestations. Certain ethnic and social groups (Donohue et al., 2021) have been historically excluded from research data and resources.

60 https://www.hrw.org/news/2018/12/03/un-wars-impact-people-disabilities.

10.1 AI Systems and Disability Support

It's important to highlight that ethically developed and implemented assistive technologies can eliminate particular social barriers and create more accessible workplaces, hiring and learning experiences, and accommodation practices.

For instance, in order to support physical impairments, AI algorithms can be used to augment smart wheelchairs (Rahimunnisa, 2024), walking sticks (Guo et al., 2021), geolocation and city tools, bionic and rehabilitation technologies. In the case of sensory impairments, it includes facial and sign recognition for sign language identification and support of deaf individuals (Adeyanju, 2021), and computer vision algorithms that can interpret images and videos and then translate that information into braille or audio output to help individuals with visual impairments.

In the area of cognitive impairments, it includes social robotics and algorithms for emotional training for students with autism (Kouroupa, 2022), wearables and devices that improve emotion recognition (Haber et al., 2020), and adaptive platforms that support dyslexia and attention deficit and hyperactivity disorders. Such technologies can serve to support the general population as well, including further advancement of healthcare, education, labor and city systems, and support of elders, neurodisabled groups and individuals with psycho-emotional disorders.

10.2 Data, Models and Errors of autonomous systems

Algorithms do not create biases themselves but perpetuate societal inequities and distortions. The reasons behind it include lack of access to data for target populations, the models trained to demonstrate efficiency for broader objectives, but lacking accuracy for specific groups or conditions, historical exclusion from research and statistics, simplification and generalization of the target group's parameters (proxies), subjectiveness introduced to labelled data or models' objectives.

For instance, AI systems are known to be less accurate towards individuals with facial differences or asymmetry, different gestures, gesticulation, speech impairment, or different communication

patterns. It especially affects groups with physical disabilities ("Disability, Bias, and AI", 2019), cognitive and sensory impairments, and autism spectrum disorders. There are examples of direct life-threatening scenarios when police and autonomous security systems (Figueroa et al. 2022), or military AI may falsely recognize assistive devices as a weapon or dangerous objects, or misidentify facial or speech patterns. These concerns were raised by UN Special Rapporteur[61] on the Rights of Persons with Disabilities, disability organizations such as EU Disability Forum.

There are a variety of physical, cognitive and social parameters that may lead to errors or inaccuracies towards individuals with disabilities. These errors can be grouped into several categories, including recognition, identification and cues, aids, semantic errors:

- Assistive tools and devices — individuals with disabilities may use a wheelchair, walking stick, rehabilitation or assistive devices, bionic hands or legs, or other tools and devices of different shapes, forms and patterns that may not be properly recognized by autonomous systems;
- Assistance and users — solutions, addressing individuals with disabilities frequently involves not only one end-user but an "ecosystem" of users, such as family members, and caregivers. For instance, specialized solutions for autism frequently involve two interfaces — one for the parent, and one — for the child. Public and city systems may not take it into consideration;
- Physical impairments. A person with a disability may lack particular limbs, or have different body shape, posture, and movement pattern, making it more difficult for proper recognition;
- Visual impairments. Blind persons and those with a visual impairment may not not properly understand visual cues given by automated systems;
- Hearing impairments. Individuals with hearing impairments may not hear and comply with audible commands or warnings, making it especially cautions for police and law- enforcement systems;

61 https://documents-dds-ny.un.org/doc/UNDOC/GEN/N22/433/14/PDF/N2243314.pdf?OpenElement.

- Speech impairments. Neurological conditions may affect speech and the ability to communicate, thus not meeting "typical" speech patterns;
- Cognitive impairments. Individuals with cognitive disabilities may communicate differently, lack emotional recognition or social skills;
- Behavioral and psychomotor patterns — individuals with disabilities may exhibit a different pattern of user behavior related to attention span, activities and cognitive parameters;
- Facial recognition that may not identify persons with eye deviation or facial neuropathy;
- Tactile recognition that is built on the assumption that everyone has hands, fingers, and fingerprints and has similar tactile parameters excludes many individuals with disabilities
- Semantic, intersectional, age and other biases — systems may add negative connotations to disability keywords for individuals of particular ethnicities. Besides, algorithms may perpetuate existing ageism (Stypinska, 2023).

Each parameter alone or in combination with others may lead to greater inaccuracies presented by autonomous systems.

These risks might be also affected by parameters of computing or physical chains. In particular, *supervised learning* (Packin, 2021), a category of machine learning that uses labeled datasets to train algorithms to predict outcomes and recognize patterns, is known for human-induced errors during the selection, labeling or existing in pretrained models (smart glasses and computer vision, visual objects), *unsupervised learning* — statistical lack of input, representation, raw data can reinforce social disparities and dismiss particular populations (e.g. DNA data clustering for medical solutions), *reinforcement learning* — environment driven errors, "problem of initial experience", experiment's limitations (e.g. learning based on a "reward system", social robotics and assistants). Data points may not exist for certain groups, identities or communities. People who collect or label data may introduce subjectiveness (reporting, selection, systemic or group attribution errors), lack evidence or access to target population. Errors can be also driven by model objectives and constraints.

As for physical chain, risks can be affected by physical human-robot Interaction, issues in balance and stability, durability and robustness, dexterity and haptic manipulation, motion and sensing components safety (servos and kinematics components related to the robots physical reliability and agility; touch, feedback, visual or voice sensors, 3D/depth cameras, LiDARs to collect data for mobility and task processing analysis, spatial intelligence), power components and environmental safety, quality of production and training cycle –planning and control, testing and simulation, sensing and perception.

10.3 Generative AI and language models — opportunities and risks

Generative AI and language-based models further expand this impact and the R&D behind it. In particular, such systems may fuel existing assistive ecosystems, health, work, learning and accommodation solutions, requiring communication and interaction with the patient or student, social and emotional intelligence and feedback. Such solutions are frequently used in areas involving cognitive impairments, mental health, autism, dyslexia, attention deficit disorder and emotion recognition impairment, which largely rely on language models and interaction.

With the growing importance of web and workplace accessibility, Generative AI-based approaches can be used to create digital accessibility solutions, associated with speech-to-text or image-to-speech conversion. It may also fuel accessible design and interfaces involving adaptive texts, fonts and colors benefiting reading, visual or cognitive impairments. Similar algorithms can be used to create libraries, knowledge and education platforms that may serve the purpose of assistive accommodation, social protection and micro-learning, equality training and policing. Finally, approaches explored through building such accessible and assistive ecosystems may help to fuel the assistive pretext — when technologies created for groups with disabilities can be later adapted for a broader population, including fueling new forms of interaction, learning and creativity, involving biofeedback, languages and different forms of media.

When compared to existing AI systems, however, language-based platforms require even more attention and ethical guidance. In particular, they can imitate human behavior and interaction, involve more autonomy and pose challenges in delegating decision-making. They also rely on significant volumes of data, a combination of machine-learning techniques and the social and technical literacy behind it.

There are different ways, in which generative AI-associated systems (Urbina, 2024) may pose risks for individuals with disabilities. In particular:

- They may fuel bias in existing systems, such as automated screening and interviews, public services involving different types of physical and digital recognition and contextual and sentiment bias.
- They may lead to manipulative scenarios, cognitive silos and echo chambers. For instance, algorithms were used to spread misinformation among patients during the COVID-19 pandemic.
- Language-based systems (Glazko, 2024) may add a negative connotation to disability-related keywords and phrases or provide wrong outcomes due to a public data set containing statistical distortions or wrong entries.
- Privacy — in some countries, governmental agencies were accused of using data from social media without consent to confirm patients' disability status for pension programmes.

10.4 Human-capacity Centered AI Policy and regional contexts

Addressing the AI policy towards groups with disabilities requires complex oversight and assessment. In particular, disability-centered deployment is *multimodal and multisensory* — it involves visual, hearing, cognitive parameters, necessity of accuracy for different modalities, *It's modular* — may involve interconnected devices and interfaces, It's *multistakeholde*r — it may involve families, caregivers. It also requires *Identifying misuse*, actions and non-actions (omissions), manipulation, addictive design, specific attention to data, models and systems oversight, privacy and consent.

For instance, AI-driven dashboards for children with cognitive disabilities may have 2 interfaces — one for the child and one for the parent, solutions can be *data-interconnected* (dashboards and interfaces for analytics and tracking, compact wearable trackers, smart glasses helping with recognition and learning, social assistants and companions). users can have tactile impairment, differences in the accuracy of color memory and search, sound and sight sensitivity)

As for the assistive systems regulation, some facial recognition systems (Benzaoui, 2023) used ear shape or the presence of ear canal to determine whether or not an image included a human face. However, this system didn't learn from sufficient patterns to recognize people who lack these parts or suffer craniofacial syndromes. Medical assessment and analytics systems are known to be created based on "normalized attributes" demographic and health groups. However, it may predominantly exclude some conditions or parameters for younger patients, attributing it only to older groups.

As for the medical data, in some countries immigrants with disabilities tend (Hacker, 2015) to *avoid medical examinations* and tests in fear to being deported or face high medical costs which lead to misrepresentation in available medical data sets. People with disabilities may have additional conditions and impairments which do not exist in data sets (e.g. allergies, digestive system disorders). Particular social groups *more likely report concerns* related to cognitive disabilities due to the better medical and educational access. Conditions affecting general population are presented with more sufficient evidence and statistics than rare genetic disorders. Infrastructure and urban datasets used for city planning are known to be "*gender-blind*", affecting accuracy of solutions for women patients.

This complex nature can be addressed through the combination of legal and policy frameworks. For instance, in the European Union disability cases and safety considerations are potentially affected in AI Act, Digital Services Act, data regulation and specific frameworks such as Accessibility Act.

- *Classifications and taxonomies* — Accessibility Act ("European accessibility act", n.d.) and Standardization directives (e.g. Regulation — 1025/2012)

- *Data profiling, manipulation, addictive design* — AI Act, DSA ("The EU's Digital Services Act", 2022), GDPR
- *Identifying "high-risks" for systems* related to certain critical infrastructures, medical devices, systems to determine access to educational, institutions or for recruiting people, law enforcement — AI Act
- *"Specific transparency risk".* AI systems such as chatbots or assistive companions should notify users that they are interacting with a machine — AI Act
- *Prohibiting particular use* of affective computing and emotion recognition for publicly accessible spaces — workplaces and educational institutions, law enforcement and migration — AI Act
- *Ensuring code of conduct* for minimal-risk systems, including accessibility ones which meet its requirements.

However, these complex efforts face several challenges at regional level.

- *Local AI solutions.* It's known that even the leading AI models (with 100, 400B
- fail in accuracy for non-English languages (Petrić Howe, 2024) or specific environments — indigenous populations, R&D, health, educational environments
- *Accessibility* –The World Health Organization (WHO) estimates that only 1 in 10 people ("Assistive technology", n.d.) have access to the assistive technology they need
- *Necessity of "Guardian" models* — specialized models addressing fairness and transparency-related features which may complement / track existing ones
- *Area specific literacy and frameworks*. Current efforts include Unesco — AI ethics frameworks and literacy in education ("What you need to know about UNESCO's new AI competency frameworks for students and teachers", n.d.), WHO — AI in health ("Ethics and governance of artificial intelligence for health", n.d.), OECD — disability, AI and labor markets ("Using AI to support people with disability in the labour market", n.d.), accidents repositories, UNDP's Digital Inclusion in a dynamic world).

- *Controlling vendors influence* — when the same companies invest in data centers and hyperscalers across regions ("Hyperscalers in crosshairs for anti-competitive pricing and lock-in", n.d.), creating data and market silos, limiting competition, and access
- *Other challenges* include digital and physical infrastructure (such as energy and water scarcity), limited cases and taxonomies, not reflecting the uniqueness of historical and social patterns for health and public solutions.

10.5 Way forward. Disability-centered policy, risks and impact assessment

Disability is not a monolith, but a spectrum, affected by underlying conditions, demographic, socio-economic and historical criteria. This complexity poses an important reminder that disability exclusion is a social issue first and only then — algorithmic. Existing AI policies and acts attempt to categorize and describe systems through primarily generalized visions of technologies, scenarios and posed risks. These categories do not address specific groups, physical or cognitive differences, unequal access to medical support or education, or economic status.

With more risks of emerging data silos and monopolization of AI development posed by corporate agents, there is an urgent need for collective action to address disability representation in policy development. It includes Introduction of *AI safety institutes, regulatory sandboxes* and testbeds (which involve units of regulation and compliance, accessible engineering and policy coordination), *risk-based categories* (unacceptable, high, low, minimum), *scenarios* (workplaces, education, law-enforcement, immigration), *specific systems considerations* (affective computing, biometrics), *systems taxonomies, frameworks and accidents repositories,* accessible digital and physical infrastructure, specialized policies and Minors Protection (e.g. 193 countries signed commitment to effectively implement children's digital safety — UN's General Assembly's Third Committee[62]).

62 "With Children at Higher Risk…", n.d.

10.6 References

Adeyanju, I. A., Bello, O. O., & Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, *12*, 200056. https://doi.org/10.1016/j.iswa.2021.200056.

Assistive technology. (n.d.). Retrieved 25 November 2024, from https://www.who.int/news-room/fact-sheets/detail/assistive-technology.

Benzaoui, A., Khaldi, Y., Bouaouina, R., Amrouni, N., Alshazly, H., & Ouahabi, A. (2023). A Comprehensive survey on ear recognition: Databases, approaches, comparative analysis, and open challenges. *Neurocomputing*, *537*, 236–270. https://doi.org/10.1016/j.neucom.2023.03.040.

Brouwers, E. P. M. (2020). Social stigma is an underestimated contributing factor to unemployment in people with mental illness or mental health issues: Position paper and future directions. *BMC Psychology*, *8*(1), 36. https://doi.org/10.1186/s40359-020-00399-0.

Chen, J. L., Leader, G., Sung, C., & Leahy, M. (2015). Trends in Employment for Individuals with Autism Spectrum Disorder: A Review of the Research Literature. *Review Journal of Autism and Developmental Disorders*, *2*(2), 115–127. https://doi.org/10.1007/s40489-014-0041-6.

Disability and Employment. (n.d.). UN DESA. Retrieved 25 November 2024, from https://www.un.org/development/desa/disabilities/resources/factsheet-on-persons-with-disabilities/disability-and-employment.html.

Disability, Bias, and AI. (2019). NYU Center for Disability Studies. https://disabilitystudies.nyu.edu/disability-bias-and-ai-report/.

Donohue, M. R., Childs, A. W., Richards, M., & Robins, D. L. (2019). Race influences parent report of concerns about symptoms of autism spectrum disorder. *Autism*, *23*(1), 100–111. https://doi.org/10.1177/1362361317722030.

Ethics and governance of artificial intelligence for health. (n.d.). Retrieved 25 November 2024, from https://www.who.int/publications/i/item/9789240029200.

European accessibility act. (n.d.). Retrieved 25 November 2024, from https://ec.europa.eu/social/main.jsp?catId=1202.

Figueroa, M., Orozco, A., Martínez, J., & Jaime, W. (2022, November 30). *The risks of autonomous weapons: An analysis centred on the rights of persons with disabilities*. International Review of the Red Cross. http://international-review.icrc.org/articles/the-risks-of-autonomous-weapons-analysis-centred-on-rights-of-persons-with-disabilities-922.

Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., & Mankoff, J. (2024). *Identifying and Improving Disability Bias in GPT-Based Resume Screening* (No. arXiv:2402.01732). arXiv. https://doi.org/10.48550/arXiv.2402.01732.

Guo, X., He, T., Zhang, Z., Luo, A., Wang, F., Ng, E. J., Zhu, Y., Liu, H., & Lee, C. (2021). Artificial Intelligence-Enabled Caregiving Walking Stick Powered by Ultra-Low-Frequency Human Motion. *ACS Nano*, *15*(12), 19054–19069. https://doi.org/10.1021/acsnano.1c04464.

Haber, N., Voss, C., & Wall, D. (2020, March 26). *Upgraded Google Glass Helps Autistic Kids "See" Emotions*. IEEE Spectrum. https://spectrum.ieee.org/upgraded-google-glass-helps-autistic-kids-see-emotions.

Hacker, K., Anies, M., Folb, B. L., & Zallman, L. (2015). Barriers to health care for undocumented immigrants: A literature review. *Risk Management and Healthcare Policy*, *8*, 175–183. https://doi.org/10.2147/RMHP.S70173.

Hyperscalers in crosshairs for anti-competitive pricing and lock-in. (n.d.). CIO. Retrieved 25 November 2024, from https://www.cio.com/article/648048/hyperscalers-in-crosshairs-for-anti-competitive-pricing-and-lock-in.html.

Kouroupa, A., Laws, K. R., Irvine, K., Mengoni, S. E., Baird, A., & Sharma, S. (2022). The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis. *PLOS ONE*, *17*(6), e0269800. https://doi.org/10.1371/journal.pone.0269800.

Learning disability statistics: Mental health problems. (2016, October 28). Foundation for People with Learning Disabilities. https://www.learningdisabilities.org.uk/learning-disabilities/help-information/learning-disability-statistics-/187699.

Packin, N. G. (2021). *Disability Discrimination Using AI Systems, Social Media and Digital Platforms: Can We Disable Digital Bias?* (SSRN Scholarly Paper No. 3724556). Social Science Research Network. https://doi.org/10.2139/ssrn.3724556.

Petrić Howe, N. (2024). ChatGPT has a language problem — But science can fix it. *Nature*. https://doi.org/10.1038/d41586-024-02579-z.

Rahimunnisa, K., Atchaiya, M., Arunachalam, B., & Divyaa, V. (2020). AI-based smart and intelligent wheelchair. *Journal of Applied Research and Technology*, *18*(6), 362–367. https://www.redalyc.org/journal/474/47471676003/html/.

Stypinska, J. (2023). AI ageism: A critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & SOCIETY*, *38*(2), 665–677. https://doi.org/10.1007/s00146-022-01553-5.

Syaodih, E., & Aprilesti, L. (2020). Disability-friendly public space performance. *IOP Conference Series: Materials Science and Engineering*, *830*, 022028. https://doi.org/10.1088/1757-899X/830/2/022028.

The EU's Digital Services Act. (2022, October 27). https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.

Urbina, J. T., Vu, P. D., & Nguyen, M. V. (2024). Disability Ethics and Education in the Age of Artificial Intelligence: Identifying Ability Bias in ChatGPT and Gemini. *Archives of Physical Medicine and Rehabilitation*. https://doi.org/10.1016/j.apmr.2024.08.014.

Using AI to support people with disability in the labour market. (2023, November 23). OECD. https://www.oecd.org/en/publications/using-ai-to-support-people-with-disability-in-the-labour-market_008b32b7-en.html.

What you need to know about UNESCO's new AI competency frameworks for students and teachers. (n.d.). UNESCO. Retrieved 25 November 2024, from https://www.unesco.org/en/articles/what-you-need-know-about-unescos-new-ai-competency-frameworks-students-and-teachers.

With Children at Higher Risk of Severe Rights Violations, Third Committee Emphasizes Need to Regulate Digital Spheres, Boost Protection Regimes. (n.d.). Meetings Coverage and Press Releases. Retrieved 25 November 2024, from https://press.un.org/en/2023/gashc4377.doc.htm.

PART 3

GLOBAL MAJORITY FACING AI

11 Reparative Algorithmic Impact Assessments: A Decolonial, Justice-Oriented Accountability Framework for AI and the Global Majority

Elise Racine

Abstract

While artificial intelligence (AI) promises transformative societal benefits, it also presents significant challenges in ensuring equitable access and value for the Global Majority. Building on emerging research on algorithmic reparations, algorithmic impact assessments, and participatory AI, this paper introduces Reparative Algorithmic Impact Assessments (R-AIAs) — a novel framework that combines robust accountability mechanisms with a reparative praxis to form a more culturally sensitive, justice-oriented methodology. By further incorporating decolonial, Intersectional principles, R-AIAs move beyond merely centering diverse perspectives and avoiding harm to actively redressing historical, structural, and systemic inequities. This includes colonial legacies and their algorithmic manifestations. Using the example of an AI-powered mental health chatbot in rural India, we explore concrete strategies through which R-AIAs can achieve these objectives, fostering equity for the Global Majority in the process.

Keywords: Artificial Intelligence; Accountability; Participatory Governance; Algorithmic Impact Assessments; Algorithmic Reparations; Algorithmic Harm; Algorithmic Colonialism; Decolonial AI; Intersectionality; Global Majority.

Introduction

Artificial intelligence (AI) has emerged as a transformative force across sectors, offering immense potential to tackle complex global challenges (Vinuesa et al., 2020). But AI's use also raises pressing ethical concerns, including the possibility for algorithmic systems to amplify biases, reproduce injustices, and exacerbate global inequities

(Ashar, Ginena, Cipollone, Barreto, & Cramer, 2024; Davis, Williams, & Yang, 2021; Igarapé Institute, 2024; Racine, 2024). These concerns are especially acute in the Global South, where "wicked" problems marked by resource constraints, infrastructure limitations, and unique socio-cultural considerations are more prevalent.

Despite these complexities, however, the Global Majority — comprising diverse communities across Africa, Asia, Latin America, and other regions — remain underrepresented in the design, development, deployment, research, and governance of AI-powered technologies (Igarapé Institute, 2024). This underrepresentation has led to systems that not only inadequately serve but often harm large portions of the world's population. To truly harness AI's potential for the benefit of all, we must prioritize the development of inclusive, equitable algorithmic systems that center the Global Majority. Reparative accountability mechanisms, grounded in decolonial and Intersectional principles, can play a crucial role in achieving this goal.

Drawing from emerging research on algorithmic reparations, algorithmic impact assessments, decolonial AI, and participatory AI, we propose a novel framework: Reparative Algorithmic Impact Assessments (R-AIAs). This approach emphasizes meaningful engagement from diverse communities throughout the AI lifecycle, surpassing traditional notions of algorithmic fairness to redress historical, structural, and systemic inequities.

11.1 Challenges in Ensuring AI Benefits the Global Majority

The dominant discourse in Western technological spaces is one of hype, where the promise of AI to address global challenges and improve lives worldwide is emphasized (Crawford, 2021; Dežman, 2024). But in reality, these benefits are not equally distributed (Benjamin, 2019; Igarapé Institute, 2024; Mohamed, Png, & Isaac, 2020). Several key challenges prevent AI from serving the Global Majority. One of the most substantial challenges is the lack of Global Majority involvement throughout the AI lifecycle — even when directly impacted. This extends to the AI workforce itself (Okolo, 2023). Critical AI functions like data labeling and content moderation are routinely outsourced to the Global Majority, often

subjecting workers to traumatic conditions and low pay (Igarapé Institute, 2024; Okolo, 2023; Perrigo, 2022, 2023).

There is also insufficient culturally sensitive data documenting the full depth and vibrancy of lived experiences from the Global Majority. Instead, algorithmic systems are trained on datasets that often reflect and amplify existing biases. For example, gender and skin-type bias in commercial facial-analysis technologies are well documented, with these tools performing consistently worse for individuals who are not white cisgender men (Birhane, 2022; Buolamwini & Gebru, 2018; Scheuerman, Paul, & Brubaker, 2019). Such misclassification has resulted in discrimination, privacy violations, wrongful policing, the reinforcement of harmful stereotypes, and a host of other harms. Moreover, AI-powered systems developed primarily in Western contexts often fail to account for the diverse cultural norms, values, and social structures of the Global Majority and Indigenous communities. This can lead to inappropriate or even harmful applications when these systems are implemented in different contexts. For instance, AI-powered content moderation systems may struggle to accurately interpret culturally specific expressions or nuances, leading to censorship or the spread of harmful matter (Sambasivan et al., 2021).

The dominance of Western epistemologies in AI design, development, deployment, research, and governance has also given rise to concerns about algorithmic colonialism. This phenomenon describes how AI-powered systems can impose particular ways of knowing and categorizing the world, potentially erasing or marginalizing indigenous and alternative knowledge systems. Birhane (2020) identifies several key manifestations: exploitative data practices, Western knowledge dominance, technological reliance, and cultural standardization (see also Mohamed et al., 2020). Crucially, algorithmic systems tend to operate from hetero-cis-normative, colonial, and capitalist epistemic positions, illustrating how these power structures can be extended via these tools (Mohamed et al., 2020; Racine, 2024).

Marginalized and minoritized groups, like sexual and gender minorities, are especially vulnerable to these epistemic impositions. For example, commercial facial-analysis technologies only return binary labels, entirely excluding non-binary/genderqueer individuals (Scheuerman

et al., 2019). However, these Western labels (e.g., man/woman, homosexual/heterosexual) may be unsuitable in other cultural settings and repress invaluable diversity and complexity (Young & Meyer, 2005; Racine, 2023, 2024). This includes local self-determined identities that operate outside these binary categorizations, such as the bacha bereesh of Afghanistan, hijra of India, and māhū and fa'afafine of the Pacific Islands. Not only are these identities and rich histories erased when Western frameworks/norms are imposed, but such acts of epistemic violence can perpetuate significant, long-lasting harm.

This epistemic injustice is compounded by the concentration of AI power in the hands of a few tech giants. As of July 2024, 14 of the 15 largest AI companies by market cap were US-based, with the remaining based in Israel (Stash, 2024). These entities often extract data and economic value from the Global Majority while imposing technological dependence and cultural homogenization. For instance, major tech platforms routinely collect personal data from users in the Global South, using it to train AI-powered systems without transparent data practices or proper compensation. Meanwhile, the concentration of AI talent, compute resources, research funding, and infrastructure in the hands of Global Minority powers more broadly — predominantly the United States, United Kingdom, European Union, China, Japan, and South Korea — limits access to the knowledge and tools necessary for technological sovereignty for many in the Global Majority (Igarapé Institute, 2024; Lehdonvirta, Wu, & Hawkins, 2024). The result is a multi-faceted form of colonialism operating on both epistemic and economic levels.

Furthermore, many algorithmic systems operate as "black boxes," with decision-making processes that are opaque and difficult to scrutinize. This lack of transparency makes it challenging to identify and address biases or errors as they arise, especially when these systems are deployed in critical domains like healthcare, criminal justice, or financial services. The absence of robust accountability mechanisms exacerbates this issue, hindering affected communities from seeking redress. Tackling these challenges requires a fundamental shift in how we approach all stages of the AI lifecycle. The following sections will explore how R-AIAs can provide a pathway towards more inclusive and equitable AI-powered technologies/systems that benefit the Global Majority.

11.2 Reparative Algorithmic Impact Assessments

Algorithmic Impact Assessments (AIAs) have emerged as a promising participatory accountability mechanism for evaluating the potential societal impacts (e.g., social, environmental, economic, cultural) of algorithmic systems before their implementation (Ada Lovelace Institute, 2021; Stahl et al., 2023; Watkins, Moss, Metcalf, Singh, & Elish, 2021). These assessments can generate greater accountability, explainability, transparency, and reflexivity (Ada, 2021; Ashar et al., 2024; Metcalf, Moss, Watkins, Singh, & Elish, 2021; Reisman, Schultz, Crawford, & Whittaker, 2018; Selbst, 2021; Watkins et al., 2021; Stahl et al., 2023). Consequently, they can also mitigate risks, maximize benefits, and foster increased understanding of and trust in AI-powered technologies (Ada, 2021; Ashar et al., 2024) — including for the purposes of sustainable development. And when designed with diversity and accessibility in mind, they can be a powerful advocate for inclusion and equity. However, as highlighted in the systematic review by Stahl et al. (2023), the field of AIAs is still maturing, with a lack of full agreement on the structure, content, and implementation of these assessments. This underscores the need for clearer frameworks and more cohesive, context-specific strategies.

For these assessments to be effective, they must incorporate diverse perspectives. The key is meaningful, active engagement that goes beyond tokenism, where lived experiences directly inform AI design, development, and deployment. As it stands, marginalized voices have been routinely omitted from accountability mechanisms and traditional algorithmic fairness efforts, a gap that is well-documented in both AI fairness literature and critiques of current participatory approaches (Birhane, 2021, 2022; Birhane et al., 2022; Davis et al., 2021; Racine, 2024). Moreover, traditional AIAs often fall short in ameliorating the deep-rooted inequities that shape the context in which these systems operate. This is where the concept of algorithmic reparations becomes vital. As Davis et al. (2021) articulate, algorithmic reparations aim to “name, unmask, and undo allocative and representational harms as they materialize in sociotechnical form.” This approach goes beyond technical performance to (a) consider how power flows through these systems and (b) place these developments within broader

patterns of oppression, privilege, marginalization, and disadvantage (Johnson, 2021; Kalluri, 2020; Racine, 2024).

Building on these concepts, we propose a novel, transformative approach: Reparative Algorithmic Impact Assessments (R-AIAs). R-AIAs combine the structured participatory evaluation process of AIAs with the justice-oriented focus of algorithmic reparations. They seek to actively rectify historical imbalances and ongoing disparities in technological development and deployment, particularly centering experiences and knowledge from the Global Majority. This approach is grounded in the understanding that AI does not operate in a vacuum but is embedded in complex social, economic, and political contexts shaped by histories of global power dynamics (Birhane, 2022; Davis et al., 2021; Kalluri, 2020; Racine, 2024).

The key components of R-AIAs are:

1. Deep consideration of historical context,
2. Thorough analysis of power dynamics and asymmetries,
3. Commitment to meaningful community engagement,
4. Incorporation of decolonial, Intersectional principles that recognizes the complex interplay of various aspects of identity, and 5. Focus on sustainable development and long-term impacts.

These assessments should not be viewed as a one-time evaluation but as an ongoing process that allows for continuous learning and adaptation (Ada, 2021; Watkins et al., 2021). They aim to go beyond simply identifying potential harms or biases in algorithmic systems and a narrow focus on technical fairness to actively securing justice and equity for the Global Majority.

11.3 Incorporating Decolonial, Intersectional Principles

To foster inclusive and equitable AI for the Global Majority, adopting decolonial, Intersectional principles in AI design, development, deployment, research, and governance is essential. Intersectionality, as introduced by Crenshaw (1991), recognizes that individuals experience overlapping forms of discrimination/oppression based on aspects of their identity, such as race/ethnicity, gender, class, sexual orientation, disability, and religion. A reparative praxis builds

on this foundation by not only addressing these intersections but actively working to repair associated harms (Davis et al., 2021; Racine, 2024). By weaving Intersectionality into every step of assessment process and centering marginalized voices, R-AIAs aim to not only dismantle inequitable structures, but drive material benefits and systemic change for those most affected by algorithmic injustice.

Decolonial thinking challenges the dominance of Western epistemologies, calling for a fundamental shift toward recognizing and incorporating diverse knowledge systems (Miller, 2022; Mohamed et al., 2020; Zimeta, 2023). This is critical to make certain that AI is not solely driven by Western-centric values but instead reflects the needs, values, and priorities of communities from the Global Majority and other marginalized groups.

For R-AIAs, we propose the following principles for decolonial, Intersectional AI:

1. Epistemological diversity: Actively incorporating diverse knowledge systems into AI development.
2. Data sovereignty: Respecting the rights of communities to control their data.
3. Technological self-determination: Empowering communities to develop and deploy AI that aligns with their values and needs.
4. Cultural preservation: Ensuring AI respects and promotes cultural diversity.
5. Reciprocity: Establishing mutually beneficial relationships between AI developers and communities.

11.4 From Principles to Practice: Implementing Decolonial R-AIAs

R-AIA implementation requires a systemic overhaul of both mindset and methodology. Below, we outline several key steps for operationalizing R-AIAs. To contextualize these further, we have used the example of a US-based technology company piloting an AIpowered chatbot to provide 24/7 mental health support to underserved communities across rural India. Each also include sample strategies/practice(s) that align or misalign with this reparative, decolonial approach. With an emphasis on including voices from the

Global Majority throughout the process, R-AIAs demands diverse, interdisciplinary teams.

11.4.1 Socio-Historical Research

Conducting thorough research into socio-historical contexts to understand the complex backdrop against which AI-powered systems operate is an essential first step. This includes desk-based investigations into past harms caused by similar technologies — particularly for marginalized and minoritized groups (Partnership on AI, 2024). This research helps identify not only possible impacts, power asymmetries, and reparative actions, but participants for engagement (PAI, 2024).

- *Reparative:* Employ librarians and information specialists with data curation and archival expertise play a key role, grounding the research in socio-historical realities/injustices (Davis et al., 2021; Racine, 2024). Appropriately incorporate Indigenous and non-Western knowledge systems.

11.4.2 Participant Engagement and Impact/Harm Co-Construction

Impacts should be co-constructed and rigorously mapped to potential harms through non-tokenistic engagement that (re)distributes power; this redistribution is essential to producing accountability (Metcalf et al., 2021). This engagement must meaningfully center the lived experiences of those most affected by AI-powered technologies. Whether it should be continuous is debated. It is paramount to mitigate the toll repeated consultants take on participants, especially for vulnerable populations and regarding sensitive topics like mental health (PAI, 2024). To safeguard participants' well-being, ethical guidelines (e.g., informed consent) must be followed. Fostering equitable collaborations that prioritize knowledge exchange and capacity building between institutions in the Global Majority and Minority can set the groundwork for effective consultations.

- *Reparative:* Utilize socio-historical research to inform participant recruitment. Implement mechanisms to navigate divergent values, iterate based on new knowledge, and alleviate burden for participants (e.g., offer compensation, mental health resources,

flexible participation options). Make certain accessibility needs are met and participant feedback directly shapes chatbot functionality

- *Harmful:* Define and assess impacts based on superficial consultations with highprestige experts while neglecting input from affected communities (PAI, 2024).

11.4.3 Sovereign and Reparative Data Practices

Data governance frameworks must respect Indigenous data sovereignty principles and ensure communities retain control over their information (Carroll, Duarte, & Liboiron, 2024; Kukutai & Taylor, 2016). Furthermore, developing inclusive, reparative data methods can address underrepresentation of diverse, minoritized experiences, correct historical exclusion, and, ultimately, contribute to long-term, community-driven outcomes.

- Reparative: Establish community-controlled data trusts, allowing local communities to decide how their data is used.
- Harmful: Extract data from affected communities without their consent or participation in decision-making, perpetuating data colonialism. Limit training to data from urban populations or Western mental health models, reinforcing disparities for rural communities.

11.4.4 Ongoing Monitoring and Adaptation

Finally, R-AIAs emphasize the importance of continuous monitoring and adjustment of AI-powered systems based on real-world impacts, acknowledging that the work of equity and justice is ongoing and iterative.

- Reparative: Develop new ways of measuring chatbot performance that incorporate diverse cultural values and priorities, challenging the dominance of Western-centric benchmarks in AI evaluation.

11.4.5 Redress

Moving beyond merely identifying issues, R-AIAs call for concrete, actionable plans that actively redress deep-rooted inequities (Davis et al., 2021) and algorithmic coloniality.

- Reparative: Partner with local AI hubs and research institutes that empower communities to develop their own AI capabilities. Offer scholarships and fellowships to underrepresented communities in rural India. Increase access to compute resources by investing in local infrastructure or providing cloud-based solutions that rural communities can use to develop and refine AI models tailored to their specific needs.

11.5 Concluding Remarks

Shaped by colonial, Western paradigms, AI-powered systems can reinforce global inequities. By combining culturally sensitive participatory methods with a reparative praxis and decolonial, Intersectional principles, the R-AIA framework moves beyond merely avoiding harm to actively contributing to reparative outcomes for the Global Majority. This approach fosters justice and equity while providing concrete strategies for addressing and redressing colonial legacies and their algorithmic manifestations.

11.6 References

Ada Lovelace Institute. (2021). Algorithmic impact assessment: A case study in healthcare. Ada Lovelace Institute. Retrieved from https://www.adalovelaceinstitute.org/wp-content/uploads/2022/02/Algorithmicimpact-assessment-a-case-study-in-healthcare.pdf.

Ashar, A., Ginena, K., Cipollone, M., Barreto, R., & Cramer, H. (2024). Algorithmic impact assessments at scale: Practitioners' challenges and needs. Journal of Online Trust and Safety, 2(4). https://doi.org/10.54501/jots.v2i4.206.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Malden, MA: Polity.

Birhane, A. (2020). Algorithmic colonization of Africa. SCRIPTed, 17(2), 389-409. https://doi.org/10.2966/scrip.170220.389.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2, 1-9.

Birhane, A. (2022). The limits of fairness. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (p. 2). New York, NY: Association for.

Computing Machinery. https://doi.org/10.1145/3514094.3539568.

Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI.

In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22) (17 pages). ACM, New York, NY. https://doi.org/10.1145/3551624.3555290.

Birhane, A. (2022). The unseen Black faces of AI algorithms. Nature, 610, 451-452. https://doi.org/10.1038/d41586-022-03050-7.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html.

Carroll, S. R., Duarte, M., & Liboiron, M. (2024). Indigenous data sovereignty. In Keywords of the datafied state (pp. 207-223). Data & Society.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review, 43(6), 1241-1299. https://doi.org/10.2307/1229039.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. New Haven, CT: Yale University Press.

Davis, J. L., Williams, A., & Yang, M. W. (2021). Algorithmic reparation. Big Data & Society, 8(2), 20539517211044808. https://doi.org/10.1177/20539517211044808 Igarapé Institute. (2024). Responsible Artificial Intelligence efforts in the Global South. Igarapé Institute.

Johnson, K. (2021). A move for 'algorithmic reparation' calls for racial justice in AI. Wired. Retrieved from https://www.wired.com/story/move-algorithmic-reparationcalls-racial-justice-ai/.

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature, 583(7815), 169-169. https://doi.org/10.1038/d41586-020-02003-2.

Kukutai, T., & Taylor, J. (2016). Indigenous data sovereignty: Toward an agenda. ANU Press. https://doi.org/10.22459/CAEPR38.11.2016.

Lehdonvirta, V., Wu, B., & Hawkins, Z. (2024, August 22). Compute North vs. Compute South: The uneven possibilities of compute-based AI governance around the globe. https://doi.org/10.31235/osf.io/8yp7z.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & 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). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445935.

Miller, K. (2022, March 21). The movement to decolonize AI: Centering dignity over dependency. Stanford HAI. Retrieved from https://hai.stanford.edu/news/movement-decolonize-ai-centering-dignity-overhttps://hai.stanford.edu/news/movement-decolonize-ai-centering-dignity-over-dependencydependency.

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659-684. https://doi.org/10.1007/s13347-020-00405-8.

Moss, E., Watkins, E., & Metcalf, J. (2021). Governing with algorithmic impact assessments: Six observations. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. https://ssrn.com/abstract=3584818.

Okolo, C. T. (2023, November 1). AI in the Global South: Opportunities and challenges towards more inclusive governance. Brookings.

Partnership on AI. (2024). Guidelines for Participatory and Inclusive AI. Partnership on AI.https://partnershiponai.notion.site/1e8a6131dda045f1ad00054933b0bda0?v=dcb890146f7d464a86f11fcd5de372c0.

Perrigo, B. (2022, February 14). Inside Facebook's African sweatshop. TIME. https://time.com/6147458/facebook-africa-content-moderation-employeehttps://time.com/6147458/facebook-africa-content-moderation-employee-treatment/treatment/.

Perrigo, B. (2023, January 18). OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. TIME. https://time.com/6247678/openai-chatgpthttps://time.com/6247678/openai-chatgpt-kenya-workers/kenya-workers/.

Racine, E. (Forthcoming). Que(e)rying artificial intelligence use for infectious disease surveillance: The need for a reparative algorithmic praxis. Big Data & Society.

Racine, E.E. (2023). Sexuality and gender within Afghanistan's bacha bereesh population. Equality, Diversity and Inclusion, 42(5), 580-609. https://doi.org/10.1108/EDI-04-2022-0096.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15). https://doi.org/10.1145/3411764.3445518.

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis and image labeling services. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 144, 33 pages. https://doi.org/10.1145/3359246.

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. Harvard Journal of Law & Technology, 35(1), 117-190. https://doi.org/10.2139/ssrn.3584818https://doi.org/10.2139/ssrn.3584818.

Stahl, B.C., Antoniou, J., Bhalla, N., et al. (2023). A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review, 56, 12799–12831. https://doi.org/10.1007/s10462-023-10420-8.

Stash. (2024, August 8). 15 largest AI companies in 2024. Stash Learn. Retrieved from https://www.stash.com/learn/top-ai-companies/.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 233. https://doi.org/10.1038/s41467-019-14108-y.

Vrabič Dežman, D. (2024). Promising the future, encoding the past: AI hype and public media imagery. AI Ethics, 4, 743-756. https://doi.org/10.1007/s43681-024-00474x.

Young, R. M., & Meyer, I. H. (2005). The trouble with "MSM" and "WSW": Erasure of the sexual-minority person in public health discourse. American Journal of Public Health, 95(7), 1144-1149. https://doi.org/10.2105/AJPH.2004.046714.

Zimeta, M. (2023). Why AI must be decolonized to fulfill its true potential. The World Today. Chatham House. https://www.chathamhouse.org/publications/the-worldtoday/2023-10/why-ai-must-be-decolonized-fulfill-its-true-potential.

12 AI Ethics for the global majority: lessons from decolonial feminist bioethics

Alice Rangel Teixeira

Abstract

Artificial Intelligence presents both opportunities and challenges in promoting human flourishing. While AI has the potential to reduce inequalities and improve outcomes, its applications often reinforce biases, especially against marginalized groups. This paper critically examines the dominant principle-based approach to AI ethics, which neglects power imbalances and social context of AI applications. Drawing from decolonial feminist bioethics, the paper proposes an alternative model for AI ethics that addresses structural injustices and centers the needs of the global majority. Through a critical analysis of existing AI ethics frameworks, the paper highlights their limitations in addressing power asymmetries and exclusionary practices. It argues for a shift towards an ethical framework that incorporates decolonial feminist theories and methods,developed in the field of bioethics as an alternative to the principlist approach, to ensure equitable and socially just AI development.

Keywords: AI Ethics, Social Justice, Global Majority, Decolonial Feminist Bioethics.

Introduction

Artificial Intelligence (AI) has been celebrated for potentially empowering humans and stimulating human flourishing, from applications that save human time with automation in areas overloaded such as the judiciary system, healthcare and business, to the prevention, diagnostic and treatment of diseases, the reduction of poverty, inequalities in healthcare and education and gender-based discrimination (Baclic et al. 2020; Floridi et al. 2018; García-Micó & Laukyte 2023; Goralski & Tan 2023; Topol 2019). However, its applications are often reported to be biased and discriminatory, untrustworthy, harming individuals and marginalized groups, and reinforcing inequalities (López Belloso

2022; Mhlambi & Tiribelli 2023; Mohamed et al. 2020; Morondo Taramundi 2022; Ricaurte 2022). These ethical concerns and the recent developments on AI capabilities have stimulated the debate of AI ethics and the publication of several frameworks and guidelines from different sectors such as business, institutional and governmental (Floridi & Cowls 2019).

This paper discusses an alternative ethical framework that addresses these shortcomings by centering the needs of the global majority. Through a critical analysis of the dominant approach, AI principle-based ethics, this paper will explore how feminist and decolonial perspectives can provide tools to develop a more inclusive and just framework. The analysis is structured as follows: first, the principle-based AI ethics and its relationship between bioethics' principlism is discussed; next, the main criticisms of this approach, including its neglect of power asymmetries and social justice issues, are outlined. Finally, feminist and decolonial bioethics is presented as offering critical perspectives on principlism, along with alternative theories and methods that can inform the development of a more inclusive and just AI ethics framework.

12.1 The principle-based approach to AI ethics

Analyzing the global landscape of AI ethics guidelines, Jobin et al. (2019) identified over 11 principles across 84 documents, highlighting that the dominant approach to AI ethics is principle-based. The work also demonstrates how unclear these principles are, with each containing an abundance of codes. The principle of "justice & fairness", for example, includes 16 codes that present ill-defined or broad terms such as reversibility, challenge, inclusion, and equity. As the authors observe, this diversity indicates divergences in how AI ethical challenges are addressed. Furthermore, because most guidelines come from the Global North, it raises concerns on how well-equipped these strategies are to deal with the global scenario of AI without neglecting the particularities of knowledge, needs and interests of underrepresented regions. Mhlambi and Tribelli (2023) observe that despite attending to several principles, guidelines tend to prioritize autonomy, perpetuating historical and abusive practices of racial and gender control and oppression.

Floridi and Cows (2019), argue that the abundance of AI principles enables ethical washing with minimal action. To counter this, they propose unifying AI principles with those of bioethics — autonomy, beneficence, non-maleficence, and justice — while adding explainability. They note a convergence between 47 principles from six key initiatives and these bioethical principles, as bioethics most closely parallels digital ethics in addressing new agents, patients, and environments. Bioethics emerged as a discipline in the late 1960s and early 1970s to address ethical issues emerging from modern medical practices. Shortly after, *The Principles of Biomedical Ethics* (Childress and Beauchamp, 1979) was published, introducing a principle-based approach that seeks to adjust the balance between particular judgments and general norms by focusing first. Beauchamp and Childress argue that the four principles represent a set of essential values shared in our 'common morality ' leading us to instinctively rely on them in decision-making (Tong, 2019). Over time, this principle-based approach became known as "mainstream bioethics" (Scully et al., 2010).

Floridi (2021) also notes that the principle-based approach proposed by the European Union (EU) "high-level expert group on artificial intelligence", influenced the design of the European legislation on AI, the EU AI ACT. However, there is little discussion on how these principles should be interpreted or the philosophical theories behind them (Mohamed et al., 2020). Analyzing 221 journal articles on AI ethics, Bakiner (2023) a lack of theoretical grounding in AI ethics, with a prevailing view that no theory is needed, and little attention to social and justice issues. To Lin and Chen (2022), AI ethics fails to address systemic injustice by focusing on mitigating bias in the algorithm. They highlight that power asymmetries shape AI, citing healthcare as an example where common datasets are biased towards US and European data, and practitioners' biases — such as racial and LGBT biases — which in turn affect AI's performance, as they generally provide the standards for its evaluation in cases such as disease diagnosis.

12.2 Critiques of the principle-based approach

Because AI ethical frameworks are predominantly from Global North regions and neglect power asymmetries and systemic injustice,

they allow the deployment of technologies that reinforce coloniality and the matrix of domination (Collins, 2000), contributing to the economic, social and epistemic oppression of marginalized social groups. This can be observed in the uneven effects of AI that disproportionately exploit human labor and natural resources of the Global South regions, while benefits disproportionately benefit regions of the Global North (Couldry & Mejias, 2019; Ricaurte, 2023; Van Dijck, 2014). Scholars focused on the effects of coloniality and neo-colonialism through data or computation colonialism, point to problems in the Western philosophical traditions that serve as foundations to the principle-based approach. Arguing that these traditions have served the interest of those in power as a tool for continuous oppression of coloniality. It is therefore necessary to consider a diversity of epistemologies instead of assuming a core shared value system (Mhlambi & Tiribelli, 2023; Mohamed et al., 2020; Ricaurte, 2022; Valente & Grohmann, 2024). Feminist scholars approaching the subject of ethical AI from gender analysis have also pointed to the same problems. As they argue, because feminism has been historically attuning to power inequities, it presents itself as a framework that can be applied to AI, allowing the continuous examination of its power asymmetries as a method to prevent exacerbating oppression (Ciston, 2019; D'Ignazio & Klein, 2020; Hancox-Li & Kumar, 2021; Katell et al., 2020).

There has been less work on proposing moral and epistemological theories that could sustain these models, which can be problematic. By leaving untouched the philosophical assumptions behind a proposed framework, there is a risk of repeating the problems on how to interpret the concepts that underpin the main concerns of the field, contributing to the current scenario of confusion and ethical-washing. If the assumed interpretation is left undiscussed, it also risks favoring a specific standpoint while neglecting others. Finally, it closes the possibility of learning from similar fields by looking to the theories developed or strengthened by them. This can be observed, for instance, by the lack of cross-work between feminist and decolonial AI scholars, or the absence of feminist, racial and decolonial theoretical foundations, with a few exceptions such as the work of Ricaurte (2022) and Birhane (2021). These works greatly

contribute to the cohesive critique of AI principle-based approaches that takes into account different systems of oppression, however they are less focused on discussing ethical theory. The complexity and potential impact of AI technology can be better understood through a careful examination of its risks and possibilities that can in turn guide the regulation and governance of the technology with a clear social-political direction (Floridi, 2018). Noteworthy, these fields have been an essential part of feminist bioethics' long standing criticism of the principle-based approach in bioethics and crucial contributors in feminist ethics (Rogers et al., 2022) that can serve as lessons to inform a decolonial feminist AI ethics.

12.3 From bioethics to AI ethics: lessons from feminist bioethics

Feminist bioethics argues that mainstream bioethics is based on ontological and epistemological foundations that favor culturally masculine ways of being and knowing. Bioethical principles are presented as universal rules that apply equally to generic and interchangeable people, while its ethical analysis has often concentrated on the rights and interests of an abstract, disembodied individual, isolated from social and historical context. As a consequence, it neglects the interests and needs of women and other marginalized social groups, relegating politically vulnerable groups to a position of moral inferiority, and compounding inequities. Feminist bioethics, though diverse, seeks non-oppressive and inclusive alternatives (Lindemann, 2022; Scully et al., 2010). This section explores their key critiques of the principlist approach and presents theoretical and methodological alternatives for feminist and decolonial AI ethics.

12.3.1 Feminist epistemology and methodology

Feminist bioethics critiques the methods and conclusions of scientific research by highlighting the relationship between knowledge and power. It challenges biased assumptions, such as those found in eugenics and the pathologization of women's bodies, or racial differences in pain threshold, arguing that abstract reasoning, detached from social context, is impossible (Ganguli-Mitra, 2022; Hutchison, 2022). Scholars like Patricia Hill Collins and Maria Lugones,

building on Standpoint Theory and intersectionality, propose alternative models that prioritize those in the margins that are often oppressed by these bodies of knowledge (Collins, 2000; Hutchison, 2022; Lugones et al., 1983; Stoetzler & Yuval-Davis, 2002). Others, influenced by postmodern feminist thought, claims the impossibility that true objectivity knowledge can ever be achieved (Anderson, 2024), they focus instead on the current situation that people actually face (Hutchison, 2022), prioritizing empirical research over normative judgment, but with a feminist methodology that takes into account the power structures that delineate any research, including the relationship between researcher and subjects (Scully, n.d.).

The lack of epistemic diversity in AI ethics amplifies inequities (Birhane, 2021; Mhlambi & Tiribelli, 2023; Mohamed et al., 2020; Ricaurte, 2023). Creating a disconnect between AI policies and empirical evidence (Carter et al., 2024; Frost et al., 2024), that might prevent the use of technology in ways that are meaningful for individuals and communities. For instance, a systematic review for medical AI applications highlighted that patients' concerns do not align with the current focus on autonomy (Tang et al., 2023). Feminist research and practice is also self-reflective, acknowledging that traditional knowledge systems have oppressed marginalized groups and regularly critiques its own work (Scully et al., 2010), this self-reflectiveness helps to avoid the universalist trap. While limited in scope, this indicates the benefits of adopting a feminist epistemology when constructing AI ethical frameworks.

12.3.2 Justice

Feminist critiques of justice in bioethics highlight the influence of Rawls's Theory of Justice, Utilitarianism, and Distributism. They criticize Rawls for prioritizing abstract principles over practical realities (Fourie, 2022), neglecting social injustices like gender and race (Jaggar, 2009)). Utilitarianism, focused on maximizing benefits, similarly overlooks what equality and well-being might mean (Scully et al., 2010), while distributism often emphasizes resource distribution without addressing non-quantifiable injustices, such as epistemic injustice — the devaluation of marginalized knowledge systems (Fricker, 2007). Current AI ethics, with its focus on resource

distribution, leading to the under-analysis of structural injustice and how it relates to AI technology (Lin & Chen, 2022)). Instead, Iris Young's model of shared responsibility (Young, 2011), which addresses systemic oppression collectively, offers a more inclusive approach to justice in AI that does neglect systemic injustice and places responsibility in the collective (Lin & Chen, 2022), diverging from current discussions of AI's responsibility that are limited to the liability model.

12.3.3 Neo-liberal autonomy vs relational autonomy

Feminist analysis of autonomy critiques the dominant libertarian view, which equates autonomy with maximizing individual choice (Scully et al., 2010; Stoljar & Mackenzie, 2022). Taken as a proficiency equally possessed by all competent adults in all circumstances, autonomy is tentatively reduced to a patient's informed consent, failing to account for the contextual conditions that influence a patient's decision, such as power hierarchies and economic disparities. In response, feminist bioethics advocates for relational autonomy (Stoljar & Mackenzie, 2022), which sees the self as socially constituted and considers values, social, historical, and emotional factors (Marway & Widdows, 2015).

While there is growing support for relational autonomy in AI ethics (Mhlambi & Tiribelli, 2023), simply shifting from a liberal to a relational view without rethinking the principle-based approach will not fully address its limitations. Likewise, the proposition of a relational ethics approach to AI (Birhane, 2021) without a discussion of what this moral framework entails, from moral agency to moral responsibilities, risks to obscure biased views that even if unintended might reinforce oppressions

12.3.4 Universalism and particularism

Starting with a critique of the abstraction and generalization of the principle-based approach, feminist bioethics, intersecting gender with social identities such as race, ethnicity, and sexuality, has had to incorporate non-Western perspectives and accommodate these differences through its history of activism and self-reflection. This critique of universalism, combined with the adoption of a relational

view of the self promotes a focus on the local and particular, seeking to extract context-specific experience to make explicit the social process in which gender and other differences are transformed into inequities (Marway & Widdows, 2015). AI ethics needs to draw from this same framework, concerning itself more with the particularities of moral life in its local applications instead of applying universalist models that invisibilize the differences in power.

12.3.5 Crisis problems and mundane problems

Feminist empirical work provides concrete evidence for the existence of ethical issues that otherwise would be dismissed, either because researchers neglect the experience of the women, or they fail to account for power asymmetries, including within the research process itself, failing to create an environment where marginalized voices can be heard. By doing so, feminists have broadened the scope of bioethics, which often focuses on high-profile "crisis issues," to include everyday concerns, deemed too mundane for ethical consideration, such as doctor-patient relationships and the impact of caregiving on family caregivers.

Current approaches that seek to legislate AI application, such as the EU AI Act focus on high-risk applications, meaning those that present a risk to fundamental rights, or that might negatively impact the health and safety of people and the environment, and limited risk, those that might present risk of manipulation and deceit. This approach has come under scrutiny, particularly from civil societies, that identify loopholes where applications deemed low-risk can still threaten fundamental rights (Edwards, 2022; Jonathan Day et al., 2024). The large scope of feminist bioethics could be a useful lens to investigate "mundane problems".

12.4 Conclusion

This paper has demonstrated that mainstream AI ethics frameworks, dominated by Global North perspectives, follow a principle-based approach influenced by mainstream bioethics. A review of feminist bioethics provided insight on how principle-based ethics neglect power asymmetries and perpetuate systems of oppression. The alternative theories and methods developed in decolonial feminist bioethics

offer significant advances beyond the principlism. Their emphasis on power relations, relational autonomy, shared responsibility, empirical evidence and local contexts creates opportunities for rethinking the ethical AI, offering distinctive tools for addressing structural injustice. By integrating insights from decolonial feminist bioethics, it is possible to build an AI ethics that challenges existing systems of oppression, centering decision-making on the needs and interests of the global majority.

12.5 References

Anderson, E. (2024). Feminist Epistemology and Philosophy of Science. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/feminism-epistemology/.

Bakiner, O. (2023). What do academics say about artificial intelligence ethics? An overview of the scholarship. *AI and Ethics*, *3*(2), 513–525. https://doi.org/10.1007/s43681-022-00182-4.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2). https://doi.org/10.1016/j.patter.2021.100205.

Carter, S. M., Aquino, Y. S. J., Carolan, L., Frost, E., Degeling, C., Rogers, W. A., Scott, I. A., Bell, K. J., Fabrianesi, B., & Magrabi, F. (2024). How should artificial intelligence be used in Australian health care? Recommendations from a citizens' jury. *Medical Journal of Australia*, *220*(8), 409–416. https://doi.org/10.5694/mja2.52283.

Ciston, S. (2019). *Intersectional Artificial Intelligence Is Essential: Polyvocal, Multimodal, Experimental Methods to Save AI*. *11*(2).

Collins, P. H. (2000). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Psychology Press.

Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, *20*(4), 336–349. https://doi.org/10.1177/1527476418796632.

D'Ignazio, C., & Klein, L. (2020). The Power Chapter. In *Data Feminism*. https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4.

Edwards, L. (2022). *Regulating AI in Europe: Four problems and four solutions*. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf.

Floridi, L. (2018). Soft Ethics and the Governance of the Digital. *Philosophy & Technology*, *31*(1), 1–8. https://doi.org/10.1007/s13347-018-0303-9.

Floridi, L. (2021). The European Legislation on AI: A Brief Analysis of its Philosophical Approach. *Philosophy & Technology*, *34*(2), 215–222. https://doi.org/10.1007/s13347-021-00460-9.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1.

Fourie, C. (2022). "How could anybody think that this is the appropriate way to do bioethics?" Feminist challenges for conceptions of justice in bioethics. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Fricker, M. (2007). Testimonial Injustice. In M. Fricker (Ed.), *Epistemic Injustice: Power and the Ethics of Knowing* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.003.0002.

Frost, E. K., Bosward, R., Aquino, Y. S. J., Braunack-Mayer, A., & Carter, S. M. (2024). Facilitating public involvement in research about healthcare AI: A scoping review of empirical methods. *International Journal of Medical Informatics*, *186*, 105417. https://doi.org/10.1016/j.ijmedinf.2024.105417.

Hancox-Li, L., & Kumar, I. E. (2021). Epistemic values in feature importance methods: Lessons from feminist epistemology. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 817–826. https://doi.org/10.1145/3442188.3445943.

Hutchison, K. (2022). Feminist epistemology. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Jaggar, A. M. (2009). L'imagination au Pouvoir: Comparing John Rawls's Method of Ideal Theory with Iris Marion Young's Method of Critical Theory. In L. Tessman (Ed.), *Feminist Ethics and Social and Political Philosophy: Theorizing the Non-Ideal* (pp. 59–66). Springer.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Jonathan Day, Iwańska, K., Simon, E., & Willamo, K. (2024). *Packed with loopholes: Why the AI Act fails to protect civic space and the rule of law*. Civil Liberties Union for Europe e.V. https://civic-forum.eu/wp-content/uploads/2024/04/AI_Act_RoL_Analysis-0424.pdf.

Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Bintz, C., Raz, D., & Krafft, P. M. (2020). Toward situated interventions for algorithmic equity: Lessons from the field. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 45–55. https://doi.org/10.1145/3351095.3372874.

Lin, T.-A., & Chen, P.-H. C. (2022). Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly*, *8*(3/4), Article 3/4. https://doi.org/10.5206/fpq/2022.3/4.14191.

Lindemann, H. (2022). Feminist bioethics: Where we've come from. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Lugones, M. C., Spelman, E. V., Lugones, M. C., & Spelman, E. V. (1983). Have we got a theory for you! Feminist theory, cultural imperialism and the demand for 'the woman's voice.' *Women's Studies International Forum*, *6*(6), 573–581. https://doi.org/10.1016/0277-5395(83)90019-5.

Marway, H., & Widdows, H. (2015). Philosophical Feminist Bioethics: Past, Present, and Future. *Cambridge Quarterly of Healthcare Ethics*, *24*(2), 165–174. https://doi.org/10.1017/S0963180114000474.

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms. *Topoi*, *42*(3), 867–880. https://doi.org/10.1007/s11245-022-09874-2.

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, *33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8.

Ricaurte, P. (2022). Ethics for the majority world: AI and the question of violence at scale. *Media, Culture & Society*, *44*(4), 726–745. https://doi.org/10.1177/01634437221099612.

Ricaurte, P. (2023). Epistemologias de dados, colonialidade do poder e resistência. *Dispositiva*, *12*(22), 6–26. https://doi.org/10.5752/P.2237-9967.2023v12n22p6-26.

Rogers, W. A., Scully, J. L., Carter, S. M., Entwistle, V. A., & Mills, C. (2022). Introduction. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Scully, J. L. (n.d.). Feminist Empirical Bioethics. In *Empirical Bioethics Theoretical and Practical Perspectives* (pp. 195–221). https://doi.org/10.1017/9781139939829.013.

Scully, J. L., Baldwin-Ragaven, L. E., & Fitzpatrick, P. (2010). *Feminist Bioethics: At the Center, on the Margins*. Johns Hopkins University Press.

Stoetzler, M., & Yuval-Davis, N. (2002). Standpoint theory, situated knowledge and the situated imagination. *Feminist Theory*, *3*(3), 315–333. https://doi.org/10.1177/146470002762492024.

Stoljar, N., & Mackenzie, C. (2022). Relational autonomy in feminist bioethics. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Tang, L., Li, J., & Fantus, S. (2023). Medical artificial intelligence ethics: A systematic review of empirical studies. *DIGITAL HEALTH*, *9*, 20552076231186064. https://doi.org/10.1177/20552076231186064.

Tong, R. P. (2019). *Feminist Approaches to Bioethics: Theoretical Reflections and Practical Applications*. Routledge.

Valente, J. C. L., & Grohmann, R. (2024). Critical data studies with Latin America: Theorizing beyond data colonialism. *Big Data & Society*, *11*(1), 20539517241227875. https://doi.org/10.1177/20539517241227875.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776.

Young, I. M. (2011). Two Structure as the Subject of Justice. In I. M. Young & M. Nussbaum (Eds.), *Responsibility for Justice* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.003.0002.

13 Exploitation All the Way Down: Calling out the Root Cause of Bad Online Experiences for Users of the "Majority World."

Zeerak Talat and Hellina Hailu Nigatu

Abstract

Global Majority users are exposed to multitudes of harm when interacting with online platforms. This essay illuminates how exploitation in the advances of Artificial Intelligence is tied to historical exploitation and how the use of blanket terminology overshadows the layers of exploitation and harm "Global Majority" populations face. It first discusses the multitude of harm content moderators from the Global Majority face, arguing against the current trend of protection through exploitation, then it illustrates the nuances and differences within the Global Majority, and finally, it outlines actionable items to move away from such harm.

Keywords: Exploitation, Artificial Intelligence, Global Majority, Content Moderators, Content Moderation.

Introduction

Global Majority users are disproportionately affected by the more extreme harms caused due to harmful content online. For instance, failures in moderation on Facebook have resulted in physical harm and escalation of violence in countries like Myanmar and Ethiopia (Akinwotu, 2021) the spread of misinformation on WhatsApp led to violent attacks on minorities in India (Samuels, E. 2020); and YouTube users from countries that do not have English as their primary language are at 60% higher rate of being exposed to content they will "regret" watching (McCrosky et. al. 2021). Such lackluster moderation and failure of automatic detection for the majority of the world's languages emboldens malicious content creators to post policy-violating videos (Nigatu et. al, 2024).

Platforms use a combination of automated systems and human moderators to moderate content (Roberts 2019). Generally, automated content moderation involves using trained machine learning models

to determine if a post should be sanctioned due to breaches of policy, e.g., on hate speech and toxicity. However, not all users are protected equally (Dias Oliva, 2020). The field of natural language processing (NLP) has paid little attention to non-European languages, which has lead to a lack of data and technological resources to train robust automated detection systems. Moreover, platforms focus their efforts disproportionately on Western countries. For instance, in 2020 while 90% of its users live outside of the United States (US) and Canada, Meta (then Facebook) spent 87% of its time moderating posts in the US (Tworek, 2021). Such disparity is also reflected in moderation personnel: YouTube reports that 89.2% of its human moderators operate in English (Google, 2023), neglecting that 67% of videos are posted exclusively in languages other than English and 5% in multiple languages including English (Van Kessel et al, 2019).

The harm that speakers of the majority of the world's languages face in relation to content moderation extends beyond exposure to harmful content as users of online platforms. Big Tech companies hire content moderators from the Global Majority, which appears like an increased effort to protect users from those communities. However, these moderators often operate under deplorable working conditions and without fair compensation for conducting deeply traumatizing work (Perrigo, 2022). Such workers, who are often employed from African, South American, South East Asian, and South Asian countries, also provide labeled data for guardrails of Large Language Models like ChatGPT (Perrigo, 2023), models which do not work well in languages spoken by the Global Majority (Ojo et al., 2023), or are entirely unavailable.

Understanding and implementing effective policy to protect users of Global Majority must begin by uncovering what lies beneath blanket terminology that serves to obscure nuances; starting with the term Global Majority. While the term has been adopted as a reclaiming of power by appealing to the number of people grouped under it, it is still a blanket term covering several geographies, hundreds of cultures, and thousands of languages whose common predicament is exploitation by the powers on the other side — a concern that remains unresolved by the adoption of the term. Prior work has demonstrated the cultural nuances that result in the under-moderation

or over-moderation of online users from the "Global Majority" or "Global South" (Shahid et al, 2023). Hence, to effectively impact practical policies, we must start by examining these nuances and uncovering what is underneath the blanket terminologies.

In this essay, we first dive deeper into multitudes of harm faced by content moderators from the Majority world, reflecting on how the common denominator is exploitation. Then, we examine the current alternatives in online moderation which pose a false dichotomy for moderation to be effective, for which surveillance is an inevitable consequence. We call out the root problem that presents these alternatives as the only options. Next, we detail the social, political, and economic structures within the "Global Majority" to illustrate the nuances in different communities that would render blanket policies ineffective. Finally, we put forth a call to action to ensure the effective protection of "Global Majority" users on online platforms. We argue that what ties the experience of Global Majority people is the continued exploitation and disregard for well-being by Big Tech and states outside of the Global Majority, which bears similarities to exploitation by colonial bodies during the period of European colonization.

13.1 Discussion

13.1.1 The Cycle of Harm In Moderation and Inclusion

In 2021, Meta (then Facebook) faced scrutiny after a whistleblower, Frances Haugen, leaked internal documents detailing the harms the platform was fostering, in some cases not taking action to rectify the situation even after becoming aware of it (Horwitz, 2021). One trend in the moderation landscape has been to hire moderators in Global Majority countries, sometimes through third-party companies. However, the working conditions of the moderators are usually dire (Perrigo, 2022). While cases brought directly against companies like Microsoft and Meta have resulted in settlement payments and some policy changes for moderators hired directly by the companies (Newton, 2020), moderators hired by third-party companies risk mass layoffs and threats against forming unions (Perrigo, 2022). This double standard is a parallel to other exploitative work performed in "Global Majority" countries (e.g. the externalization of "Global

Minority" pollution and trash to the "Global Majority" (Liboiron, 2021)), where workers are treated differently for the same work when it is performed in "Global Minority" countries. The exploitation does not stop there. Perhaps ironically, such moderators are hired to moderate OpenAI models like ChatGPT, which do not work for the African languages that they speak (Ojo et al, 2023). In fact, ChatGPT was not available in countries like Ethiopia until November 2023 (Shega, 2023). In this way, the labor of the "Global Majority" is extractive, and the conditions under which moderators work are for the benefit of the privileged few who can operate the internet in languages like English and Spanish.

Communities from the "Global Majority" are exposed to harm (1) while using the platforms, due to weak platform policy enforcement and limited performance of technologies used in the moderation pipeline; (2) while moderating harmful content by virtue of exposure to traumatic content; (3) through poor working conditions and exploited labor; and (4) through technologies that exploit their labor but leave out their whole communities from whatever benefit the technology might provide. At the center of this cycle of harm is the exploitation and neglect of the wide swath of communities. The current systems that sustain the digital landscape are an extension of the history of colonization and exploitation that have ravaged the "Global Majority" (Kwet, 2019). Even when these communities are included in Artificial Intelligence research, they are treated as "bottom billion petri dishes"(Sambasivan et al, 2021, p.320)–their diversity and the weak policies protecting them make them an attractive test-bed for evaluating model robustness with little-to-no consequence or cost.

13.1.2 False Dichotomies of Harm: Either you are surveilled or you are left in the trenches.

Communities that have largely been excluded from policy and technological advances in the moderation space are exposed to harmful content daily. These unmoderated harmful content could be due to (1) policies that exist but are not enforced properly for these communities, or (2) policies that do not exist since the design of policies takes place under contexts that do not account for the diverse realities of "Global Majority." When policies do exist and are

under-enforced, malicious actors exploit the under-enforcement to propagate policy-violating content. As such, communities who have already been exploited by global structures are exploited again in our failure to effectively moderate online spaces.

When policies that reflect the diverse cultural context in the "Global Majority" simply do not exist, entire communities and cultures are left in a vacuum. Indeed, some companies seek to enforce a single standard upon all users, disregarding cultures, customs, and traditions. For instance, Facebook's one-size-fits-all approach resulted in the removal of a post of village kids swimming in a pond for violating the platform's policy against child nudity; although in the context of the poster, it is a common activity for children to swim naked in their local ponds to avoid "being scolded by their parents" (Shahid & Vashistha, 2023, p. 5).

With the rapid advances of Large Language Models and the "low-resource language" NLP community trying to increase the representation of these languages, harmful, toxic, and culturally nonrepresentative content on online spaces risks trickling down to model development and deployment. Generative models are trained using data from YouTube, Twitter, and general web scraping (Cole, 2024). However, training models for the majority of the world's languages present a particular risk as effective content moderation technologies and practices are not deployed for such languages. Thus, risks of harm are compounded by a lack of appropriate moderation, thereby compounding the risks of harm that have been documented for English (Talat et al. 2022).

Platforms that benefit from their users should adhere to their end of the bargain and provide a "positive experience for everyone on [their] platforms no matter where they [the users] are in the world" (Google, 2023, p. 8). Effective content moderation infrastructures, both human and automated, are required for safely building language technologies and content moderation technologies. However, many language technologies have risks of dual-use (Kaffee et al., 2023), including the risk of surveillance (Solaiman et al. 2023). It is therefore particularly important to consider how technologies are deployed and used, in addition to how data is gathered for the technologies themselves.

Here we would like to pause and reflect on what exactly effective moderation is, especially in the current context of the moderation pipeline. If the premise of moderation was not capitalistic and exploitative, could we have safer online experiences that put the power in users and not in companies that are out for profit?

13.1.3 What Lies Under Blanket Terminologies?

The degree and type of harm communities from the Global Majority face are shaped by the social, political, and economic realities of each community. Take two YouTube users studied by Nigatu & Raji, (2024) who studied the experiences of Ethiopian women on YouTube: a migrant domestic worker and a software engineer in the United States. Both users are Ethiopians, women, and of the Global Majority; yet have completely different realities. Migrant domestic workers cross borders to countries like Qatar and Lebanon en masse, either legally or via human traffickers. Once there, most of these women are subject to inhumane treatment, and sexual harassment and are often left without access to legal or medical services (Diab et al., 2023). Nigatu & Raji, (2024) show how these migrant domestic workers are exposed to harm through exposure to graphic and sexual videos while seeking medical help on online platforms. On the other hand, the Ethiopian Software Engineer living in the US is exposed to the same policy-violating content as the migrant workers when they search in their language. That is, a shift of location does not indicate a shift in types of policy-violating content. Change in policy enforcement might, for instance, remove policy-violating posts that expose both sets of users to harm. However, removal would not satisfy the need for information from the migrant worker, in this case, medical advice.

Political responses of different countries towards platform policies, or failures of platform policies also vary drastically. Countries like Ethiopia, Somalia, and Sudan ban online platforms when policies do not align with their values or when policies do not protect citizens from violent content. However, this has little impact on the actual problem as users resort to VPN services to access the platforms. Additionally, representatives for these platforms are most often subject to regulatory scrutiny in Global Minority countries,

even when the harms are primarily impacting people in the Global Majority. It is clear the platforms respond to the callouts by powerful governments; Europe has constantly been praised for the GDPR and its requirements against online harm to its citizens.

While the term "Global Majority" is an evolution from prior binaries based on social and economic status or geographic location (Khan et al., 2022), it is still a binary. The realities–and needs–of Indigenous and Aboriginal communities who continue to suffer the consequences of colonization and occupied land are different from those of African and Asian countries that faced the brunt of exploitation colonialism. Within the Global Majority several layers of class, ethnicity, and power result in the exploitation and harm of some communities over others. There is no single "AI from the Global Majority" because the "Global Majority" is many.

Call to Action: Throughout this essay, we have discussed the degree and depth of harm and exploitation that Global Majority users face. However, Global Majority users are not idly waiting for the mercy of the powers that be; to the extent that they can, they devise ways to protect themselves from harm[63]. We can augment their efforts by designing interventions that support them and relying on methods like participatory design as we build AI tools. Additionally, members of the Global Majority face layers of barriers to entering academic and policy spaces at a Global scale (Septiandri et al., 2023). Those who do make it, ourselves included, have degrees of privilege not afforded to the many who are organizing on the ground. Hence, we are responsible for engaging with community organizations–to the degree they are interested–to connect the academic and policy space with community organizing.

13.2 Conclusion

The manifold of communities that the "Global Majority" encompasses makes it challenging to enforce one-size-fits-all policies. The harms members of these communities face vary across the diverse social,

63 Instagram users create Fake-Instagram or "Finsta" accounts to share more intimate content with a close group of friends. A YouTube user in Nigatu & Raji (2024) study created multiple accounts for different aspects (religious, educational, and general) becuase she did not "want to be hit with disturbing content when [I] was watching a religious sermon or looking at a lecture.".

economic, and political axes each community has. Most of the current policies for protecting users in the digital age have been designed, tried, and tested in the "Global Minority" context. Our response to the fact that we have ignored the majority of the world's population in policy making and implementation should not be to blindly extend these policies to the communities we ignored. In moving from neglect to blind inclusion, we risk the exploitation of community members at several levels of the pipeline. Instead, we should focus our efforts on augmenting community efforts and building interventions that center community needs.

13.3 References

Akinwotu, E. (2021). Facebook's role in Myanmar and Ethiopia under new scrutiny. The Guardian. Retrieved from https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny.

Cole, S. (2024). Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI. 404 Media. https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project.

Diab, J. L., Yimer, B., Birhanu, T., Kitoko, A., Gidey, A., & Ankrah, F. (2023). The gender dimensions of sexual violence against migrant domestic workers in post-2019 Lebanon. Front. Sociol., 36741584. Retrieved from https://pubmed.ncbi.nlm.nih.gov/36741584.

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. Sexuality & Culture, 25(2), 700–732. doi: 10.1007/s12119-020-09790-w.

Google. 2023. U Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report. Technical Report. https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf.

Horwitz, J. The Facebook Files. (2021, October 01). The Wall Street Journal. Retrieved from https://www.wsj.com/articles/the-facebook-files-11631713039.

Kaffee, L.-A., Arora, A., Talat, Z., & Augenstein, I. (2023). Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing. ACL Anthology, 13977–13998. doi: 10.18653/v1/2023.findings-emnlp.932.

Khan, T., Abimbola, S., Kyobutungi, C., & Pai, M. (2022). How we classify countries and people — and why it matters. BMJ Global Health, 7(6). doi: 10.1136/bmjgh-2022-009704.

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. Race & Class. doi: 10.1177/0306396818823172.

McCrosky, J., Geurkink, B., Zawacki,K., Jay,A., Afoko, C., Gahntz, M., and Bennet, O.(2021) YouTube Regrets.https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.

Newton, C. (2020). Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. Verge. Retrieved from https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health.

Nigatu, H. & Raji, I.D. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. | Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. (2024, June 05). https://doi/10.1145/3630106.3658546.

Ojo, J., Ogueji, K., Stenetorp, P., & Adelani, D. I. (2023). How good are Large Language Models on African Languages? arXiv, 2311.07978. Retrieved from https://arxiv.org/abs/2311.07978v2.

Perrigo, B. (2022). Facebook Faces New Lawsuit Alleging Human Trafficking and Union-Busting in Kenya. Time. Retrieved from https://time.com/6147458/facebook-africa-content-moderation-employee-treatment.

Perrigo, B. (2023). Universities Are Wondering How to Adapt New Artificial Intelligence Tool ChatGPT. Time. Retrieved from https://time.com/6247678/openai-chatgpt-kenya-workers.

Roberts, S.T. Behind the Screen. (2024, August 12). Retrieved from https://yalebooks.yale.edu/book/9780300261479/behind-the-screen.

Sambasivan, N. & Arnesen, E. & Hutchinson, B. & Doshi, T. & Prabhakaran, V. Re-imagining Algorithmic Fairness in India and Beyond | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. (2021, March 01). https://doi/10.1145/3442188.3445896.

Samuels, E. (2020). How misinformation on WhatsApp led to a mob killing in India. Washington Post. Retrieved from https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india.

Septiandri, A.A., Constantinides, M., Tahaei, M., Quercia, D., 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT? In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 160–171. https://doi.org/10.1145/3593013.3593985.

Shahid, F. & Vashistha, A. (2023) Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? | Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. (2024, July 01). Retrieved from https://doi/10.1145/3544548.3581538.

Shega Team. (2023). ChatGPT Now Available in Ethiopia. https://shega.co/post/chatgpt-now-available-in-ethiopia.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., Daumé III, H., Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., Lin, M., Lin, X., Luccioni, S., Mickel, J., Mitchell, M., Newman, J., Ovalle, A., Png, M.T, Singh, S., Strait, A., Struppek, L., Subramonian, A. (2023). Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv, 2306.05949. Retrieved from https://arxiv.org/abs/2306.05949v4.

Talat, Z., Neveol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ..Van Der Wal, O. (2022). You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. ACL Anthology, 26–41. doi: 10.18653/v1/2022.bigscience-1.3 and Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv, 2403.00742. Retrieved from https://arxiv.org/abs/2403.00742v1.

Tworek, H. (2021). Facebook's America-centrism Is Now Plain for All to See. Centre for International Governance Innovation. Retrieved from https://www.cigionline.org/articles/facebooks-america-centrism-is-now-plain-for-all-to-see.

Van Kessel, P., Toor, S. and Smith, A. (2019). 1. Popular YouTube channels produced a vast amount of content, much of it in languages other than English. Pew Research Center. Retrieved from https://www.pewresearch.org/internet/2019/07/25/popular-youtube-channels-produced-a-vast-amount-of-content-much-of-it-in-languages-other-than-english/#:~:text=Meanwhile%2C%2067%25%20posted%20videos%20exclusively,the%20first%20week%20of%202019.

14 Countering False Information: Policy Responses for the Global Majority in the Age of AI

Isha Suri and Shiva Kanwar

Abstract

False information including misinformation and disinformation is being recognized as a severe global risk anticipated over the coming years. Access to generative artificial intelligence (AI) has dramatically increased the capacity for creating and disseminating falsified information. This is further compounded by algorithmic promotion of divisive content and creation of filter bubbles, leading to a precarious environment. We analyse the role of AI in exacerbating the false information crisis, evaluate regulatory responses to false information across various jurisdictions, and propose strategic policy recommendations for the Global Majority to effectively counter the threats of misinformation and disinformation in the age of AI.

Keywords: Misinformation; Disinformation; False Information; Artificial Intelligence; Algorithms; Algorithmic Bias; Recommender Systems; Intermediary Liability; Platform Governance; Content Moderation.

Introduction

The proliferation of false information poses a significant threat to societal cohesion and democratic integrity in the contemporary digital landscape. As open access to advanced technologies, particularly artificial intelligence (AI), becomes increasingly prevalent, the capacity for generating and disseminating falsified information has increased dramatically. Sophisticated AI models have democratized creation of synthetic content, including realistic images and videos, voice cloning and counterfeit websites, blurring lines between authentic and fabricated narratives. This phenomenon is further compounded by an erosion of trust in information sources and institutions, leading to a precarious environment where societal cohesion and legitimacy of electoral processes and governance are jeopardized.

In response, governments worldwide are implementing evolving regulatory frameworks to curb dissemination of false information online. These measures often grapple with the delicate balance between safeguarding free speech and mitigating risks associated with falsified information. Particularly in global majority nations, where the intersection of digital authoritarianism and false information may exacerbate political repression, the need for tailored policy responses becomes imperative. This essay seeks to analyse the role of AI in exacerbating the false information crisis, evaluate regulatory responses to falsified information across various jurisdictions, and propose strategic policy recommendations enabling the Global Majority to counter pervasive threats of false information in the AI age.

World Economic Forum's Global Risks Report 2024 recognises false information as the most severe global risk anticipated over the next two years (World Economic Forum, 2024; Ezrach, Stucke, 2022). As such, different jurisdictions have been grappling with this menace and its ability to undermine democratic ideals for the past decade, albeit with limited success. Today, a handful of dominant technology firms are largely responsible for how users traverse the internet. Acting as gatekeepers, these firms control access to digital markets, (Khan, 2019) including internet-based communication services. Social media companies operate in multi-sided markets, as intermediaries for distinct user groups. Predominantly, on one side they interact with users accessing the platform for generating content (the 'free side' of the market), and on the other, they sell placement for digital advertisers. (Stasi, 2023). Therefore, advertising-led business models dependent on massive data collection, profiling, and personalisation are a major source of revenue for social media platforms. (Article 19, 2023) Research suggests that toxic and fabricated content is likely to be more engaging, with one study reporting that disinformation was likely to spread six times faster than the truth. (Vosoughi et al., 2018) And recent research demonstrates that AI-generated disinformation may be more convincing than human-generated disinformation. (Williams, 2018) Multiple studies have demonstrated that social media is designed to reward and amplify divisive content, hate speech and disinformation. (O'Carroll, Elsayed-Ali, 2024). With algorithms designed to maximise user engagement, content likely to trigger user

attention is amplified including extreme content and content that contributes to formation of filter bubbles. (Ezrachi, Stucke, 2022) For instance, an internal study by Facebook revealed that its News Feed algorithms exploit the human brain's attraction to divisiveness and if left unchecked it would feed users "more and more divisive content to gain user attention and time over platform" (ibid,). Similarly, employees at Google sought to improve issues pertaining to filter bubbles and enhance diversity of content by modifying YouTube's recommendation algorithm. However, it reduced viewer retention (which would eventually reduce advertising income) because of which the change was suspended (ibid.). Therefore, owing to their integrated structures, and profit maximising incentives, platforms continue to employ algorithms that recommend divisive content.

14.1 Regulatory Responses

Regulatory responses to tackle false information can vary, For instance, some nations have proposed or enacted laws specifically targeting false information such as the European Union's Digital Services Act (DSA), while others ground their proposed amendments or legal frameworks on existing legislation, including penal codes, civil law, electoral law, or cybersecurity law (UNESCO et al., 2020). These regulatory frameworks either aim to hold perpetrators accountable as purveyors of false information or transfer the obligation to internet communication corporations to oversee or eliminate specific types of content (ibid.). The paper discusses DSA since it is the only regulation on recommender systems, and is therefore helpful in addressing various nuances involved in regulating algorithmic recommender systems. Germany's Network Enforcement Act is also discussed in an effort to contrast its intermediary liability framework with India, as outlined in the case study.

Ascribing criminal liability, particularly in cases where false information is defined broadly, carries significant risks of censorship (ibid.). For instance, Malaysia passed the Malaysia Anti-Fake News Act which criminalised the publication and dissemination of false news, punishable by up to six years in jail and a fine of $128,000. The law which was repealed in December 2019, made online service providers responsible for third-party content on their platforms (Poynter,

n.d.). Sri Lanka also amended its penal code in 2019 to prohibit fake news and hate speech that is "harmful to harmony between nations and national security" and enabled prosecution for spreading false statements or hate speech (id.). This law has been criticised for its potential to stifle free speech, usher in censorship, and facilitate mass surveillance (Schiffrin, Cunliffe-Jones, 2022).

Electoral regulations have also been used to combat false information. In the run-up to the 2018 general elections, Brazil introduced several draft bills criminalizing electoral misinformation with penalties ranging from fines to imprisonment for crimes ranging from spreading fake news stories on social media to publishing inaccurate press accounts (Poynter, n.d.). In 2019, Brazil amended its electoral code to define the crime of "slanderous denunciation for electoral purpose", with a penalty of two to eight years of imprisonment (UNESCO et al., 2020).

More recently, instances of AI deepfakes being used to manipulate political narratives and public opinion have been reported in countries including Moldova, Slovakia, and Bangladesh (Swenson, Chan, 2024). And, deepfakes pose a grave threat to democratic processes with consequences such as voter confusion and manipulation (ibid.). While measures such as labelling AI-generated content are being developed to combat deepfakes, they are ineffective in preventing the spread of false information (ibid.).

Legislative proposals have also sought to tackle this issue through intermediary liability for online platforms regarding false information or hate speech. Germany's Network Enforcement Act, 2017 mandates that for-profit social media platforms with over two million registered users are required to act against hate speech and offences outlined in the German criminal code. Such entities are required to implement transparent procedures for reporting content and managing complaints, and remove/block "manifestly unlawful" content within 24 hours and "unlawful" content within 7 days (UNESCO et al., 2020). Countries have also established specialized task forces to monitor and investigate false information campaigns. In 2018, Indonesia established the National Cyber and Encryption Agency intending to assist intelligence agencies and law enforcement to combat online misinformation and hoaxes

in anticipation of nationwide regional elections, although the specific authorities granted to this agency remain ambiguous (Poynter, n.d.).

The European Union's DSA introduces due diligence and transparency obligations regarding algorithmic decision-making by online platforms. It applies to all "intermediaries"[64] providing services in the EU and deems platforms and search engines with over 45 million monthly users in the EU as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) (European Commission, 2023). Failure to comply with any obligation under the DSA can result in a fine of up to 6 per cent of the annual worldwide turnover in the preceding financial year.[65] The DSA mandates enhanced transparency in recommender systems and advertising by requiring intermediary service providers to disclose their content moderation tools and algorithmic decision-making processes in their terms and conditions.[66]

Apart from regulatory responses, fact-checking, especially on social networks, is also being used to counter false information. While published fact-checks provide people with an authoritative source of information, they often receive fewer shares on social media than the mis/disinformation they aim to debunk (UNESCO et al., 2020).

Meta has the only large-scale international "third-party verification" programme among the dominant technology companies. Launched after the 2016 US presidential elections, the programme collaborates with independent fact-checking organizations, to assess accuracy of information on Facebook, Instagram and WhatsApp (Meta, n.d.). Fact-checkers are compensated by Meta. However, there is lack of transparency regarding payments made to the third-party fact-checking collaborators and ambiguity around the initiative's effectiveness in curbing the spread of false information (UNESCO et al., 2020). An increasing reliance on a system where more content is flagged initially by Meta's AI tools raises concerns about

64 Intermediary here includes social media platforms, search engines, online marketplaces, and internet service providers. See Baker (2024, April 4).

65 Article 52 Digital Services Act.

66 Article 14(1) Digital Services Act.

potential algorithmic errors, and concerns about Meta developing AI tools based on data acquired through partnerships from this programme (ibid.).

Fact-checking initiatives also face added challenges in the Global Majority with low digital literacy, lack of connectivity, and rural-urban and gender divides affecting efficacy. Multilingual societies also result in misinformation in regional languages being 'ignored' (Ugwa, Jain, 2023). Emerging research suggests that falsified information manifests differently across the globe, necessitating a nuanced and contextual approach to addressing the problem. For instance, during the COVID-19 pandemic, while India accounted for 16 per cent of global misinformation, the nature of content differed from that in the West. In the West, anti-vaccine related fake information gained traction, however, in India the myths ranged from using home remedies for treatment of COVID-19, thereby requiring distinct tactics from regulators and advocacy groups (Orsek, 2023).

14.2 India: Case Study

India relies predominantly on the Information Technology Act, 2000 (IT Act) and the recently amended Information Technology Rules to curb false information related harms in the country. Provisions such as Sections 69-A of the IT Act enable State Authorities to send content takedown orders to intermediaries whenever they find it "necessary or expedient" for national security, integrity, friendly relations with foreign states, and prevention of offences related to these grounds.[67] Any intermediary failing to comply with such an order is liable to pay a fine and/or face imprisonment for up to seven years.[68] While corresponding blocking rules provide a framework for implementation of the law, experience suggests that the rules cause an excessive restriction on freedom of speech and expression. For instance, research has highlighted that content creators are rarely notified or afforded a hearing (Gupta, 2015; Sakar, Grover, 2020). Furthermore, the blocking

67 Section 69A(1), Information Technology Act, 2000.

68 Section 69A(3), Information Technology Act, 2000.

rules also require confidentiality, effectively preventing content creators from viewing or challenging orders issued under them (Mukhopadhyay, 2019). Recently X also challenged the blocking rules in the Karnataka High Court arguing, inter alia, that blocking orders did not contain reasons recorded in writing and were not communicated to users, consequently preventing users from effectively challenging them. X also claimed that directions of the Ministry of Electronics and Information Technology (MeitY) to block entire accounts rather than specific tweets were disproportionate and excessive. However, a Single Judge Bench of the Court rejected X's arguments and dismissed their challenge and imposed a penalty of Rs. 50 lakh (US$59,655) (*X Corp* v. *Union of India*, 2022). An appeal against the order is currently pending before a Division Bench of the High Court ("Karnataka High Court stays order…", 2023).

Furthermore, empirical evidence from India confirms that online blocking solely at the discretion of the executive has far-reaching effects on freedom of expression (Sehgal, Grover, 2023). If an intermediary is legally obligated to respond to an overwhelming number of content takedown requests under the fear of losing its legal immunity, they are likely to over-comply to avoid sanction. This has the potential to chill online expression.[69]

Indian law has no regulations dealing with algorithms used by intermediaries and their potential harms, thereby limiting its ability to effectively counter AI-fuelled false information. Although reports suggest that the proposed Digital India Act will have provisions pertaining to algorithmic accountability, there remains ambiguity around the legislation and its timelines for implementation (Ministry of Electronics and Information Technology, 2023).

69 See Dara (2011).

Table 1 A Brief Outlook on Regulatory Responses to False Information

Country	Instrument/Response	Status	Criminal Sanctions	Intermediary Liability	Transparency and Accountability Provisions
Argentina	Bill on creating a Commission for the Verification of Fake News 2018	Not passed	✓	✓	
Bangladesh	Digital Security Act 2018	In force	✓		
Brazil	Draft Bills 2018	Not passed	✓		
Brazil	Amendment to electoral code 2019	Passed by Congress	✓		
Egypt	Regulating the Press and Media 2018	In force	✓		
Egypt	Anti-Cybercrime Law 2018	In force	✓	✓	
Egypt	Penal Code	In force	✓		
European Union (EU)	Digital Services Act 2022	In force	✓	✓	✓
Germany	Network Enforcement Act 2017	In force	✓		✓
India	Information Technology Act, 2000 Information Technology Rules, 2009	In force	✓	✓	
Kenya	Computer Misuse and Cybercrimes Act 2018	In force	✓		
Malaysia	Anti-Fake News Act 2018	Repealed	✓	✓	
Nigeria	Anti-social Media Bill; Originally titled Protection from Internet Falsehood and Manipulation Bill 2019	Not passed	✓	✓	
Singapore	Protection from Online Falsehoods and Manipulation Act, 2019	In force	✓	✓	✓
Sri Lanka	Amended Penal Code 2019	In force	✓		

Source: Compiled by Authors

14.3 Conclusion and Way Forward

Countering the menace of fake information poses significant challenges for policymakers worldwide. And the advent of generative AI tools has further exacerbated the problem. Amongst other things, it requires treading a fine balance between restricting harmful speech without violating fundamental rights to free speech and expression. However, it will require efforts from all stakeholders including platforms, policymakers, and regulators to effectively address these threats. Based on our research, we recommend:

Consider unbundling platforms: The internet today is characterised by large social media platforms controlling the flow of information and communication between users. These dominant platforms rely on advertising as a source of revenue, with most of them being the largest

providers of online advertising services, thereby creating a conflict of interest that requires intervention by regulators (Stasi, 2023). To counter this problem, regulators should consider unbundling content curation services[70] (excluding content moderation services) from content hosting services within a platform (Statsi, 2021). Given that large social media platforms are global in nature, decisions taken in one jurisdiction are likely to have a spillover effect in another. For example, in Germany, as part of a remedy to respond to competition concerns from third-party sellers, Amazon agreed to amend its terms of business for sellers on Amazon's online marketplaces across Europe, North America, and Asia ("Amazon in deal with German watchdog…", 2019). Such cross-country benefits could be further leveraged by promoting international cooperation between antitrust authorities across jurisdictions. This could also create opportunities to shift revenue models away from advertising and disincentivise promoting user engagement with divisive content.

Adopting a co-regulatory approach: Regulations alone will struggle to eliminate false information from digital platforms; a comprehensive strategy involving efficient regulatory interventions, along with self-regulation by platforms is required. A co-regulatory response involving government and platforms working together is likely to achieve better results. Through such a collaboration, regulators could gain access to information on algorithmic recommender systems, and make better decisions on how to shape their design to achieve desired outcomes. A co-regulatory response tailored to each jurisdiction's requirements is also likely to make enforcement easier (ibid.).

Develop inclusive AI-assisted tools for content moderation: Automated hate speech detection systems that have shown success in English and European languages struggled in countries such as Myanmar, India, and Ethiopia, due to lack of cultural contextualisation (Udupa et al., 2022). Miscreants evade keyword-based machine

70 This paper the term "content curation" has been defined as the measures taken by social media platforms that affect the availability, visibility and accessibility of content, such as ranking, promotion, demotion. These measures are performed by fully or partially automated systems based on algorithms. Content curation differs from content moderation, which usually indicates the activities undertaken by social media platforms to detect, identify and address illegal content or content incompatible with their terms and conditions, such as demotion and removal.

detection through astute combinations of words, misspellings, satire, changing syntax and coded language (ibid.). Platforms must develop AI tools using diverse high-quality data sets, and employ local teams proficient in the local language and cultural context.

Global majority countries must consider domestic realities before succumbing to the Brussels effect: The DSA introduces due diligence and transparency obligations regarding algorithmic decision-making by online platforms that complement other EU AI regulatory efforts such as the AI Act (Chander, 2023). Adoption of such regulatory provisions without accounting for contextual realities, especially by authoritarian regimes and fragile democracies can leave nations vulnerable to potential misuse (ibid.). It could also incentivize platforms to adopt uniform content moderation policies that align with European standards, which, while promoting global consistency, may inadvertently suppress local norms and practices in global majority countries, resulting in over-censorship (ibid.). Furthermore, emulating complex regulations such as the DSA may pose challenges for developing countries, which often lack administrative and judicial capacities required for effective implementation, thereby increasing the risk of inconsistent application and exploitation by powerful entities. While comparative regulatory analysis is helpful, countries should tailor these regulations through studies grounded in their jurisdictions and also enhance regulatory capacity, where required.

Promote transparency in content recommender systems: Most regulatory and legislative responses focus on content moderation from the lens of eliminating potentially harmful user-generated content without addressing how individual pieces of content achieve high impact through recommender systems. Prioritizing transparency in recommender systems is essential to tackle harms that arise from algorithms promoting divisive content. It enhances comprehension of algorithmic decisions, fosters trust, and alleviates bias and privacy concerns while ensuring compliance with ethical AI standards.

14.4 References

Amazon in deal with German watchdog to overhaul marketplace terms. (2019, July 17). CNBC. Retrieved October 21, 2024, from https://www.cnbc.com/2019/07/17/amazon-in-deal-with-german-watchdog-to-overhaul-marketplace-terms.html.

Article19. (2023, January). *Taming Big Tech: A pro-competitive solution to protect free expression*. Retrieved from https://www.article19.org/wp-content/uploads/2023/02/Taming-big-tech-UPDATE-Jan2023-P05.pdf.

Baker, G. (2024, April 4). The EU Digital Services Act: A Win for Transparency. Retrieved from https://freedomhouse.org/article/eu-digital-services-act-win-transparency.

Dara, R. (2011). Intermediary Liability in India: Chilling Effects on Free Expression on the Internet. *Centre for Internet & Society*. Retrieved from https://cis-india.org/internet-governance/intermediary-liability-in-india.pdf.

European Commission. (2023, April 25). Press Release — Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413.

Ezrachi, A., & Stucke, M. E. (2022). How Big-Tech Barons Smash Innovation — and How to Strike Back. Harper Collins.

Ezrachi, A., & Stucke, M. E. (2022). How Big-Tech Barons Smash Innovation — and How to Strike Back. Harper Collins.

Gupta, A. (2015, March 27). But what about Section 69A? *The Indian Express*. Retrieved from https://indianexpress.com.

Karnataka High Court stays order imposing Rs 50 lakh fine on X Corp. (2023, August 10). *Business Standard*. Retrieved from https://www.business-standard.com.

Khan, L. M. (2019). The Separation of Platforms and Commerce. *Columbia Law Review*, *119*(4). Retrieved from https://columbialawreview.org/content/the-separation-of-platforms-and-commerce.

Meta. (n.d.). Meta's Third-Party Fact-Checking Program. Retrieved from https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking.

Ministry of Electronics and Information Technology. (2023). Proposed Digital India Act, 2023. Retrieved from https://www.meity.gov.in/writereaddata/files/DIA_Presentation%2009.03.2023%20Final.pdf.

Mukhopadhyay, D. (2019, December 16). Delhi HC issues notice to the government for blocking satirical Dowry Calculator website. Retrieved from https://internetfreedom.in/delhi-hc-issues-notice-to-the-government-for-blocking-satirical-dowry-calculator-website/.

O'Carroll, T., Elsayed-Ali, S. (2024, August 20). Musk is the symptom, Big Tech is the crisis. Retrieved from https://www.linkedin.com/pulse/musk-symptom-big-tech-crisis-sherif-elsayed-ali-wppge/?trackingId=ZoQKBSHBQ4GP6PxMlTygHQ%3D%3D.

Orsek, Baybars. (2023, June 5), How India can show the way in combatting fake news in the Global South. Retrieved from https://indianexpress.com/article/opinion/columns/india-show-way-combatting-fake-news-global-south-8646961/.

Poynter. (n.d.). A guide to anti-misinformation actions around the world. Retrieved August 25, 2024, from https://www.poynter.org/ifcn/anti-misinformation-actions/#malaysia.

Sakar, T., & Grover, G. (2020, February 15). How India is using its Information Technology Act to arbitrarily take down online content. *Scroll.in*. Retrieved from https://scroll.in/article/953146/how-india-is-using-its-information-technology-act-to-arbitrarily-take-down-online-content.

Schiffrin, A, Cunliffe-Jones, P. (2022). Online Misinformation: Policy Lessons from the Global South. In H. Wasserman, D. Madrid-Morales (Ed.). *Disinformation in the Global South*. (pp. 161-178). USA: Wiley-Blackwell.

Sehgal, D., & Grover, G. (2023, April). *Online Censorship: Perspectives From Content Creators and Comparative Law on Section 69A of the Information Technology Act*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4404965.

Stasi, M. L. (2023). Social media markets: A pro-competitive approach to free speech challenges. [Doctoral Thesis, Tilburg University].

Stasi, M.L. (2021). Unbundling hosting and content curation on social media platforms: between opportunities and challenges. UCLA Journal of Law and Technology, 28 (2).

Swenson, A., & Chan, K. (2024, March 14). Election disinformation takes a big leap with AI being used to deceive worldwide. Retrieved from https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd.

Udupa, S., Maronikolakis, A., Schütze, H., & Wisiorek, A. (2022, June). *thical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence*. The Shorenstein Center on Media, Politics and Public Policy. Retrieved from https://shorensteincenter.org/wp-content/uploads/2022/06/Ethical-Scaling.pdf.

Ugwa, J., & Jain, M. (2023, December 18). Big tech 'failing' to curb fake news in global South. Retrieved from https://www.scidev.net/global/scidev-net-investigates/big-tech-failing-to-curb-fake-news-in-global-south/.

UNESCO, International Telecommunication Union, & Broadband Commission for Sustainable Development. (2020). Balancing Act: Countering Digital Disinformation while Respecting Freedom of Expression: Broadband Commission Research Report on 'Freedom of Expression and Addressing Disinformation on the Internet'. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379015.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. Retrieved from https://www.science.org/doi/10.1126/science.aap9559.

Williams, R. (2023, June 28). Humans may be more likely to believe disinformation generated by AI. MIT Technology Review. Retrieved October 21, 2024, from https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/.

World Economic Forum. (2024, January). The Global Risks Report 2024. weforum.org. Retrieved from https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf.

X Corp v. *Union of India*, W.P. No. 13710 of 2022, Karnataka High Court, Order dated 30-06-2023.

15 Addressing the Challenges of AI Content Detection in the Global South

Richard Ngamita

Abstract

The rapid adoption of artificial intelligence (AI) in content creation has raised significant content moderation challenges, particularly in the Global South, where 'cheapfakes' — manipulated media created with basic tools — pose a serious threat. Existing detection systems, primarily designed for deepfakes, are inadequate for cheapfakes, which exploit low-tech environments to spread misinformation. To address this, initiatives must focus on developing AI models trained on local data, enhancing research and development, and implementing inclusive content moderation policies. These efforts protect civic participation and democracy in the Global South.

Keywords: Deepfakes, Cheapfakes, AI, content moderation, Global South.

Introduction

The widespread use of artificial intelligence (AI) in content creation has posed significant challenges for content moderation, particularly in the Global South. While much attention has been given to detecting deepfakes, there is growing concern about the more common threat of 'cheapfakes' — AI-manipulated media created using basic editing tools. These cheapfakes can have serious consequences in regions with limited technological infrastructure, where misinformation or disinformation can easily incite violence and political instability. Current detection mechanisms, primarily designed for deepfakes, are insufficient for identifying cheapfakes, which include manipulated audio and video created with minimal resources. These types of content can be easily spread across social media platforms, making them difficult to detect and regulate (Paris & Donovan, 2019).

In 2023, Chinese smartphone brands such as iTel, Infinix, Huawei, and Tecno captured a 48% market share in Africa (Statista, 2023). While

these devices have made digital technology more accessible, they often produce low-quality video content. This presents a challenge for automated detection systems, which may mistakenly flag these videos as fake, not due to manipulation, but simply because of their inherently poor quality. This issue highlights the limitations of current detection technologies, which are often ill-equipped to consider the context in which content is created and consumed, especially in the Global South.

Adding to this complexity is the significant geopolitical influence of China, which plays a significant role in shaping Africa's technological landscape. China's strategic economic and political engagement in Africa has facilitated the widespread adoption of its smartphone brands. While these devices are affordable and provide much-needed access to technology, they have raised concerns about surveillance and propaganda. Chinese technology companies, often influenced by state directives, may embed software that enables data tracking and collection. This duality — affordable access alongside the potential for digital surveillance — complicates the benefits of these smartphones, particularly in terms of privacy and control over information flow.

The issue is further compounded by the infrastructural challenges faced by countries in the Global South. Limited access to high-speed internet and reliance on low-end smartphones result in a higher prevalence of low-quality content. Videos created under these conditions are often flagged as suspicious by AI detection tools — not because they have been tampered with, but because poor video quality is mistakenly linked to inauthenticity. This not only leads to false identifications but also undermines the credibility of legitimate content from these regions. Thus, the interplay of technological limitations, geopolitical influences, and infrastructure challenges creates a precarious digital environment, where access to technology can both empower and marginalize.

While the Global South faces challenges related to infrastructure and cheapfakes, the Global North contends with the more sophisticated threat of deepfakes. Politically motivated deepfakes have increasingly been used to manipulate public opinion. For example, a recent instance in the U.S. involved a fake voice message falsely claiming to be from President Joe Biden, which was sent to voters in New Hampshire during the primary election to discourage voting. Although

the nature of manipulated media varies between the Global North and South, the dangers remain significant in both contexts — cheapfakes in the South, given their ease of creation, and deepfakes in the North, due to their technical complexity (Lewandowsky et al., 2012).

For example, a poorly edited video showing a political figure endorsing a controversial policy could spread quickly, especially in places with limited access to reliable news sources. A cheapfake featured Donald Trump endorsing Umkhonto we Sizwe (MK) and encouraging South Africans to vote for the party. Another involved an AI-generated video of Joe Biden falsely claiming that if the ANC won the election, the USA would impose sanctions on South Africa. Additionally, a manipulated image of Julius Malema of the Economic Freedom Fighters (EFF) appeared to show him crying after a perceived political defeat.

15.1 Results

Platforms like Meta, YouTube, and TikTok have introduced content moderation guidelines to address manipulated media, but these measures are largely focused on deepfakes. For example, Meta's manipulated media policy applies primarily to deepfakes, while YouTube's misinformation policy targets content that poses a risk of egregious harm (Meta, 2023; YouTube, 2023). TikTok prohibits AI-generated realistic scenes of fake people unless labelled by the creator. However, these policies inadequately address the proliferation of cheapfakes, which pose a more immediate threat to civic participation in the Global South.

Another challenge for detecting and moderating AI-generated content in the Global South is the region's linguistic and cultural diversity. Many AI detection tools are trained on datasets that primarily consist of content in English or other widely spoken languages. This limits the effectiveness of these tools in detecting manipulated content in languages underrepresented in training data, leading to gaps in detection capabilities across different regions.

Moreover, the Global South's socio-political context presents additional content moderation challenges. Over 70% of the world's population lives under authoritarian regimes, primarily in low- and middle-income countries (Freedom House, 2023). In these

environments, the disclosure of AI-generated content or the identity of the content creator could lead to severe repercussions, including imprisonment or worse.

15.2 Recommendations

To effectively address these challenges, several initiatives can be proposed to enhance AI research and development focused on content detection in the Global South to address these challenges. One key approach is developing AI models trained on local data. This would involve collecting and annotating large datasets of content from the Global South, including texts, images, videos, and audio in local languages and dialects. By training AI models on this data, detection tools would be better equipped to recognize the nuances of manipulated content in these regions.

Collaborative efforts between local governments, academic institutions, and international organizations are essential to support research and development in this area. Funding should be directed towards building the necessary infrastructure for data collection and analysis, as well as for training local researchers and developers. This would not only improve the detection of AI-generated content but also empower local communities to participate in the global conversation on AI ethics and regulation.

One such initiative is Thraets, a company that is actively involved in combating the spread of AI-generated misinformation and disinformation, particularly in Africa. Through initiatives like the 'Safeguarding African Elections' project, Thraets is working to develop open-source AI tracking tools and knowledge hubs that focus on monitoring AI-generated content related to elections. This is particularly significant in regions where the proliferation of cheapfakes — manipulated media created with basic tools — poses a threat to civic participation and democratic processes (Thraets, 2024). Thraets also trains journalists and civil society organizations to detect and counter AI-generated disinformation. This capacity-building effort is especially crucial in regions where resources and expertise are often limited, and where the impact of misinformation can be particularly destabilizing. Thraets' efforts

represent a significant step forward in the fight against AI-generated disinformation in the Global South.

An important initiative can be the development of clearer and more inclusive content moderation policies by social media platforms. These policies should explicitly address the issue of cheapfakes and outline specific measures for detecting and mitigating their spread. Platforms should also invest in tools that allow users to report suspected manipulated content and provide clear guidelines on how this content will be reviewed and acted upon.

It's important to prioritize raising awareness about the dangers of manipulated media in the Global South. Educational campaigns should aim to improve digital literacy and critical thinking skills among the population to reduce the impact of misinformation. These campaigns should be conducted in local languages and customized to the specific cultural contexts of different regions.

15.3 Conclusion

To effectively combat the issue of cheapfakes and ensure digital inclusivity, it is particularly essential to develop AI models trained on local data, support research and development initiatives, and implement clearer and more inclusive content moderation policies. We can better protect the citizens of the Global South from the harmful effects of manipulated media and ensure that they can participate fully in the digital age by taking these steps.

15.4 References

Freedom House. (2023). *Freedom in the World 2023*. Retrieved from https://freedomhouse.org/report/freedom-world/2023/global-2023.

Paris, B., & Donovan, J. (2019). *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. Data & Society.* Retrieved from Data & Society — Deepfakes and Cheap Fakes (datasociety.net).

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131. Retrieved from https://doi.org/10.1177/1529100612451018.

Meta. (2023). *Community Standards on Manipulated Media*. Retrieved from https://www.facebook.com/communitystandards/manipulated_media.

Statista. (2023). *Smartphone market share in Africa in 2023*. Retrieved from https://www.statista.com/statistics/1171017/smartphone-market-share-by-vendor-in-africa/.

Thraets. (2024). *Thraets secures grant to protect African elections from AI-generated mis/disinformation*. Retrieved from Thraets.

YouTube. (2023). *Misinformation policy*. Retrieved from https://support.google.com/youtube/answer/10834785?hl=en.

16 Bridging the gap between the North and South in the governance of dual-use artificial intelligence technologies

Guangyu Qiao-Franco and Mahmoud Javadi

Abstract

This article examines the complex challenges of regulating dual-use artificial intelligence (AI) technologies within international arms control frameworks, amid a growing divide between the Global North and Global South. The intangible nature and dual-use potential of AI make traditional monitoring, verification, and classification methods ineffective. Developed nations are integrating civilian AI research into defense applications and imposing strict access controls to maintain military advantages, which exacerbates geopolitical tensions and stifles global innovation. In contrast, many Global South countries, unable to match these technological advancements, advocate for outright bans on autonomous weapons systems to mitigate their disadvantages. This dynamic undermines global cooperation and increases the risk of interstate conflict. The article advocates for a paradigm shift toward inclusive AI governance that addresses the needs and aspirations of both developed and developing nations. By fostering international dialogue, capacity building, and equitable access to AI technologies, it proposes establishing a transparent, multilateral framework for responsible AI use to bridge the North-South divide, reduce tensions, and promote global security and prosperity.

Keywords: Artificial Intelligence, Dual-Use Technologies, North-South Divide, AI Governance.

Introduction

Artificial intelligence (AI), a key driver of economic growth, holds significant implications for international peace and security. In the early 2010s, concerns about the autonomous use of force enabled by AI prompted intergovernmental negotiations on arms control under the United Nations Convention on Certain Conventional Weapons

(CCW). These discussions have revealed a growing divide between the Global North and the Global South regarding the military use of AI and regulatory approaches. Over a decade later, this gap appears to be widening rather than closing.

The dual-use nature of AI, which allows for both civilian and military applications, further complicates the path to a comprehensive arms control agreement. Developed countries are increasingly integrating civilian research and development (R&D) into defence, raising concerns about the military use of dual-use technologies by adversaries. This has led to stricter access controls, such as the United States tightening semiconductor export restrictions to China, supported by Japan and the Netherlands (Allen, et al., 2023). These restrictions permeate the civilian domain and raise security concerns.

In response, many Global South countries, unable to develop AI weapons, have opted for an outright ban on the use of autonomous systems to offset their technological disadvantage (Bode & Qiao-Franco, 2024). Meanwhile, emerging economies have taken a more rigid stance in military AI governance due to fears that broader control measures might be imposed under the guise of national security. For instance, China's unexpected abstention on a UN General Assembly resolution concerning lethal autonomous weapons systems (LAWS) in 2023 contradicted its earlier support for a legal ban on LAWS at the UNCCW. This has contributed to growing distrust and tension between states, undermining efforts to build confidence and coordinate on AI governance, while increasing the likelihood of extreme responses that could trigger interstate conflict.

To prevent this negative trajectory, a stepwise paradigm shift is needed in arms control regarding dual-use AI. Measures must account for the needs of both the Global South and Global North. International dialogue and partnerships should be fostered to promote capacity building, knowledge transfer, and inclusivity. These initiatives would help create an incentive structure encouraging responsible AI use and broader engagement. Ultimately, whether AI is used for peaceful or military purposes is determined by social factors. A new arms control paradigm should address the current insecurity dynamic, reduce the push for rival states to accelerate civil-military technology transfers, and pave the way for a 'global AI order' (Kissinger & Allison, 2023).

This article outlines the inherent challenges of controlling dual-use technologies and emphasises the different economic conditions and aspirations of the Global North and Global South. It concludes by proposing measures for achieving a harmonised approach to AI security regulation, aiming to build an inclusive arms control regime for AI safety.

16.1 Intricacies and Challenges of Arms Controls for Dual-Use AI

AI is an intangible technology, unlike other tangible and recognisable technologies, making traditional restrictive measures less, if not entirely, applicable and effective. Three main reasons justify this challenge. First, the intangible nature of AI software enables effortless cross-border transfer, circumventing monitoring by enforcement agencies (Brockmann, 2022). Unlike physical goods, AI algorithms can be transmitted digitally across borders with little to no physical trace, making it difficult for authorities to track and regulate their movement effectively.

Secondly, the verification of AI capabilities is complex due to the extensive lines of code involved, rendering it challenging for enforcement agencies to assess (Kaur et al., 2023). Unlike conventional technologies where physical characteristics can be examined, AI systems often consist of intricate algorithms with millions of lines of code, making it daunting to verify their functionalities, especially when those functionalities could have both benign and harmful applications.

In addition to the monitoring and verification challenges, AI is increasingly provided as a service rather than a standalone product, complicating export controls and oversight of its use across multiple countries (Klein & Patrick, 2024). With the rise of cloud computing and Software-as-a-Service (SaaS) models, AI capabilities can be accessed remotely, blurring the lines of jurisdiction and making it challenging for regulators to enforce compliance with arms control and usage restrictions (Cespedes & van der Kooij, 2023).

The dual-use nature of AI introduces another layer of hurdles in classification and regulation. Unlike other revolutionary technologies, whose progress relies heavily on government investments, AI

technologies are propelled forward by private actors, ranging from technologists and entrepreneurs to corporations (Perifanis & Kitsios, 2023). Restrictive measures will likely pose significant risks to global commerce and can provoke dissent among private sectors reliant on overseas markets.

States, primarily from the Global North, have developed national frameworks and transnational regimes — such as the Wassenaar Arrangement and the Australia Group — to maintain control lists for dual-use items. However, the composition of these lists remains subjective and politically driven, largely due to the absence of international consensus on the definitions and scope of dual-use technologies (Benson & Putnam, 2023). In the absence of established criteria for controlling dual-use AI within existing transnational regimes, these Global North states have increasingly asserted their authority by imposing restrictions on access to AI technologies, their components, and applications. Managing AI items on these lists is particularly challenging given the widespread use of general-purpose AI software. Excessive access controls designed to limit the export or use of AI technologies with dual-use potential risk stifling innovation, hindering economic growth, and unnecessarily escalating geopolitical tensions.

16.2 The Widening Gap between the Global North and Global South

While some emerging economies, such as China, India, and Turkey, are becoming leading technology innovators, most of the Global South, particularly the poorer regions, can only adopt AI technologies previously developed in the Global North. The significant technological gap has led to differing views on issues such as the adequacy of the existing legal framework to regulate autonomous weapons, the permissible forms of AI use in armed conflict, and measures to ensure human control, as discussed during the UNCCW negotiations over the regulation of military AI (Bode et al., 2023). While several Global South countries, in collaboration with a few small developed states and civil society, have succeeded in securing a mandate to negotiate a new legally binding instrument on lethal autonomous weapons at the UN General Assembly (UNGA, 2024), this instrument is unlikely

to be endorsed by developed nations, which seek to modernise their armed forces to maintain a military advantage.

Instead of focusing their efforts on UN negotiations, several AI safety initiatives have emerged in the Global North, including those within the G7, OECD, NATO, and the EU. Other inclusive multilateral frameworks, led by countries such as the Netherlands, South Korea, the UK, and the US — such as REAIM (Government of the Netherlands, 2023), the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (U.S. Department of State, 2023), and the Bletchley Declarations (UK Prime Minister's Office, 2023) — have not been well received in the Global South. In contrast, Global South forums such as BRICS, ASEAN, and the African Union have primarily concentrated on AI's developmental potential, particularly its impact on the digital economy, with security concerns often receiving less attention (See e.g., Jin, 2024; ASEAN Secretariat, 2024; African Union, 2024). Consequently, the North-South divide in the global governance of dual-use AI technologies appears to be widening.

Programmes aimed at transferring technology to bridge capability gaps in various sectors have proven difficult to sustain in the field of AI, largely due to concerns in developed countries about the potential for malicious use. The Global North, particularly the United States, prioritises maintaining its qualitative edge in AI, often monopolising technology and securitising access to prevent its diffusion from civilian to military applications. Common measures include export controls, foreign investment reviews, and the suspension of R&D partnerships (Moller-Nielsen, 2024). Notable examples include US de-risking policies (The White House, 2023a), NATO's Defence Innovation Accelerator for the North Atlantic (DIANA) (NATO, 2024), and the Action Plan on Synergies between Civil, Defence, and Space Industries (European Economic and Social Committee, 2021).

The need for access control conflicts with the desire for rapid advancements in the Global South, especially among emerging economies subject to these stringent measures. China, for instance, has persistently sought to acquire and develop AI technologies, using them to advance various domestic and international agendas. The 2024 remarks delivered by Chinese Prime Minister Li Qiang at the World Economic Forum highlight these diverging perspectives,

criticising the restrictions on technology access and innovation while advocating for more open technological cooperation (WEF, 2024).

The restrictions on access to AI technologies, even in civilian domains, have exacerbated geopolitical tensions, diminishing the sense of security among states and making meaningful progress in cooperative military AI governance increasingly unlikely. The US's "chip war" with China offers a pertinent example. On 7 October 2022, the Biden administration issued new regulations (U.S. Department of Commerce, 2022) limiting US exports of advanced AI chips and Chinese acquisitions of companies capable of producing chips smaller than 14 nm. This was followed by an Executive Order in August 2023, establishing mechanisms to limit outbound investment in sectors such as semiconductors, quantum information, and AI in China and other designated countries of concern (The White House, 2023b). In response, the US undertook extensive efforts to dissuade countries in the Middle East and Africa from maintaining ties with Chinese technology companies.

China's reaction was swift: it imposed licensing requirements on the export of rare-earth metals, such as gallium and germanium, and their derivatives, which are essential for semiconductor manufacturing (Shivakumar et al., 2024). Additionally, following the restrictions from Washington and its allies, China has refocused its military-civil fusion-driven semiconductor investment policies to enhance state autonomy (Waldie, 2022). These policies have supported less competitive enterprises, facilitated the substitution of outdated foreign chips with domestically produced alternatives in critical military equipment, and allowed military-focused research to continue without fear of foreign embargoes. In a likely response to Western restrictions, China, the world's second-largest military spender, allocated an estimated €270 billion to its military in 2022, accounting for 13 per cent of global military spending. This represents a significant 63 per cent increase since 2013 and a 4.2 per cent rise from 2021 (Tian et al., 2023).

Although national measures like those adopted by Washington and Beijing — while not exclusive to these countries (Sterling, 2023) — aim to control access to dual-use AI, they risk reinforcing protectionism

and isolationism, worsening global geopolitical dynamics rather than effectively managing dual-use AI regulation.

In addition to triggering a securitisation spiral that reduces both the Global North's and Global South's sense of security, this imbalanced regulatory approach may lead to a race to the bottom in AI safety standards. States may be incentivised to adopt more lax safety regulations to attract investment in AI industries, while competitive pressures could prompt AI producers to release products prematurely, sacrificing thorough testing and risk management.

16.3 Towards A Paradigm Shift for Dual-Use AI Governance

Governing dual-use technologies, particularly AI, necessitates a paradigm shift that reconsiders the multifaceted benefits and threats these technologies pose to nations across both the Global North and Global South. This shift involves identifying shared interests and common challenges to foster international collaboration and build consensus. Scholarly analyses and policy proposals (Kissinger & Allison, 2023; Reppy, 2006) emphasise the urgency of this approach, a sentiment echoed by the adoption of the United Nations' Global Digital Compact in September 2024 (Reiland, 2024).

AI's pervasive impact on various dimensions of human life — economic, social, and political — makes its governance especially critical. Implementing AI export controls and arms control mechanisms is vital to prevent the malicious proliferation of AI technologies that could compromise global security. However, when states exploit and weaponise AI against one another, it undermines efforts to establish a global AI governance framework essential for maximising benefits while minimising risks.

For developing nations, AI offers unprecedented opportunities for economic growth and social advancement. To realise these benefits, it is imperative that the Global South is actively included in global AI governance discussions. Inclusive policies are crucial to prevent the widening of the technological divide and to ensure that AI contributes to poverty eradication and sustainable development in less-developed regions.

To this end, Track Two and Track 1.5 diplomacy — facilitated by epistemic communities such as technologists, scientists, and industry leaders — provide promising avenues for initial engagement (Qiao-Franco, 2022). These non-governmental channels can foster mutual understanding, build trust, and promote informed discussions on managing dual-use AI technologies. This is particularly important in contexts where influential nations, such as the United States and China, may perceive each other antagonistically. By facilitating nuanced debates and identifying common ground, these communities can develop pragmatic solutions that balance national security concerns with the imperatives of innovation and competitiveness.

These efforts can lay the groundwork for an inclusive and transparent dual-use AI control framework within a multilateral setting, open to all states and viewpoints. Incorporating measures to bridge the North-South gap — such as technology transfer agreements, capacity-building initiatives, and equitable access to AI advancements — can promote understanding and trust between developed and developing nations.

The proposed framework should aim to fulfil the security needs of developed nations by preventing malicious AI use while simultaneously addressing the goals of developing nations for economic and social development. This includes supporting poverty eradication through AI-driven solutions in sectors like healthcare, education, and agriculture. By fostering a global AI order free from weaponisation and politicisation, AI can serve as a tool for global good rather than a source of conflict.

Ultimately, mitigating geopolitical tensions and enhancing global stability reduces the impetus to convert civilian technologies into military applications. By actively bridging the North-South gap and cultivating an inclusive international environment, the international community can harness AI's transformative power to promote global prosperity and security for all nations.

16.4 References

African Union. (2024). Continental artificial intelligence strategy. Retrieved October 21, 2024, from African Union website: https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

Allen, G. C., Benson, E., & Putnam, M. (2023, April 10). Japan and the Netherlands announce plans for new export controls on semiconductor equipment. Retrieved October 21, 2024, from Center for Strategic and International Studies website: https://www.csis.org/analysis/japan-and-netherlands-announce-plans-new-export-controls-semiconductor-equipment.

ASEAN Secretariat. (2024). ASEAN Guide on AI Governance and Ethics. Retrieved October 21, 2024, from ASEAN Secretariat website: https://asean.org/book/asean-guide-on-ai-governance-and-ethics/.

Benson, E., & Putnam, M. (2023, April 11). Export controls and intangible goods. Retrieved October 21, 2024, from Center for Strategic and International Studies website: https://www.csis.org/analysis/export-controls-and-intangible-goods.

Bode, I., Huelss, H., Nadibaidze, A., Qiao-Franco, G., & Watts, T. F. A. (2024). Algorithmic warfare: Taking stock of a research programme. *Global Society*, *38*(1), 1–23. https://doi.org/10.1080/13600826.2023.2263473.

Bode, I., & Qiao-Franco, G. (2024). The geopolitics of AI in warfare: Contested conceptions of human control. In R. Paul, E. Carmel, & J. Cobbe (Eds.), *Handbook on Public Policy and Artificial Intelligence* (pp. 281–294). Cheltenham: Edward Elgar Publishing. https://doi.org/10.4337/9781803922171.00030.

Brockmann, K. (2022). Applying export controls to AI: Current coverage and potential future controls. In T. Reinhold & N. Schörnig (Eds.), *Armament, Arms Control and Artificial Intelligence* (pp. 193–209). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11043-6_14.

Cespedes, F. V., & van der Kooij, J. (2023, April 18). The rebirth of software as a service. *Harvard Business Review*. Retrieved from https://hbr.org/2023/04/the-rebirth-of-software-as-a-service.

European Economic and Social Committee. (2021, March 28). Action Plan on synergies between civil, defence and space industries. Retrieved October 21, 2024, from European Economic and Social Committee website: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/action-plan-synergies-between-civil-defence-and-space-industries.

Government of the Netherlands. (2023, February 13). REAIM 2023. Retrieved October 21, 2024, from Government of the Netherlands website: https://www.government.nl/ministries/ministry-of-foreign-affairs/activiteiten/reaim.

Jin, Z. (2024, October 18). Tapping AI's potential. Retrieved October 21, 2024, from China Daily website: https://www.chinadailyhk.com/hk/article/595673#Tapping-AI%E2%80%99s-potential-2024-10-18.

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804.

Kissinger, H. A., & Allison, G. (2023, October 13). The path to AI arms control. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/united-states/henry-kissinger-path-artificial-intelligence-arms-control.

Klein, E., & Patrick, S. (2024, March 21). Envisioning a global regime complex to govern artificial intelligence. Retrieved October 21, 2024, from Carnegie Endowment for International Peace website: https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en.

Moller-Nielsen, T. (2024, January 25). EU reveals new economic security plan to resist "fierce" Chinese tech competition. Retrieved October 21, 2024, from Euractiv website: https://www.euractiv.com/section/economy-jobs/news/eu-reveals-new-economic-security-plan-to-resist-fierce-chinese-tech-competition/.

NATO. (2024, July 5). Defence Innovation Accelerator for the North Atlantic (DIANA). Retrieved from NATO website: https://www.nato.int/cps/en/natohq/topics_216199.htm.

Perifanis, N.-A., & Kitsios, F. (2023). Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review. *Information*, *14*(2), 85. https://doi.org/10.3390/info14020085.

Qiao-Franco, G. (2022, May 25). Can track II dialogues be the new "ping-pong" diplomacy to thaw the sino-us relationship on military ai? — Autonorms. Retrieved October 21, 2024, from AutoNorms website: https://www.autonorms.eu/can-track-ii-dialogues-be-the-new-ping-pong-diplomacy-to-thaw-the-sino-us-relationship-on-military-ai/.

Reiland, P. (2024, October 1). United nations: Global digital compact adopted by UN member states. Retrieved October 21, 2024, from Friedrich Naumann Foundation website: https://www.freiheit.org/human-rights-hub-geneva/global-digital-compact-adopted-un-member-states.

Reppy, J. (2006). Managing dual-use technology in an age of uncertainty. *The Forum*, *4*(1), 000010220215408884 1116. https://doi.org/10.2202/1540-8884.1116.

Shivakumar, S., Wessner, C., & Howell, T. (2024, February 21). Balancing the ledger: Export controls on u. S. Chip technology to China. Retrieved October 21, 2024, from Center for Strategic and International Studies website: https://www.csis.org/analysis/balancing-ledger-export-controls-us-chip-technology-china.

Sterling, T. (2023, June 30). Dutch curb chip equipment exports, drawing Chinese ire. *Reuters*. Retrieved from https://www.reuters.com/technology/amid-us-pressure-dutch-announce-new-chip-equipment-export-rules-2023-06-30/.

The White House. (2023a, April 27). Remarks by national security advisor Jake Sullivan on renewing American economic leadership at the Brookings institution. Retrieved October 21, 2024, from The White House website: https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/04/27/remarks-by-national-security-advisor-jake-sullivan-on-renewing-american-economic-leadership-at-the-brookings-institution/.

The White House. (2023b, August 9). Executive order on addressing united states investments in certain national security technologies and products in countries of concern. Retrieved October 21, 2024, from The White House website: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/08/09/executive-order-on-addressing-united-states-investments-in-certain-national-security-technologies-and-products-in-countries-of-concern/.

Tian, N., opes da Silva, D., Liang, X., Scarazzato, L., Béraud-Sudreau, L., & Assis, A. (2023). *Trends in world military expenditure, 2022*. Solna: SIPRI. Retrieved from SIPRI website: https://www.sipri.org/publications/2023/sipri-fact-sheets/trends-world-military-expenditure-2022.

UK Prime Minister's Office. (2023, November 2). The Bletchley declaration by countries attending the ai safety summit. Retrieved October 21, 2024, from UK Prime Minister's Office website: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

UNGA. (2024). *Lethal autonomous weapons systems: Report of the secretary-general*. UNGA. Retrieved from https://docs-library.unoda.org/General_Assembly_First_Committee_-Seventy-Ninth_session_(2024)/A-79-88-LAWS.pdf.

U.S. Department of Commerce. (2022, October 7). Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC). Retrieved from U.S. Department of Commerce website: https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file.

U.S. Department of State. (2023, November 9). Political declaration on responsible military use of artificial intelligence and autonomy. Retrieved October 21, 2024, from U.S. Department of State website: https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

Waldie, B. (2022, April 1). How military-civil fusion steps up china's semiconductor industry. Retrieved October 21, 2024, from DigiChina website: https://digichina.stanford.edu/work/how-military-civil-fusion-helps-chinas-semiconductor-industry-step-up/.

WEF. (2024, January 17). Davos 2024: Special Address by H.E. Li Qiang, Premier of the State Council of the People's Republic of China. Retrieved from World Economic Forum website: https://www.weforum.org/agenda/2024/01/li-qiang-china-special-address-davos-2024/.

PART 4

SOCIAL CHALLENGES OF AI

17 From AI Bias to AI By Us: A Case Study from MIT Critical Data

Catherine Bielick, Rodrigo Gameiro and Leo Celi

Abstract

This paper advocates for inclusive AI development, emphasizing its necessity for global equity, ethical soundness, and social relevance. We detail MIT Critical Data's approach to equitable AI development, focusing on healthcare. Our methods prioritize diverse collaboration and community engagement. Through global datathons, open-source datasets, and accessible education, we empower the global majority to actively participate in shaping AI that benefits all. Significant results, including numerous publications and established community hubs, demonstrate the impact of this approach. We argue that inclusivity in AI is not only achievable but crucial for its future success and fairness, particularly in serving the global majority.

Keywords: AI Bias, Open-source Datasets, Accessible Education, Global Datathlon, Fairness, Inclusivity.

Introduction

The Importance of AI from the Global Majority

While artificial intelligence (AI) influences society globally, its development and deployment are concentrated in technologically and economically dominant regions, leaving the majority of the global population underrepresented (World Health Organization, 2024). This disparity results in AI systems that do not reflect the diversity of the global majority. Consequently, these systems may perpetuate and exacerbate biases and inequities, further marginalizing already vulnerable populations (Shaffer, Alenichev, & Faure, 2023). The potential for AI to drive positive change is immense, but only if it is developed responsibly through a process that is participatory, inclusive, reflexive and reflective.

Inclusivity in AI development is not only a matter of equity, but is essential for any system that is ethically sound, socially relevant, and

economically beneficial to all (Hendl & Shukla, 2024; Jansky, Hendl, & Nocanda, 2024). By involving diverse voices in the AI creation process we, as a society, can ensure that these technologies are reflective of and responsive to the varied experiences and needs of different populations. At MIT Critical Data, we have taken these challenges head-on by employing a grassroots, local-first approach that prioritizes diversity and inclusivity in AI development. Through our initiatives, we aim to build a more equitable AI landscape that benefits everyone, not just a privileged few. This paper highlights our methods and the tangible outcomes of our work, demonstrating how inclusivity in AI is not only possible but essential for the technology's future success and fairness.

MIT Critical Data's Approach to Achieving Equitable Development, Transparency, and Accountability for AI

At MIT Critical Data, we recognize that engaging diverse communities is essential to combating bias in healthcare AI. Our approach is derived from five distinct core values: (1)rigorous and innovative research, (2)multi-level and accessible teaching, (3)building and networking communities of primary stakeholders, (4)reimagining legacy systems of power, and (5)advocacy for epistemic humility and health equity. We strive to unite the full range of professional, empirical, and cognitive backgrounds to foster collaborative imagination.

17.1 Discussion

17.1.1 Pioneering Research Methods in Healthcare AI

We conduct our research under the premise that AI has both the capacity to revolutionize healthcare, and to harm it. It is clear that relying solely on model prediction accuracy as the final arbiter for its implementation is short-sighted, not generalizable, and risks significant harm to populations traditionally excluded from research and model training (Futoma, Simons, Panch, Doshi-Velez, & Celi, 2020). Rather than merely developing highly accurate models using robust methodologies, we prioritize addressing foundational challenges in machine learning for healthcare and incorporating any model development into the broader context of the data. Recognizing the many biases inherent to healthcare AI across all

stages of the pipeline (Gichoya et al., 2023), we have made efforts to create guidelines for responsible AI development, such as a well-validated checklist called TRIPOD-LLM (Gallifant et al., 2024). This specific tool helps quantify the severity of bias in published studies using LLM models, and also serves as a framework for responsibly designing prospective healthcare LLM studies. Key considerations for responsible AI development include identifying and involving community members who would be most impacted by it, collaborating with co-authors from diverse backgrounds, openly discussing conflicts of interest, deeply understanding the data's story and fidelity, mitigating "hidden signals" in the data (Gichoya et al., 2023), and committing to the replicability of digital research through open science (Seastedt et al., 2022; Watson et al., 2023).

Nevertheless, it is important to underscore that AI research transcends any single cognitive or organizational domain and should not be developed, appraised, or regulated in a vacuum. Given its vast applicability, as we see, there are no individual experts in AI, only collective wisdom. For that reason, our research ranges widely, including large language models, AI model error interrogation, causal reinforcement learning, scientometric analysis, network science, epistemic research, time-series deep learning of electronic medical record data, ethics, vector embeddings, and implementation science. As such, to ensure a holistic approach, we collaborate globally with a diverse array of experts including social scientists, computer engineers, network scientists, ethicists, philosophers, physicians, veterinarians, pharmacists, data scientists, and statisticians. We believe that healthcare AI research should be cultivated within the global majority through a crowd-sourced approach that bridges communities and disciplines, and advances the decentralization and democratization -inclusion- of health equity research. This mission is furthered by teaching knowledge and skills, empowering others to pass this understanding forward.

17.1.2 Multi-level and Accessible Teaching

To nurture collaboration, one of our focuses is on teaching. Our approach includes a wide range of training, education, expertise, age groups, and demographics. We partner with local and distant

high schools and community colleges to advance healthcare AI education, sharing model development coding notebooks, providing access to open data sets, and offering tools to assess expected bias and harm. Our lab hosts a rotating cohort of visiting students from all over the world. We also teach at the Harvard School of Public Health, MIT, offer a freely available edX course, involve medical residents at Beth Israel Deaconess Medical Center, and engage in many more educational venues. We then translate this approach to durable, community-focused educational initiatives, particularly through global datathons, as discussed below.

17.1.3 Investment and Networking of Relevant Communities

As discussed, AI in healthcare cannot succeed without recentering the global majority to the forefront. To collaborate towards that goal, our approach centers on elevating primary actors involved in AI model development by initiatives such as incorporating their perspective into the TRIPOD-LLM bias assessment tool and validating a team scorecard applicable to any healthcare AI project. Also, we work to establish community hubs –organically scaled networks that bring together people from neighboring countries and regions. These hubs serve as grassroots initiatives, fostering a community of individuals committed to advancing equitable AI. By building those networks, we ensure to connect and empower the capacity that is mostly already present within the communities. Furthermore, at the local level, we nurture the next generation of AI leaders, equipping them with the critical perspectives needed to challenge prevailing biases in healthcare datasets. Our ethos is that critical thinking cannot thrive in a room where everyone thinks the same way. We believe that diversity in thought and experience is key to developing AI that is truly inclusive and effective.

One of our main drivers to establish such networks is through a global network of datathons and policy camps (Aboab et al., 2016). Our datathons are immersive, multi-day events held in countries across the globe. These events provide spaces where interdisciplinary teams can critically engage with open health datasets as well as collaborate to uncover and address biases that could influence AI

models, ensuring that these technologies prioritize health equity. Those are not only confined to capital cities; they are also hosted in smaller towns and regions that are often overlooked in global initiatives. This approach allows engagement of talented individuals from various backgrounds, ensuring that the AI solutions reflect the communities they are designed to serve. Furthermore, datathons and policy camps are often conducted in local languages, enabling participants to communicate and collaborate effectively, regardless of their linguistic backgrounds.

17.1.4 Reimagine Legacy Systems of Power and Expertise

When it comes to reimagining legacy systems, we stand for the decentralization of medical knowledge and the democratization of clinical data sharing. To achieve this we advocate for alternative metrics beyond the traditional impact factor to evaluate the impact of scientific journals, promote open access, and support open science to maximize scientific replicability. Our focus on data which is Findable Accessible Interoperable and Reusable (FAIR) reflects our dedication to transparency (Jacobsen et al., 2020). Moreover, aiming to diminish the barriers of data gate-keeping, we host PhysioNet, a continually-building collection of 314 large physiological and clinical datasets (at time of writing), over 50 related open-source software packages, and over 30 tutorials and reference guides. Among these datasets is the well-known Medical Information Mart for Intensive Care (MIMIC) now in its fourth iteration (MIMIC-IV), which includes data on 12,881 patients and 13,941 ICU stays from 2010-2018. Branches of this data set include raw CXR images, ECG waveforms, echocardiograms, emergency department encounters, and free-text clinical notes for large language models. All code is freely available and access is regulated through a data use agreement. As a result, preliminary data shows that MIMIC datasets are cited significantly more often than several proprietary publicly available datasets, with citation numbers ranging from 48.8-2,523.7 times higher, an advantage that grows further when adjusting for funding received.

17.1.5 Advocacy for Epistemic Humility and Digital Health Equity

Developing responsible AI in healthcare requires recognizing that this is a complex and multifaceted problem. Moreover, several common principles should be generally understood. First and foremost, when it comes to regulation, the authority to create policies around these systems must be primarily informed by those most affected. That is, when legitimized decision makers, such as regulatory agencies, are designing policies around AI, they should consult with the most affected stakeholders. For instance, if a healthcare AI model is to be trained and applied to people with HIV in South Africa, then people with HIV in South Africa must have a seat at the table for every stage of its development. Secondly, well-defined transparency standards throughout the AI model's lifecycle –from the conceptualization to implementation– must be developed. Third, rather than evaluate the AI bias of a model *post-hoc*, there may be value in mandating a prospective, systematic evaluation. However, it is important to emphasize that this is not a comprehensive list, and further initiatives are part of the iterative process towards building fairer outcomes for AI in healthcare.

As an example of future explorations towards responsible AI development, we are currently developing model interrogation tools to identify groups that might be harmed by false negative and positive predictions during the model validation stage. More classic approaches towards model performance evaluation are often insufficient, as shown by numerous studies that have identified "accurate" models in training and testing stages using conventional performance metrics, yet these models have ultimately caused harm or contributed to patient mortality when applied in real-world settings (McDermott, Hansen, Zhang, Angelotti, & Gallifant, 2024). While we have suggested some alternatives (Gallifant et al., 2023), these are still under development.

This is an iterative process, many times constrained by our collective imagination, and potential is lost when we surround ourselves with people who think exactly as we do. To counter this, we also created symposia for epistemic humility and critical thinking where individuals from any discipline can come together to discuss the broader ethical,

regulatory, and societal implications of AI in healthcare. Through those, we have learned that before regulation of AI health equity among the global majority can be more fully addressed, there are clear structural and systemic challenges to engage. We need to continue developing AI error interrogation tools and alternative performance metrics that capture the humanity inherent to the data. It is essential to incentivize peer-reviewed journals to reject manuscripts which only report accuracy of yet another new AI model. Educated community actors in AI must be involved in policy-making and. We must advocate for making science and data accessible from behind paywalls and ensuring it is understandable to those without the privilege of academic immersion. Collaboration in all forms, across disciplines, cognitive domains, cultures, religions, quantities of education, race/ethnicity, industries, and nations is essential to fully open the gates keeping AI from the global majority.

17.1.6 Results and Impact of MIT Critical Data's Approach

Our results have been significant, both in terms of academic output and real-world impact. Since 2014 we have hosted 46 datathons in 21 unique countries, including Singapore, Taiwan, the Philippines, Mexico, and more. Over 2,000 publications have been produced and a formal network effect assessment is also underway. These papers not only advance the field of healthcare AI but also ensure that contributions come from a broad spectrum of voices, particularly those from the underrepresented global majority. There are over 9,000 citations from over 40,000 people using the over 300 open-source datasets hosted on the PhysioNet Platform. These citations reflect the widespread adoption and influence of the datasets we maintain, which are used by researchers globally to develop AI solutions. Importantly, many of these citations come from researchers affiliated with low- and middle-income countries (LMICs) and minority-serving institutions (MSIs) in the United States, highlighting our success in promoting greater authorship representation from these regions.

Furthermore, the establishment of critical hubs has played a pivotal role in our initiative's success. By creating organically scaled networks that connect people across neighboring countries, we have fostered

a sustainable and resilient community of AI practitioners. These hubs are not reliant on external funding guarantees but are instead driven by the shared commitment of their members to advance equitable AI. For instance, AI. For instance, the collaboration between Mbarara University of Science and Technology (MUST) in Uganda and MIT exemplifies the transformative potential of these hubs. Rogers Mwavu, a computer scientist from Uganda and one of the key leaders in building this alliance, describes its impact:

"The MUST-MIT collaboration has significantly advanced a multidisciplinary approach to improving global health in Uganda, addressing key challenges such as maternal health, HIV/AIDS, and non-communicable diseases," Mwavu explains. This partnership has been particularly impactful in building local capacity and developing sustainable, culturally relevant solutions. By combining MIT's technological expertise with MUST's local insights, the collaboration has equipped healthcare workers, students, computer scientists, and community leaders with skills in data collection, analysis, and application. As a result of this long-term collaboration, researchers at MUST have implemented mobile-health tools for real-time patient data collection in remote areas, utilized telemedicine for expanded access to specialized care, and leveraged big data analytics to track health trends and predict disease outbreaks. This mutually beneficial partnership has not only enhanced healthcare delivery and research capabilities in Uganda but has also provided MIT students and faculty with valuable experience in applying technology to global health challenges, further demonstrating the reciprocal nature of our hub model.

Further concrete outcomes of our approach are reflected in the high-impact publications that have emerged from our initiatives (Collins et al., 2024; Ellen et al., 2024; Gottesman et al., 2019; Gottlieb, Ziegler, Morley, Rush, & Celi, 2022; Komorowski, Celi, Badawi, Gordon, & Faisal, 2018; Wong et al., 2021; Wu et al., 2022). These publications are not just a measure of academic success; they represent real-world advances in how AI can be used to improve healthcare for diverse populations. By involving diverse stakeholders in the co-creation process, we have developed AI solutions that are not only

technically robust but also aligned with the needs and realities of the communities they are designed to serve.

17.2 Conclusion: Lessons Learned and Future Directions

Through our work at MIT Critical Data, we have demonstrated that inclusivity in AI development is not only achievable but also essential for creating equitable healthcare solutions. Some challenges have included both ensuring sustained engagement from participants in underrepresented regions and bridging the gap between diverse linguistic and cultural contexts. We continually adapt our methods to ensure that our initiatives remain accessible and relevant. The need for sustained engagement underscores the importance of building long-term relationships with local communities, rather than relying on one-time events. Similarly, the diversity of linguistic and cultural contexts enriches the AI solutions developed through our initiatives, as they are informed by a broader range of perspectives and experiences.

As we look to the future, we urge the global AI community to recognize the value of engaging with diverse populations and to make a concerted effort to include voices from the global majority in AI decision-making processes. The future of AI in healthcare depends on our collective ability to build systems that are not only technologically advanced but also equitable and just. Together, we can create an AI landscape where every voice is heard, and every community benefits.

17.3 References

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. The Lancet. Digital Health, 2(9), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2.

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., … Purkayastha, S. (2023). AI pitfalls and what not to do: Mitigating bias in AI. The British Journal of Radiology, 96(1150), 20230023. https://doi.org/10.1259/bjr.20230023.

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., … Bitterman, D. S. (2024, July 25). The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use (p. 2024.07.24.24310930). p. 2024.07.24.24310930. medRxiv. https://doi.org/10.1101/2024.07.24.24310930.

Seastedt, K. P., Schwab, P., O'Brien, Z., Wakida, E., Herrera, K., Marcelo, P. G. F., ... Celi, L. A. (2022). Global healthcare fairness: We should be sharing more, not less, data. PLOS Digital Health, 1(10), e0000102. https://doi.org/10.1371/journal.pdig.0000102.

Watson, H., Gallifant, J., Lai, Y., Radunsky, A. P., Villanueva, C., Martinez, N., ... Celi, L. A. (2023). Delivering on NIH data sharing requirements: Avoiding Open Data in Appearance Only. BMJ Health & Care Informatics, 30(1), e100771. https://doi.org/10.1136/bmjhci-2023-100771.

18 The Prosumer in AI Governance: Class Antagonisms and the Social Relations of Labor

Avantika Tewari

Abstract

This paper examines "data prosumer" as an ideological construct essential to contemporary capitalism, framing users as virtual data producers who leverage personal data to assert privacy rights and engage in market activities. This abstraction helps commodify social interactions, reducing diverse human activities to exchangeable data units.

While personal data is governed by individual rights, non-personal data is appropriated by governments to create data markets that support visions of digital sovereignty in the global economy. The paper explores the reduction of labor to data prosumers in AI governance, emphasizing how digital labor markets exacerbate socio-political inequalities and informal labor conditions, especially across the Global Majority.

It critiques the global political economy's reification of individuals as "data populations" and reintroduces class analysis to challenge data commodification amid generative AI's mystification of labor. Finally, it argues that the push for digital sovereignty through data ownership obscures the exploitation inherent in capitalist, data-driven expansion.

Keywords: Prosumer, Digital Justice, Data Rights, Labor, Workerism, AI Governance, Data Sharing, Data Value.

Introduction

In the AI-driven era, the "prosumer" concept, initially developed by Alvin Toffler (1991) and George Ritzer (2019), describes individuals who both produce and consume, creating surplus value (Fuchs, 2012). I reinterpret this concept to emphasize the consumptive nature of digital social production under capitalism, where platforms abstract individuals into data values — whether as citizens, laborers, or consumers — interpellating them as a "data resource" to be reclaimed

through “data ownership.” This paper examines the transformation of global populations into data prosumers and explores its socio-economic implications for labor, using India’s policy framework and debates on digital sovereignty and AI governance as key examples.

18.1 Dual Nature of Prosumer Engagement

As concerns about AI displacing human labor grow, there is increasing advocacy for individuals to reclaim ownership of their generated data (Oliver and O’Neil, 2015). This perspective treats personal data as a compensatory asset (Birch, 2017; 2020), suggesting that ownership could help mitigate labor precarity.

The World Economic Forum even labels digital personal data as a “new asset class,” offering potential for economic and societal value creation (WEF, 2011: 5, quoted in Birch, 2020). However, this view often overlooks the exploitative dynamics within the capitalist data economy, where human activity is commodified.

Framing data ownership as a solution to labor precarity diverts attention from the systemic inequalities exacerbated by AI. Everyday activities generate data that trains AI models and sustains the attention economy (Ricardo et al., 2022). This shift transforms users into “prosumers,” whose labor, creativity, and knowledge become vital inputs for generative AI.

Data-driven societies rely on both traditional labor and emerging niche markets, where data producers meet algorithmically driven demands. This creates a paradox: while prosumer activities may seem unproductive, they are essential to the expansion of data markets.

As individuals interact with AI and the Internet of Things (IoT), the boundary between consumer and producer blurs, resulting in “data prosumption” — a form of labor critical to value creation in digital markets, yet often overlooked. Interoperable systems create digital “playgrounds” (Sukumar, 2021), seamlessly embedding data extraction into everyday life.

Platforms turn prosumer engagement into essential labor for the attention economy, while promoting the narrative that data is a form of property that the “new precariat class” (Standing, 2014)

can reclaim. However, this narrative obscures the deeper capitalist logic that drives data exploitation, preventing meaningful efforts to address growing inequalities in the digital economy.

18.2 Illusion of Data Ownership

The transition from productive labor to abstract data generation has led to two significant abstractions: reducing social interactions to quantifiable metrics and fetishizing data as a commodity. Although data ownership is often presented as a route to worker liberation, platforms increasingly exploit social relations and commodify labor, exacerbating class disparities. This exploitation is especially pronounced in the Global South, where AI development relies on vast quantities of data, often referred to as the "new oil."

The capitalist division of labor, historically measured by labor time, now manifests in prosumerism and AI economies. Capitalist strategies extend working hours to extract absolute surplus value and increase efficiency to extract relative surplus value (Marx, 1867). As a result, labor that does not produce immediate data outputs is marginalized, with "unpaid labor" being redefined through data value distribution strategies (Varoufakis, 2023).

The "prosumer ideology" obscures economic inequalities by promoting data dividends through the notion of "productive consumption" (Arvidsson, 2013) on platforms that promote "socially responsible capitalism." This ideology also encourages post-work entrepreneurialism (Webster & Dor, 2023), undermining traditional wage-labor contracts.

The valorization of digital consumption as "unpaid labor" (Fuchs, 2012) creates a paradox: passive data generation is seen as productive, overshadowing the material labor that sustains the digital economy. This disproportionately affects economies in the Global South, where critical but invisible work — such as data labeling, content moderation, and gig labor — supports AI systems but is devalued in market assessments in favor of "user-generated data" as a key commodity (Gao et al., 2021).

Despite their essential roles, workers from the Global Majority remain marginalized, while wealth generated by AI accumulates in

Global North platforms (Birhane, 2024). Claims of democratized access in the platform economy — where all individuals have equal opportunities as "data bodies" (Gurumurthy & Chami, 2021; Singh, 2019) — overlook how labor is restructured across industries, focusing too heavily on unequal data production value.

The narrative of egalitarian access conceals class conflicts, global labor divisions, and the precarity of workers sustaining networked publics. Celebrating digital access as "democratization" obscures the exploitation inherent in the platform economy, making systemic inequalities invisible. The idea of a digital "playground" (Scholz, 2013) for self-expression masks value extraction mechanisms, reinforcing myths of equal opportunity and ignoring persistent structural barriers.

18.3 Dialectics of Labor and Value in the Digital Economy

Digital platforms are designed to capture user attention and engagement, converting historical data into user profiles and economic value. This dynamic creates tensions in AI-driven economies, where consumption often overshadows the productive labor required to maintain these systems. As data use expands, critical issues arise around ownership, rights, and labor, particularly concerning the protection of personal data and governance of anonymized and non-personal data (Gupta & Naithani, 2023).

Government interventions in data sharing, ostensibly promoting innovation by breaking data silos, often entrench exploitative labor practices. Such interventions render specific work invisible while giving data businesses access to centralized public databases. For instance, the Indian government portrays itself as both a guardian of public interest and a market architect, reshaping data to serve its "digital sovereignty" aspirations (Athique, 2019: 77). This dual role supports a broader shift towards a digital economy, especially in the Global South.

Despite the reliance on user-generated data for platforms, the labor underpinning these ecosystems — such as gig work, logistics, and data services (Dzieza, 2023) — remains undervalued. Workers are obscured within the value chain as platforms prioritize data

commodification, reducing individuals to "data bodies" capable of asserting "data sovereignty" only through consent-based exchange.

India's "health stack," for example, aims to unify healthcare services by treating anonymized health data as a public good (Barlett et al., 2024; Gurumurthy & Chami, 2022, Parsheera 2024). However, the state's custodianship of data under the guise of "national public interest" paradoxically promotes data business growth while exempting certain processors from regulatory oversight to create "national champions" (Athique & Kumar, 2022; Panday, 2021).

Efforts to reclassify data based on its purpose, origin, and domain of production lack comprehensive legal clarity. Non-personal data (NPD) (Singh 2019), which cannot be directly linked to an individual's identity, often comes from "unseen workers" like content moderators, data labelers and gig workers. These workers are essential for generating and maintaining vast amounts of NPD, particularly in the Global South, where digital labor pools support multinational tech companies (Shahid, 2024; Mehrotra, 2022).

An example of how NPD is utilized in platform capitalism is real-time traffic data, often collected from gig workers such as ride-hailing drivers or food delivery couriers. This NPD enhances operational efficiencies for platforms by optimizing routes, predicting demand, and reducing delivery times.

Unlike personal data, which is often framed as empowering individuals with privacy or control, NPD is treated as a public resource that companies and governments can expropriate. This reveals the limitations of prosumer ideology, which suggests shared agency, but in reality, NPD (Verma & Gurtoo, 2021) is harvested without workers' knowledge or compensation, challenging the notion of user power over their data (Fink, 2024).

Government interventions that claim to promote data "commons" and innovation often exacerbate labor exploitation by rendering work anonymous while providing vast datasets to businesses. This approach narrows power and proprietary rights over personal data while commodifying it to serve market imperatives in a bid to claim "national champions on the global stage" as a way to assert digital sovereignty (Athique & Kumar, 2022, Panday 2021).

For example, there is growing criticism that the Indian government aligns its data governance with market goals, reshaping public databases for both public and private sector use, prioritizing economic utility over the social value of labor. This reduces individuals to mere data owners or human capital in AI systems (Mishra, 2023; Panday & Samdub, 2024). Advocating for data rights through data as a public good, in this context ignores the complexities of labor exploitation, especially for the Global Majority (Barlett et al., 2024; Gurumurthy & Chami, 2022).

By portraying "digital subjects" as entrepreneurial agents (Irani, 2019), these narratives obscure systemic exploitation, which reduces labor to generating surplus value for AI-driven optimizations. The abstraction of labor into data commodities erases critical distinctions of class, gender, race, and geography (Mohun, 1984). The prosumer ideology deepens these inequalities by framing individuals as "data bodies" (Mager & Mayer, 2019), further entrenching divisions along lines of caste, race, and gender in the Global Majority.

18.4 Prosumer Ideology as a Condition of Data Market Expansion

Intersectional feminist critiques (Gurumurthy & Chami, 2021, Radhakrishnan 2020) highlight the importance of incorporating social power differentials into data science and ethics, advocating for embodied subjectivity and democratic participation in production. This critique challenges the disembodied abstraction of labor in platform capitalism, stressing the need for equitable representation.

As generative AI becomes integral to platform business models, debates around creative labor and intellectual property reemerge, necessitating a deeper understanding of colonial legacies and neo-colonial accumulation patterns that reinforce global inequalities. Beyond addressing data denial, it is essential to analyze the class structures that perpetuate divisions within the "digital precariat" (Standing, 2014). In the Global South, the rise of "peer-to-peer" services has further platformized domestic spaces and informal labor, deepening informality through algorithmic job allocation (Dubal 2023).

Platforms often render human mediation invisible, reducing workers to mere algorithmic components. Gray and Suri (2019) argue that the "ghost workers" behind AI should be acknowledged as crucial actors in networked publics. However, while this recognition seeks to expose invisible labor, it fails to challenge capitalist structures that systematically devalue specific forms of labor across the Global Majority.

Recognizing all labor as equal does not dismantle the structural inequalities determining labor value. A rights-based approach may affirm the dignity of work, but it overlooks the dualities of exploitation and domination (Ayalew, 2024). Such approaches risk reinforcing techno-solutionism (Duberry, 2023), particularly when governance frameworks render informal workers "computable" under the guise of digital inclusion, obscuring the underlying power dynamics.

Framing the digital precariat as a unified class of data prosumers oversimplifies diverse lived experiences, masking the unequal access to resources and opportunities within the platform economy. Defining labor through precarity falsely implies equality among those engaged in "free labor" on digital platforms, ignoring the structural differences shaping their roles.

The "sharing economy" facilitates data value exchange among "peers" through AI mediations, reducing labor to abstract data value authenticated by scientific economism (Sinha 2024). Although these metrics acknowledge diverse identities, they reinforce normative categories that exclude workers who fail to meet specific algorithmic standards. Rather than promoting egalitarian market access, this dynamic entrenches class-based marginalization. Addressing platform exploitation requires not just analyzing the power of algorithms but also recognizing the ongoing expropriation of labor through them.

Furthermore, portraying citizens as part of a "data-rich" digital precariat (Nilekani, 2018) echoes Althusser's concept of the subject as an ideological construct, which obscures the material foundations of capitalist exploitation. Focusing solely on digital sovereignty and data-prosumer rights — through frameworks like consensual data-sharing (WEF, 2022; Singh & Vipra 2019) — neglects the structural exploitation embedded in platform economies. Platforms claiming

to “formalize” informal sectors (Surie & Huws, 2023) often mask class exploitation, expanding data markets while sidelining labor outside of the platforms.

Platforms like Uber and Swiggy exemplify a significant shift in the organization of labor, aggregating informal workers and commodifying every aspect of platform development. Logistics, a central component of “variegated capitalism” (Neilson & Rossiter, 2017), demonstrates how local labor regimes and consumption patterns are shaped for global exchange, reducing workers to “data bodies” and framing them as social “peers” within an abstract digital economy.

This dynamic relegates workers to fragmented roles within a rapidly expanding consumer economy, redirecting the discourse from issues of privacy and control to those of data valuation and exploitation (Singh, 2019). As a result, these systems not only deepen existing inequalities but also conceal the exploitative nature of digital labor. The ideology of prosumerism, which claims to elevate users and digital laborers as data producers (Arvidsson, 2013), simultaneously devalues the logistical labor predominantly carried out by workers in the Global South (Shanmugavelan, 2024). Framing data extraction as progressive economic development obscures the material labor that sustains it, exacerbating global inequalities (Jung, 2023).

18.5 Conclusion

Recognizing individuals as data entities has intensified claims to data rights, increasingly tying them to legal frameworks governing data exchange. This shift represents a departure from traditional notions of privacy, which emphasized withholding data from platforms, toward advocacy for individual control over data usage. However, genuine control over data necessitates collective agency in regulating digital production — a dimension that remains largely unaddressed.

Reducing users to digital prosumers or data subjects (Gandini 2021) oversimplifies deeper societal conflicts and reinforces their status as biopolitical populations (Gregory & Sadowski, 2021), treating individuals as interchangeable data points. This perspective flattens complex class relations into negotiations over data production,

framing social inequality as a matter of uneven data distribution rather than engaging with the broader socio-economic divides that underpin it.

Moreover, digital users are often conceptualized as "prosumer commodities" (Flisfeder, 2016), with their autonomy shaped by platforms in data-driven markets. This dynamic reconfigures the labor-capital relationship, heightening worker precarity and complicating the classification of user activity as productive labor. The commodification of user engagement exacerbates exploitation, further blurring the line between consumer participation and labor.

As generative AI enhances content creation, disputes over data ownership are likely to intensify, exposing contradictions in commodifying user data while maintaining the illusion of personal control. It is critical to assess AI's material impacts on labor processes rather than merely speculate on its abstract potential. While scholars such as Christian Fuchs (2013) emphasize unpaid digital labor, these frameworks risk oversimplifying class struggles, particularly in debates on AI governance.

Julie Cohen (2019) critiques how governance structures are co-opted by economic imperatives, reshaping democratic processes in favor of market interests. The "economization of governance" transforms democratic participation into a productivist role for citizens, reducing their political agency in favor of their role as economic actors — data prosumers fueling AI systems. This reduction of social interactions to data-driven abstractions turns individuals into biopolitical populations where rights and agency are claimed through data identification (Athique & Parthasarathi, 2023).

Ultimately, the rhetoric of data decolonization and digital sovereignty may, paradoxically, reinforce existing data markets by normalizing the reduction of citizens to data subjects exploited for their data potential. This normalization distracts from addressing systemic inequalities and risks entrenching structures of exploitation under the guise of empowerment.

To reclaim genuine agency, we must reject reductive distributive and productivist frameworks and advocate for an understanding

of labor, one that recognizes the complex dynamics of power, inequality, and exploitation in the digital age.

18.6 References

Marx, K. (1867). Part V: The production of absolute and of relative surplus-value. In *Capital Volume One*. Marxist Online Archive.

Toffler, A. (1991). *The third wave*. New York, NY: Bantam Books.

Mohun, S. (1984). Abstract labor and its value-form. *Science & Society, 48*(4), 388–406. http://www.jstor.org/stable/40402953.

Fuchs, C. (2012). Google capitalism. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society, 10*(1), 42–48. https://doi.org/10.31269/triplec.v10i1.304.

Arvidsson, A. (2013). The potential of consumer publics. *Ephemera, 13*(2), 367–391.

Fuchs, C. (2013). Digital prosumption labor on social media in the context of the capitalist regime of time. *Time & Society, 23*(1).

Scholz, T. (Ed.). (2013). *Digital labor: The internet as playground and factory*. Routledge.

Frayssé, O., & O'Neil, M. (2015). *Digital labour and prosumer capitalism: The US matrix*. Palgrave Macmillan.

Flisfeder, M. (2016). Digital labour and the internet prosumer commodity: In conversation with Christian Fuchs. *Alternate Routes: A Journal of Critical Social Research, 27*. Retrieved from https://alternateroutes.ca/index.php/ar/article/view/22403.

Birch, K. (2017). Rethinking value in the bio-economy: Finance, assetization, and the management of value. *Science, Technology, & Human Values, 42*(3), 460–490. https://doi.org/10.1177/0162243916661633.

Bailey, R., & Parsheera, S. (2018, October 31). Data localisation in India: Questioning the means and ends. National Institute of Public Finance and Policy. Retrieved October 14, 2024, from https://www.nipfp.org.in/media/medialibrary/2018/10/WP_2018_242.pdf.

Nilekani, N. (2018, August 14). Data to the people: India's inclusive internet. *Foreign Affairs*.

Athique, A. (2019). Digital emporiums: Platform capitalism in India. *Media Industries Journal, 6*(2). https://doi.org/10.3998/mij.15031809.0006.205.

Cohen, J. E. (2019). *Between truth and power: The legal constructions of informational capitalism*. Oxford University Press.

Ghosh, S., et al. (2019, January 22). Google says data is more like sunlight than oil, just 1 day after being fined $57 million over its privacy and consent practices. *Business Insider India*.

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Irani, L. (2019). *Chasing innovation: Making entrepreneurial citizens in modern India*. Princeton University Press. https://doi.org/10.2307/j.ctv941vd8.

Singh, P. J., & Vipra, J. (2019). Economic rights over data: A framework for community data ownership. *Development*, 62(1), 53–57. https://doi.org/10.1057/s41301-019-00207-3.

Ritzer, G. (2019). Prosumption: Contemporary capitalism and the 'new' prosumer. In F. F. Wherry & I. Woodward (Eds.), *The Oxford handbook of consumption* (pp. 73–93). Oxford University Press.

Singh, P. J. (2019). Data and digital intelligence commons (making a case for their community ownership). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3873169.

Mager, A., & Mayer, K. (2019). Body data — data body: Tracing ambiguous trajectories of data bodies between empowerment and social control in the context of health. *Momentum Quarterly, 8*(21), 95–108. https://www.researchgate.net/publication/334178972_Body_data_body_Tracing_ambiguous_trajectories _of_data_bodies_between_empowerment_and_social_control_in_the_context_of_health.

Athique, A., Parthasarathi, V., & IAMCR — International Association for Media and Communication Research (Eds.). (2020). *Platform capitalism in India*. Palgrave Macmillan.

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPT-Ed, 17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389.

Birch, K., & Muniesa, F. (Eds.). (2020). *Assetization: Turning things into assets in technoscientific capitalism*. The MIT Press.

Sukumar, A. M. (2021, February 5). Designing digital public goods and playgrounds in India: The need for theoretical and contextual analysis. ISPIRT. https://research.ispirt.in/articles/Designing-Digital-Public-Goods.

Gandini, A. (2021). Digital labour: An empty signifier? *Media, Culture & Society, 43*(2), 369–380. https://doi.org/10.1177/0163443720948018.

Gao, S., Liu, Y., Kang, Y., & Zhang, F. (2021). User-generated content: A promising data source for urban informatics. In W. Shi, M. F. Goodchild, M. Batty, M.-P. Kwan, & A. Zhang (Eds.), *Urban informatics, the urban book series* (pp. 503–522). Springer.

Gregory, K., & Sadowski, J. (2021). Biopolitical platforms: The perverse virtues of digital labour. *Journal of Cultural Economy, 14*(6), 662–674. https://doi.org/10.1080/17530350.2021.1901766.

Athique, A., & Kumar, A. (2022). Platform ecosystems, market hierarchies and the megacorp: The case of Reliance Jio. *Media, Culture & Society, 44*(8), 1420–1436. https://doi.org/10.1177/01634437221127798.

Gurumurthy, A., & Chami, N. (2022). Beyond data bodies: New directions for a feminist theory of data sovereignty. *Data Governance Network*.

Parsheera, S. (2022). India's policy responses to big tech: And an eye on the rise of 'alt big tech'. *Indian Journal of Law and Technology, 18*(1).

Sadowski, J. (2022). Lords of the platform. In *Platform labour and global logistics* (pp. 28–38). Routledge.

World Economic Forum. (2022). Sharing data to achieve decarbonization of value chains: Briefing paper. https://www3.weforum.org/docs/WEF_Data_sharing_to_decarbonize_value_chains_2022.pdf.

Alasoini, T., Immonen, J., Seppänen, L., & Känsälä, M. (2023). Platform workers and digital agency: Making out on three types of labor platforms. *Frontiers in Sociology, 8*, 1063613. https://doi.org/10.3389/fsoc.2023.1063613.

Baeza-Yates, R., & Fayyad, U. M. (2022). The attention economy and the impact of artificial intelligence. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Eds.), *Perspectives on digital humanism* (pp. 123–134). Springer International Publishing.

Dubal, V. (2023). On algorithmic wage discrimination. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4331080.

Duberry, J. (2023, January 30). Beyond techno-solutionism and silver bullets. *Geneva Policy Outlook*. Retrieved October 10, 2024, from https://www.genevapolicyoutlook.ch/beyond-techno-solutionism-and-silver-bullets/.

Dzieza, J. (2023, June 20). Inside the AI factory: The humans that make tech seem human. *New York Magazine*. Retrieved October 14, 2024, from https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html.

Gil, J., Martínez, P., & Sequera, J. (2023). The neoliberal tenant dystopia: Digital polyplatform rentierism, the hybridization of platform-based rental markets and financialization of housing. *Cities, 137*, 104245. https://doi.org/10.1016/j.cities.2023.104245.

Gupta, I., & Naithani, P. (2023). Data protection, localization and sovereignty: How is India's privacy landscape evolving? In *Indian Perspectives on Information Governance and Privacy Law* (pp. 297–312). Springer. https://doi.org/10.1007/978-981-19-9615-4.

Jung, M. (2023). Digital capitalism is a mine not a cloud: Exploring the extractivism at the root of the data economy. *TNI*.

Mishra, N. (2023). Data governance and digital trade in India: Losing sight of the forest for the trees? In A. Chander & H. Sun (Eds.), *Data sovereignty* (pp. 240–263). Oxford University Press.

Surie, A., & Huws, U. (2023). *Platformization and informality: Pathways of change, alteration, and transformation*. Palgrave Macmillan.

Varoufakis, Y. (2023). *Technofeudalism: What killed capitalism*. The Bodley Head.

Webster, E., & Dor, L. (2023). The end of labour? Rethinking the labour question in the digital age. In *Recasting workers' power* (pp. 1–29). Policy Press.

Ayalew, Y. E. (2024, June 23). A Third-World critique of the human rights-based approach to content moderation. *Tech Policy Press*. Retrieved October 10, 2024, from https://www.techpolicy.press/a-thirdworld-critique-of-the-human-rightsbased-approach-to-content-moderation/.

Bartlett, B., Ainsworth, J., Cunningham, J., Davidge, G., Harding, M., Holm, S., Neumann, V., & Devaney, S. (2024). Health data stewardship: Achieving trust through accountability in health data sharing for research. *Law, Innovation and Technology*, 1–41. https://doi.org/10.1080/17579961.2024.2392937.

de Souza, S., & Bhardwaj, K. (2024). Publisher correction to: India's conception of community data and addressing concerns for access to justice. *DISO*, 3, 17. https://doi.org/10.1007/s44206-024-00104-3.

Fink, A. (2024). Data cooperative. *Internet Policy Review*, 13(2). https://doi.org/10.14763/2024.2.1752.

Parsheera, S. (2024). Stack is the new black?: Evolution and outcomes of the 'India-stackification' process. *Computer Law & Security Review*, 52, 105947. https://doi.org/10.1016/j.clsr.2024.105947.

Panday, J., & Samdub, M. (2024, March 12). Promises and pitfalls of India's AI industrial policy. *AI Now Institute*. Retrieved August 13, 2024, from https://ainowinstitute.org/.

Shahid, F. (2024, August 26). Colonialism in content moderation research: The struggles of scholars in the majority world. *Center for Democracy and Technology*. Retrieved October 10, 2024, from https://cdt.org/insights/colonialism-in-content-moderation-research-the-struggles-of-scholars-in-the-majority-world/.

Shanmugavelan, M., et al. (2024). The formalization of social precarities. *Data & Society*. https://datasociety.net/library/the-formalization-of-social-precarities/.

Sinha, A. (2024). A general model and incomplete history of AI. *Knowing without Seeing*. Retrieved from https://www.knowingwithoutseeing.com/essays/a-general-model-incomplete-history-of-ai.

19 Cost or Benefit? The impact of AI on the work of medical practitioners

Amrita Sengupta and Shweta Mohandas

Abstract

While there is a growing interest in using AI for its speed and proposed efficiency, there are concerns over its use in the highly specialised and sensitive medical field. Through primary research with medical professionals, this essay looks at the current use of AI by medical practitioners in their research and practice, new challenges and the perceived benefits of AI for healthcare for medical professionals.This essay also briefly reviews generative AI's impact on community health workers in India. The essay suggests a more careful approach to AI adoption for healthcare so as to not cause undue burden on healthcare professionals in the short to medium term.

Keywords: Artificial Intelligence, AI in Healthcare, GenAI, AI and work.

19.1 Background

The growth of applications of AI in healthcare has proliferated globally, some of the popular use cases being radiology, telemedicine and mental health chat bots, while use of AI in drug discovery and disease surveillance have also seen an increased interest. Global studies have also suggested that AI can help in reducing treatment costs, improving health outcomes and, helping in faster diagnosis (IBM, 2024), (Alowais et al., 2023).

In the Indian context one estimate suggests that "the Indian healthcare AI market is expected to reach USD 1.6 billion by 2025"("AI In Healthcare: Changing India's Medical Landscape," n.d.). Startups like Cure.AI, Niramai and Wysa, BrainSightAI as well as big technology companies such as IBM, Microsoft and Google have already invested heavily in AI and healthcare in India (Pti, 2024). Given the rapid scale of growth and investments in AI systems, we are at a moment where adoption for AI in healthcare in India needs to be critically

examined, specifically on how it impacts the work of medical practitioners and the healthcare system at large. The demand for healthcare professionals is expected to grow given the current shortage of healthcare workers in India (with a ratio of 1.7 nurses per 1,000 people and a doctor-to-patient ratio of 1:1,500 nationwide) ("Healthcare System in India, Healthcare India — IBEF," n.d.). In a currently overburdened healthcare system, the promise of AI is that of faster, efficient and cost effective diagnosis and care. However, as a build up to it, what are the demands it will put on healthcare workers in the immediate term, with additional data annotation and labelling responsibilities, learning the use of advanced and emerging technologies, and picking up additional data management responsibilities, among others? In India, especially since the process of digitising healthcare is still nascent, there is a need to look at if and whether AI is actually living up to its promises and acting as an aid to medical professionals if not a replacement.

In addition, with the growth and large-scale adoption of Generative AI (GenAI), there has been an increased pattern of information seeking on platforms such as ChatGPT. While there are certain benefits to be derived from such use, it also raises questions on how physician's over-reliance on (GenAI) responses in clinical decision making could impact patients, medical practitioners as well as the healthcare system at large, some deliberations we hope to get into through the course of this essay.

In this essay, we present findings from our research on how medical professionals currently use AI for healthcare, the perceived benefits and pitfalls of using AI, specifically how it impacts the work of medical professionals, and a few provocations for future implementation of AI systems in India in the wake of (GenAI).

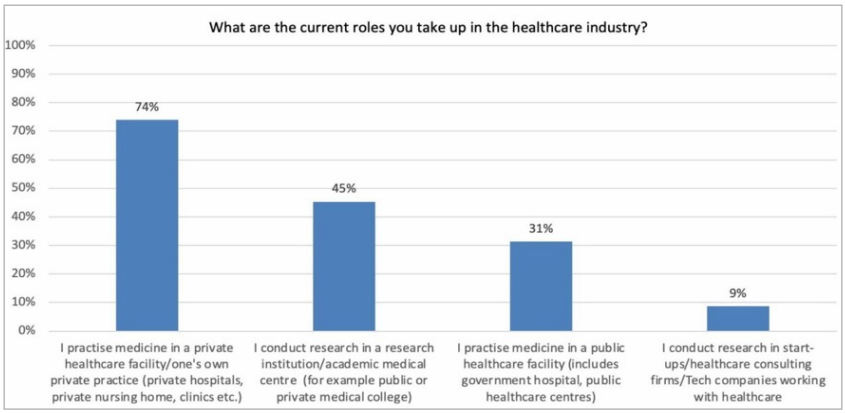
19.2 Methodology

As part of a larger mixed methods, Institutional Review Board approved study on AI and healthcare in India, we conducted three surveys with 500 respondents across three prominent stakeholder groups — medical practitioners and researchers (150 respondents), respondents from healthcare institutions (150 respondents), and respondents from technology companies and startups developing

and deploying healthcare-focused AI models in India (200 respondents). We also did 18 qualitative interviews with medical professionals, startups, technology companies, civil society members, and policy makers.

In this essay, we focus specifically on the medical practitioner and researchers' survey with 150 respondents and the interviews with five doctors, and ten technology companies and startups from the larger study. Data collection for the surveys and interviews were conducted between January and April 2024. The below chart lays out the split of the medical professionals surveyed by their roles.

Figure 1 Responses from medical professionals on their use or research in AI in particular areas.

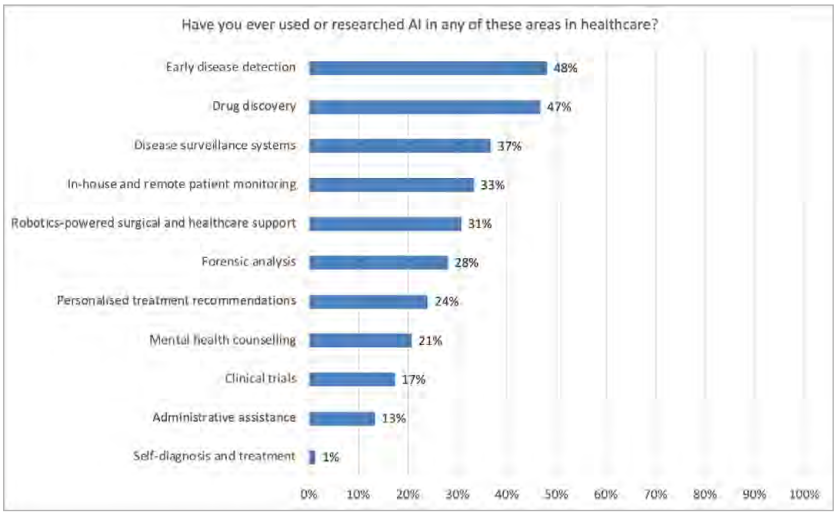


Source: CIS Survey of medical practitioners in AI and healthcare, January-April 2024, n = 150.

19.3 Current use of AI by medical practitioners in their research and practice — Insights from our study

As seen in Figure 1, our survey revealed that a lot of AI related work was limited to research in particular areas as opposed to actual implementation. Early disease detection and drug discovery were the most picked areas of use and research in healthcare. Administrative assistance was one of the lowest-picked choices, which potentially points to the lack of access to standardised processes and digitised data that could be used for training AI for administrative assistance.

Figure 2 Responses from medical professionals on their use or research in AI in particular areas. This was a multi-select question



Source: CIS Survey of medical practitioners in AI and healthcare, January-April 2024, n = 150.

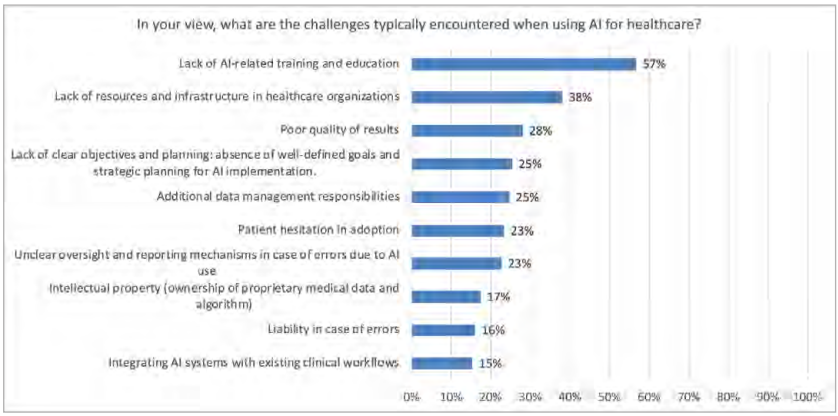
Through our interviews and interviewee profiles we gathered that the most common use cases of AI in India include diagnosis such as cancer screening, chatbots for mental health, drug discovery, and remote monitoring, and for administrative assistant and patient management functions. However all the doctors we interviewed stated that the use of AI in their workflow was not yet widespread. The use of AI was also limited to mostly private hospitals. In our interview with one doctor working in a large public hospital they stated that AI was being currently implemented in research stages and the AI use was currently limited to administrative tasks.

19.3.1 New challenges brought by AI for medical professionals

In our survey, medical practitioners expressed several concerns that they experienced while using AI for healthcare. Nearly 60% medical practitioners expressed the lack of AI-related training and education as a big barrier to adoption of AI systems. While 25% respondents also reported additional data management responsibilities as a challenge, which points to the burden that AI use is creating for medical practitioners.

Nearly one in four medical practitioners in our survey mentioned additional data management responsibilities as a challenge when it came to integrating AI into their work. Doctors have also raised concerns of the efforts and infrastructure required on their side to digitise health records, such as administrative assistance ("Ayushman Bharat Digital Mission: Boon or Bane?," 2023) and the cost of data security(Karpagam, 2021).This was pointed out in our survey as well, with 38% medical practitioners citing the lack of resources and infrastructure as a challenge while using AI for healthcare (see Figure 3).

Figure 3 Responses from medical practitioners on the question of challenges faced while using AI for healthcare. This was a multi-select question



Source: CIS Survey of medical practitioners in AI and healthcare, January-April 2024, n = 150.

While there have been reports of state governments encouraging the use of AI in healthcare with initiatives such as the screening for kidney disease ("State Government to Screen Kidney Diseases With AI-powered Mobile App," n.d.) and tuberculosis(Yasmeen, 2024); the use of AI is mostly limited to private hospitals ("Progress of Healthcare Artificial Intelligence in India," n.d.).Hence the benefits of AI like reduced costs, efficiency, and reduced burden on doctors is yet to reach the areas where it is needed the most — -the public healthcare system.

Through the AI life cycle for healthcare, medical practitioners would be required to intervene at various stages of AI implementation from data collection to train the AI systems to deployment and use of

AI systems by the medical practitioners.The following paragraphs shed light on how medical practitioners engage in these stages, with insights from our in-depth interviews.

In data collection; the lack of India specific data requires medical professionals to digitise and annotate the data in addition to their clinical and administrative work. The issue of data security especially is also more emphasised after multiple health data leaks (Singh, 2024). It was highlighted by interviewees from civil society that the medical professionals in addition to data collection had to spend out of pocket to ensure security of this data.

In development; where the AI system is made and trained by technologists, and medical practitioners are seldom involved in its creation. It was pointed out by civil society interviewees that often medical professionals are not actively involved in the development of the AI systems, thereby making them mere end users. In our interviews with doctors, however, they stated that they worked with startups in developing, and providing feedback to AI systems (Pti, 2023).

In deployment; the still high cost of AI and existing infrastructural challenges with healthcare in India, means that the doctors and hospitals are still not able to adopt AI systems as easily (Alkhaldi, 2024). It was highlighted by an interviewee from a tech company that hospitals are still grappling with the idea of accommodating AI systems in their existing workflows and still deciphering how to book AI to their expenses (whether as devices or an IT expense). They also stated that the high cost of AI had to be justified in order for hospitals to purchase them, and then proceed to make up the costs from the patients.

In terms of AI workflows, post deployment as well, these systems are currently being used as a tool that compliments the doctor's decisions. However this in turn adds another layer of work for the medical professional who cannot blindly follow the AI system's results.

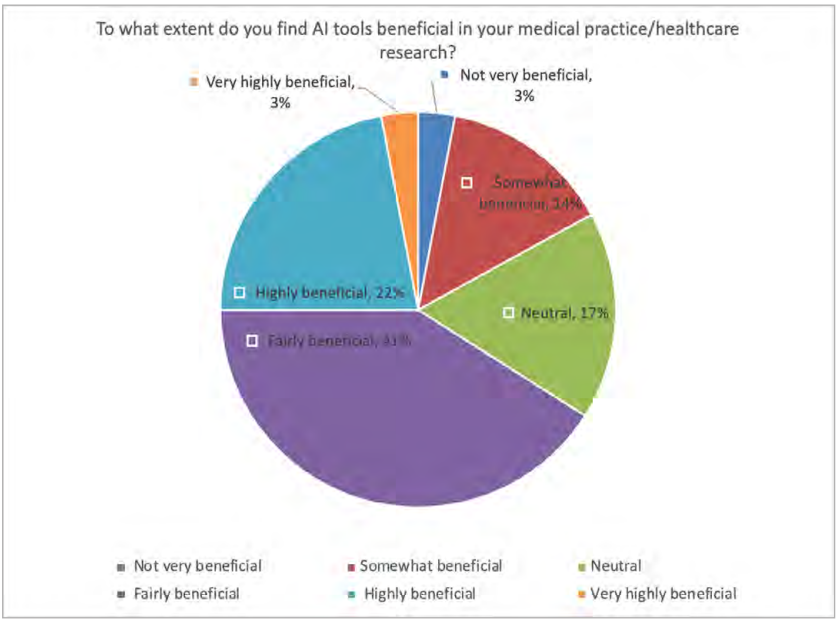
19.3.2 Perceived benefits of AI in healthcare

There are some benefits that AI in healthcare can bring. As secondary literature suggests, one such example is its use in disease surveillance, which due to the large amount of data and compute power that AI uses, offers a significant advantage. AI is being used to predict future

outbreaks as well as help public health officials be better prepared and proactive (Anjaria et al., 2023). An example of this initiative in India is the Dengue Dashboard established at IISC (Artpark, n.d.). Similarly in drug discovery, AI's potential to transform every stage of the workflow is being explored. Currently AI is involved in drug design, decision making; determining the right therapy for a patient, and managing the clinical data generated(Chun, 2023). In India AI was used to examine potential drugs for Covid -19 treatment (ET HealthWorld, 2020). The use of language data and text to speech has also been helpful in providing multilingual support through chat bots such as Wysa ("FAQ — AI Chatbot | Online Therapy," n.d.), which provides mental health support in Hindi and a few other languages, and HealthifyMe provides multilingual support to maintain nutritional goals (Saha, 2024).

In our survey, medical practitioners also shared their views on whether they find AI beneficial and in which areas (see Figure 4 and Figure 5).

Figure 4 Responses on how medical professionals view AI as a tool in their medical practice/healthcare research (question was asked on a likert scale of 1 to 7, single-select)

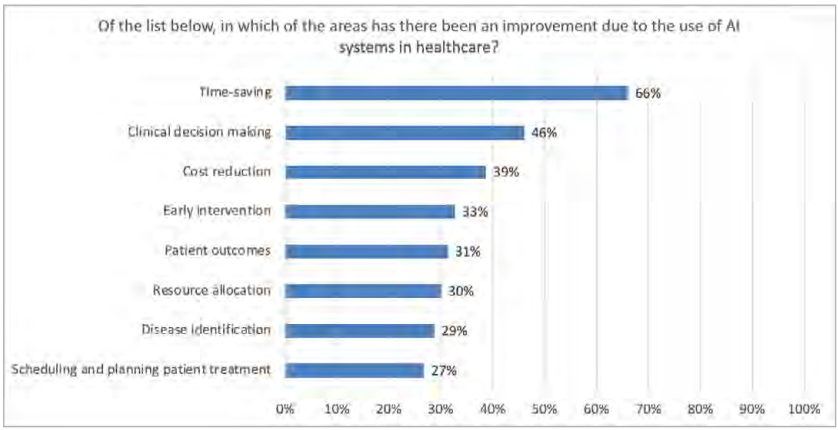


Source: CIS Survey of medical practitioners in AI and healthcare, January-April 2024, n = 150.

In our survey, 41% of medical professionals suggested that AI could be fairly beneficial in healthcare or healthcare research (see Figure 4). Further, when it came to realising the benefits of AI, medical professionals saw time saving as the most noted benefit, followed by improvement in clinical decision making (see Figure 5).

This could have certain potential risks that have been highlighted in various research studies from the last many years, including but not limited to questions of liability in case of AI based decision making, over-reliance on AI leading to loss of/negative impacts on clinical skills, representational biases that the AI models may present (Ameen, Wong, Yee, & Turner, 2022), especially so if the source of data is not the same as where the AI system is being administered.

Figure 5 Responses from medical professionals on areas where they have seen improvement due to AI. This was a multi-select question



Source: CIS Survey of medical practitioners in AI and healthcare, January-April 2024, n = 150.

19.3.3 Generative AI and its impact on community health workers in India

While our survey and interviews didn't directly or extensively investigate the use of Gen AI for healthcare, given its rapid adoption, it would be remiss to not reflect on its impact on healthcare in India. In this section, through existing secondary literature and our analysis, we look at ways in which it can have positive and some negative impacts when used for healthcare in the Indian context, especially in public health.

Accredited Social Health Activist (ASHA) workers who are community health workers are the first point of care between the family and public health system, through collecting data, providing basic curative care, and promoting universal immunisation("About Accredited Social Health Activist (ASHA)," n.d.). Due to their reach they could be provided with smartphones compatible with dedicated AI screening applications, something which has seen some success in infant care (Ai, 2023) and GenAI could be explored to help them make quick initial diagnosis and screening about a person's health, and prioritise care. Gen AI could also be used for translations to reduce the language barrier.

As stated earlier AI has seen success in disease surveillance and prediction. With the amount of data, speed and the right training of community health workers and public health professionals in Gen AI applications, it could be used to collate large amounts of data from multiple sources such as data provided by community health workers, hospitals, and social media and provide faster analysis of the spread of a disease, making policy decisions and implementation easier (Bharel, Auerbach, Nguyen, & DeSalvo, 2024).

While Gen AI has potential to improve healthcare delivery in India, there are also some perceived concerns it could bring especially with respect to medical professionals. One of the issues that could arise is over reliance on these systems in their work which could make them less attuned to their innate skills and observations. The easy access to GenAI systems could also mean that patients could also self-diagnose and self-medicate which could lead to medical emergencies (MacMillan, 2024). The absence of a liability framework and guidelines on the use of GenAI in practice could also mean that the medical professionals use this at their own risk and without proper training and support from institutions.

19.4 Conclusion

While AI is not set to replace medical professionals, there is still an uncertainty of what roles it will play in healthcare. As also seen in the survey and interview data the nascent stages of AI in healthcare in India mean that medical professionals are still using AI more as an added step to their existing workflow and spending

time improving the AI system through their feedback. On the other hand AI has also seen success in the larger context in areas with large manpower and expertise such as disease surveillance, and drug discovery, which have an immense potential to help in public health as well as reduce time it takes to make decisions and analyse trends. It is here that GenAI could improve their capacities and help regions like India where timely interventions could benefit both the public health system and the public. Hence while AI and in the future GenAI has the capacity to help healthcare, we need to prioritise areas where there could be most benefit and is in larger public interest with the least disruption to the existing workflow and be considerate of whether the costs (manpower as well as work time) outweigh the benefits.

19.5 References

About Accredited Social Health Activist (ASHA). (n.d.). Retrieved from https://nhm.gov.in index1.php?lang=1&level=1&sublinkid=150&lid=226.

AI in Healthcare: Changing India's Medical Landscape. (n.d.). Retrieved from https://indiaai.gov.in/article/ai-in-healthcare-changing-india-s-medical-landscape.

Ai, W. (2023, January 25). Empowering the Asha worker. Retrieved from https://www.wadhwaniai.org/2020/11/empowering-the-asha-worker/.

Alkhaldi, N. (2024, September 2). Assessing the Cost of Implementing AI in Healthcare — ITRex. Retrieved from https://itrexgroup.com/blog/assessing-the-costs-of-implementing-ai-in-healthcare/.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education, 23(1). https://doi.org/10.1186/s12909-023-04698-z.

Ameen, S., Wong, M., Yee, K., & Turner, P. (2022). AI and Clinical Decision Making: The Limitations and Risks of Computational Reductionism in Bowel Cancer Screening. Applied Sciences, 12(7), 3341. https://doi.org/10.3390/app12073341.

Anjaria, P., Asediya, V., Bhavsar, P., Pathak, A., Desai, D., & Patil, V. (2023). Artificial intelligence in Public Health: Revolutionizing epidemiological surveillance for pandemic preparedness and equitable vaccine access. Vaccines, 11(7), 1154. https://doi.org/10.3390/vaccines11071154.

Artpark. (n.d.). Dengue Dashboard — ARTPARK @IISc — Leading AI & Robotics startup incubation. Retrieved from https://artpark.in/health/dengue-dashboard.

Ayushman Bharat Digital Mission: Boon or Bane? (2023, May 30). Retrieved from https://accountabilityindia.in/blog/ayushman-bharat-digital-mission-boon-or-bane/.

Bharel, M., Auerbach, J., Nguyen, V., & DeSalvo, K. B. (2024). Transforming public health practice with generative artificial intelligence. Health Affairs, 43(6), 776–782. https://doi.org/10.1377/hlthaff.2024.00050.

Chun, M. (2023, March 8). How Artificial Intelligence is Revolutionizing Drug Discovery — Bill of Health. Retrieved from https://blog.petrieflom.law.harvard.edu/2023/03/20/how-artificial-intelligence-is-revolutionizing-drug-discovery/.

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. Journal of Multilingual and Multicultural Development, 23(4), 245–259. https://doi.org/10.1080/01434630208666468.

ET HealthWorld & www.ETHealthworld.com. (2020, May 15). Tech Mahindra R&D unit leverages AI for research on potential therapeutic drugs for COVID-19. ETHealthworld.com. Retrieved from https://health.economictimes.indiatimes.com.

FAQ — AI chatbot | Online Therapy. (n.d.). Retrieved from https://www.wysa.com/faq.

Karpagam, S. (2021, October 10). Why the New Ayushman Bharat Digital Mission Does Not Bode Well. Retrieved from https://science.thewire.in/health/ayushman-bharat-digital-mission-technofix-public-healthcare-privatisation/.

MacMillan, C. (2024, January 8). Generative AI for Health Information: A guide to Safe use. Retrieved from https://www.yalemedicine.org/news/generative-ai-artificial-intelligence-for-health-info.

Progress of healthcare artificial intelligence in India. (n.d.). Retrieved from https://www.isb.edu/en/research-thought-leadership/research-centres-institutes/isb-institute-of-data-science/Opinion-Pieces/Progress-of-Healthcare-Artificial-Intelligence-in-India.html.

Pti. (2023, December 11). Doctors with knowledge of artificial intelligence will hold advantage: Andrew Elder. The Economic Times. Retrieved from https://economictimes.indiatimes.com.

Saha, S. (2024, October 1). How This Startup is Making You Eat Healthy with Generative AI. Retrieved from https://analyticsindiamag.com/industry-insights/ai-startups/how-this-startup-is-making-you-eat-healthy-with-generative-ai/.

State government to screen kidney diseases with AI-powered mobile app. (n.d.). Retrieved from https://indiaai.gov.in/news/state-government-to-screen-kidney-diseases-with-ai-powered-mobile-app.

Yasmeen, A. (2024, July 31). AI-based lung cancer screening in Karnataka detects 133 lung nodule malignancy and nearly 3,000 TB-presumptive cases in last nine months. Retrieved from https://www.thehindu.com/news/national/karnataka/ai-based-lung-cancer-screening-in-karnataka-detects-133-lung-nodule-malignancy-and-nearly-3000-tb-presumptive-cases-in-last-nine-months/article68465107.ece.

20 Reimagining Education: Potential Solutions for Nomads

Faizo Elmi

Abstract

This essay examines the transformative potential of artificial intelligence (AI) in enhancing educational opportunities for nomadic populations. By leveraging AI technologies, such as adaptive learning platforms and virtual classrooms, educational access can be tailored to meet the unique needs of mobile communities. The essay explores how AI-driven tools can provide personalized learning experiences, bridge educational gaps, and support continuous learning despite geographical constraints. It also addresses the challenges of integrating AI in such contexts, including technological infrastructure and cultural considerations. Ultimately, the paper argues that AI holds significant promise for delivering equitable and flexible education to nomadic groups.

Keywords: Education, Nomads, AI, Mobile Communities, Educational Gaps.

Introduction

A common assumption with the term "nomadic" is that it is a culture that is fundamentally at odds with the modern world. For centuries, nomadic populations have resisted industrialization, modernization, and domination. In fact, nomadic populations are largely considered to be "aimless wanderers, immoral, promiscuous and disease-ridden" peoples (Hill & Randall, 2022). While this mindset is untrue and condescending, the fact remains that nomadic people live in such a way that makes traditional access and implementation of education rather difficult.

The idea of traditional education is often met with distrust among nomadic groups. Historically, the issue of educating nomadic populations has come from one of two viewpoints (Dyer, 2006). First, many academics suggest that education would allow for

nomadic populations to assimilate into settled society. The second more altruistic reason relates back to the United Nations Children Fund (UNICEF) stating that education is a human right and a key factor in reducing child labour and poverty (Dyer, 2006). Rapidly developing technology can now allow for nomadic education to no longer be approached from an either-or mindset. With some flexibility and modern technology, there are possibilities to meet in the middle. With a majority of nomads living across Africa, this issue is especially poignant.

20.1 Traditional Nomadic Approach to Education

Traditionally nomadic people have provided their children with a fulfilling education, with little to no say from institutionalized powers (Krätli, 2001). The environmental, economic, and historical knowledge that children needed to know was passed down from generation to generation. This equipped nomadic children with the appropriate skills and context to not only survive in their respective domains, but also attain professional positions within their respective societies (Krätli, 2001). The nomads became proficient in whatever local knowledge that was needed.

Beginning with the decade following the Second World War, rapid industrialization merged with post-colonial borders began to create a society that deemed nomads as being obsolete (Dyer, 2006). Modernization has made it challenging for nomadic people to continue with their traditions and many of their education systems are no longer adequate in preparing their children for their adult lives (Dyer, 2006).

Formal education was largely considered to be a waste of time by many nomadic groups (Jama,1993). School curriculums have been criticized by nomadic educators for being “made by sedentary people for sedentary people” (Dyer, 2006). For many nomads, the content that was being presented in these curricula, were not at all practical to their lived realities. Lack of applicability would eventually translate into lack of interest, and result in nomadic groups across the globe having some of the highest dropout rates in their respective countries (Dyer, 2006). Additionally, nomadic children are more likely to experience cultural alienation when they do attend

school. Nomads often have a strong sense of pride in their identity, as nomad Krätli writes:

> *"Pastoralists' strong sense of dignity is linked to pride in their own identity as pastoralists, nomads and a distinct ethnic group. Such a perception of themselves may be increasingly undermined by propaganda depicting them as ignorant, poor, dependent and powerless, made even more destructive by a feeling of being cheated in almost all interactions with the wider society" (Krätli, 2001).*

Seeing as how one of the more common solutions in educating nomadic children is boarding school, it is not shocking to think that students would begin to identify more with the dominant culture.

20.2 Challenges with Educating Nomadic Populations

There are many technical and cultural challenges involved in educating nomadic populations. As mentioned earlier, lack of applicability is one of the more common reasons, but there are also several other factors that contribute to the issue as well.

More often than not, nomadic groups are situated in remote areas that would require a great amount of effort to reach. This makes it difficult to provide both teachers and resources in those sparsely populated areas. In addition to this, migration patterns also determine when nomads will be in a certain area and for how long (Jama,1993). With many settled schools being either unable or unwilling to work to accommodate nomadic children results in high truancy rates among nomadic children (Dyer, 2006). Furthermore, seeing how nomadic groups also consistently have issues relating to poverty often mean that even those willing cannot afford to send their children to boarding schools, or pay for their upkeep in a settled village (Jama,1993). In addition to this, traditional education often undermines indigenous structures of education (Jama,1993).

Arguably most importantly is what refers to as the opportunity cost associated with sending children to schools. Children in many nomadic groups have certain responsibilities they are expected to

tend to. These often involve aspects of animal husbandry, family rearing, or other domestic tasks (Krätli, 2001). For many families this loss of a child would cause a significant burden in the household. What's more is the physical aspect of moving to and changing a family's entire course to be in closer proximity to a school is usually a hefty task. Particularly for pastoral nomadic families, who rely heavily on seasonal migratory patterns of animals (Jama,1993).

20.3 Previous Attempts at Educating Nomadic Groups

One of the more common solutions that have been utilized in the past is the creation of boarding schools. The goal was to provide suitable living conditions for nomadic children in hopes that this would improve student retention. However, boarding schools have also been unable to attract a large nomadic population. This is due to several reasons, one of which being the division within the family structure (Carr-Hill, Sedel, Eshete, & de Sousa, 2005). Children were being socialized away from their communities, resulting in a feeling of isolation from their communities. Krätli writes:

> *"when it comes to boarding school, no nomadic parents or children wish to be separated for long periods, usually with no way of communication. He also argued that the parents do not like the idea of giving custody of their children to people they do not know, to whom they are not related and whose moral integrity they often doubt" (Krätli, 2001).*

20.4 Case Study: Somalia

Somalia is largely arid and desert in climate, and a vast majority of the population were either nomadic or semi nomadic in nature (Konczacki, 1967). A devastating drought during the early 1970's caused immense damage to the Somali nomads traditional way of life in a way that it was never truly able to recover from (Shirwa, 1999). Currently, 32% of the population are still nomads (Federal Government of Somalia, 2022). These groups often move around in search of water and grazing areas for their livestock (Lewis, 2024). The semi-nomadic population remain in parts of the south living a

more agro-pastoral life, but still rely heavily on their livestock for their survival (Lewis, 2024).

While education in Somalia is a right and largely considered by the population to be a means for good, mistrust between nomadic groups and the government have resulted in very little improvement regarding the education of nomadic children. Several of the issues previously highlighted regarding reluctance in enrolling children in schools appear in conversation with Somalia nomadic parents. Household labour, lack of applicability and alienation are some of the reasons Somali parents are unwilling or unable to educate their children (Carr-Hill, 2015). The three most common reasons in this case were due to lack of availability, income, and constant migration (Carr-Hill, 2015). This paints the image that while enthusiastic about the prospect of education, it is not convenient enough for many pastoral nomads.

Technological development and education have arguably been interconnected since the printing press allowed for knowledge and information to be far more easily available. In 2022 when ChatGPT catapulted AI into the public consciousness, it sparked debates surrounding the role of AI in the classroom. To what extent could it be adapted to benefit both students and teachers? UNESCO's Global Education Monitoring Report 2023 states "these new tools can prove invaluable in providing personalized support for students, particularly those with disabilities or living in remote areas" (2023). In regards to nomadic children, AI can be particularly useful due to its personalized and flexible access. Despite their mobility and accessibility issues, nomadic children can still receive an education.

Ultimately, the goal of utilizing AI in this specific case would be to provide an education for nomadic children that caters to unique traditions and way of life. Using the Somali example, as of 2024 roughly 85% of Somali adults own a mobile phone (75, 2024). Seeing as though a majority of financial transactions in Somalia are done via the internet, it can be assumed that this is a largely technologically-literate population (75, 2024).

The solution being proposed is to take advantage of this perfect storm of able participants. Beginning with the creation of a primary

education server that can be accessed offline. Ideally, students in their primary years would be able to access a wide variety of learning resources that they would learn from while working alongside an AI tutor. Charity Help International (CHI) utilizes technology from two non profit organizations to create exactly this.

Learning Equality was founded in 2012 as a hopeful solution to the inequality present in internet access across the globe (2024). It would eventually go on to become a non-profit organization specializing in aiding educational equity through technology. With the ultimate goal behind the project being inclusivity, Learning equality emphasizes the importance of creating inclusive educational experiences for the widest range of students possible (2024). Aside from providing access to offline learning opportunities, it also utilizes Kolibiri, an open source education platform that provides educational services without the need for internet access (2024). Most importantly, educators can individually manage their own content, making this easy to tailor to local needs and curriculum (2024). The second organization is Kiwix. Which is essentially a free offline library that condenses various websites and articles in such a way that they are able to be easily downloaded and stored (2024). Together, both of these programs create an educational server that can be accessed through any piece of technology that is either Windows or Linux based. CHI in particular emphasizes that second hand technology (to a degree) is most cost efficient.

Learning Equality, and other programs of the sort have the potential to provide nomadic communities with both technological resources and access to quality education. Additionally, through the use of AI programs can be taught to create tailor made lessons for individual students that are both culturally sensitive and relevant. Not only would this contribute in bridging the educational gap, but it would also aid in improving digital literacy skills among students, and empower local communities by showing respect for their choices and traditions.

While this solution does have many benefits, it should be noted that there are certain challenges that would come along with implementing programs such as these on a wide scale. For instance there would be significant limitations on the technology itself. Whether that be

due to charging issues, storage concerns, or simple repairs, there are certain logistical issues that need to be met in order to guarantee success. Moreover, there is the nomadic community itself to consider. If they themselves are not involved during both the planning and implementation of this project, there may be some cultural disconnect that could cause unnecessary barriers.

In the case of Somali nomads, if a program such as the one mentioned above was institutionalized on a large scale, it has the potential to solve several problems surrounding educating Somalia's nomadic community. It would no longer be required for students to physically attend schools. This would mean that nomads would not have to arrange their travel plans around the accessibility of a school. Additionally, students would not have to be separated from their families for prolonged periods of time. Instead they would be able to work from the comfort of their own home. No longer would parents have to choose between alienating their child from their families or providing them with a decent education. This close proximity to home would also mean that the traditional practices of the nomads would be maintained. The funds that would have also been spent on school related necessities (uniforms, books, boarding) could also be reallocated back into the home. Without the confines of a traditional classroom, students would also be able to work at their own pace. This would make them available to participate in household labour when it was needed of them (to an acceptable degree).

Remote schooling would also allow for a certain amount of flexibility in learning. Students would be able to move through material at their own pace, this would be particularly beneficial for students who may have irregular schedules due to their travels. It would also mean continuity for these students as well. Without having to start afresh with their education every time they find themselves moving, students can move through their schooling without interruption. Even students who may have certain learning disabilities can benefit from this, as Learning Equality offers a wide range of resources in their lessons. These include interactive videos, texts, audio lectures, and more that can all help students of different abilities.

The Basic Accelerated Education Program in Somalia is a program that seeks to improve access to a quality accelerated education

for out of school youth (Federal Government of Somalia, 2022). In 2022 alone, improving access alone cost roughly 762,720$ (Federal Government of Somalia). If a program such as this was to be either partially or completely replaced by an AI remote learning platform of sorts, the costs associated with creating and new infrastructure could be reverted into investing towards remote learning programs. This would financially make accessing education easier for nomadic groups who historically spend the least on schooling, while also improving on the flexibility aspect of the program that the Ministry of Education prides itself on (Federal Government of Somalia, 2022).

In conclusion, AI offers a transformative opportunity to address educational disparities faced by nomadic populations. By integrating adaptive learning technologies and virtual classrooms, AI can tailor educational experiences to the unique needs of mobile communities, facilitating continuous and personalized learning regardless of location. However, successful implementation requires overcoming challenges related to technology access and cultural adaptation. Embracing AI in education for nomadic groups can bridge gaps and create more equitable learning opportunities, but it must be approached thoughtfully, with consideration of both technological and socio-cultural factors. As AI continues to evolve, it holds the potential to significantly enhance educational outcomes for nomadic populations.

20.5 References

Somalia — Information, Communication and Technology (ICT). Retrieved from https://www.trade.gov/country-commercial-guides/somalia-information-communication-and-technology-ict.

Carr-Hill, R. (2015). Education of children of nomadic pastoralists in Somalia: Comparing attitudes and behaviour. *International Journal of Educational Development*, *40*, 166–173. doi:10.1016/j.ijedudev.2014.10.001.

Carr-Hill, R. A., Sedel, C., Eshete, A., & de Sousa, A. (2005). *The Education of Nomadic People in East Africa: Synthesis Report.* African Development Bank | UNESCO-IIEP.

Dyer, C. (2006). *The Education of Nomadic Peoples: Current issues, future perspectives*. New York, NY: Berghahn Books, Incorporated.

Explore offline Wikipedia and educational content with Kiwix. (2024). Retrieved from https://kiwix.org/en/.

Federal Government of Somalia. 2021. Education Sector Analysis. Dakar. UNESCO. Retrieved From https://moe.gov.so/wp-content/uploads/2022/07/Somalia-Education-Sector-Analysis-Jan-2022-1.pdf.

Global Education Monitoring Report Team. (2023). *Global Education Monitoring Report 2023 technology in Education — a tool on whose terms? United Nations Educational, Scientific and Cultural Organization*. Erscheinungsort nicht ermittelbar: United Nations.

Hill, A. G., & Randall, S. (2022). Issues in the study of the demography of sahelian pastoralists and Agro-Pastoralists. *Population, Health and Nutrition in the Sahel*, 21–40. doi:10.4324/9781315831794-2.

Imagine a world in which all learners develop their agency, create positive transformation, and flourish. (2024). Retrieved from https://learningequality.org/.

Jama, M. A. (1993). Strategies on Nomadic Education Delivery. State of The Art Review. Somalia: UNICEF.

Konczacki, Z. A. (1967). Nomadism and Economic Development of Somalia: The Position of the Nomads in the Economy of Somalia. Canadian Journal of African Studies / Revue Canadienne Des Études Africaines, 1(2), 163–175. https://doi.org/10.2307/483530.

Krätli, S. (2001). *Education Provision to Nomadic Pastoralists*. Brighton, UK: Institute of Development Studies.

Lewis, I. M. (2024). Somalia. Retrieved from https://www.britannica.com/place/Somalia.

Ministry of Education (Somalia). (2022). *Education Sector Strategic Plan (ESSP) 2022-2026*. Ministry of Education (Somalia). https://moe.gov.so/wp-content/uploads/2022/07/ESSP-2022-2026.pdf.

TSUI, A. O., RAGSDALE, T. A., & SHIRWA, A. I. (1991). THE SETTLEMENT OF SOMALI NOMADS. Genus, 47(1/2), 131–152. http://www.jstor.org/stable/29789048.

21 The Need for Transnational Perspectives on the Social, Legal and Environmental Impact of Artificial Intelligence

Jess Reia, Rachel Leach and Anuti Shah

Abstract

The popularization of artificial intelligence (AI) models imposes various challenges, from human rights violations to energy consumption. While sustainability has been part of the AI agenda for years, environmental justice (EJ) is still making its way into AI regulatory frameworks. This essay discusses the need for transnational, climate-centered perspectives on AI regulation. Three questions guide this work: (1) Are EJ concerns considered when regulating and governing AI? (2) How do geopolitical power dynamics play into the environmental impact of AI? (3) How can EJ in AI regulation be improved? Here, we propose an exploratory analysis of cases in the US and Brazil.

Keywords: artificial intelligence; AI regulation; environmental justice; AI governance; climate crisis.

Introduction

Artificial intelligence (AI) models' increasing presence in the public debate, decision-making systems, and everyday life requires quick regulatory and policy answers. With a technological transformation spearheaded mostly by the private sector and its ability to shape research (Burrell & Metcalf, 2024), concerns around the technologies social, legal and environmental impact are receiving growing attention. The popularization of large language models (LLMs) and generative AI (genAI) have brought challenges from human rights violations to massive amounts of energy and water consumption.

While sustainability has been part of the AI agenda for years across countries (Wang et al., 2024), including the promise of AI as a solution to the climate crisis and tool to achieve the Sustainable Development Goals (SDGs) (Vinuesa et al., 2020), environmental justice is still making its way into AI regulatory frameworks and policies. As

environmental concerns are becoming more visible, issues of justice and sovereignty must not be overlooked. The official narrative at meetings 28 and 29 of the Conference of the Parties (COP), the United Nations (UN) Climate Change Conference, promotes "leveraging AI" for all to create opportunities for the energy and technology sector to work together (COP28 UAE, 2023). However, this narrative dismisses perspectives of other sectors and community organizations engaged in exposing the harms caused by AI systems.

The idea that AI systems exist in the cloud, disconnected from everyday life and material resources has been widely questioned by parts of academia, civil society, and grassroots movements (Bender et al, 2021; Castro et al., 2024; AlgorithmWatch, 2022). Beyond perpetuating colonial and digital extractivist practices (Ricaurte, 2019; Iyer, 2022), the expansion of big data infrastructure also poses threats to digital sovereignty (Belli & Hadzic, 2023) in global majority countries. Territories where valuable, scarce resources are available for extraction, like the lithium triangle (Chile, Bolivia and Argentina), are uniquely at risk. As Brazil regulates AI and prepares to host COP30 in 2025, concerns about digital infrastructure, sovereignty and human rights come to the forefront.

This essay's goal is to discuss the need for transnational perspectives on AI regulation that consider social and environmental justice components. The questions guiding this work are:

1. Are environmental justice concerns part of the process of regulating and governing AI internationally?
2. How do geopolitical and industry power dynamics play into the environmental impact of AI transnationally?
3. What are the first steps to improve environmental justice in AI regulatory frameworks?

To respond to these questions, this essay proposes an exploratory analysis of AI regulatory frameworks in the U.S. and Brazil. The impact of U.S. based Big Tech companies extends beyond borders, impacting the geopolitics, sovereignty and the environment in the global majority, highlighting the need for comparative and transnational studies. The findings presented here are part of a larger project funded by UVA's Environmental Institute addressing the impact

of incorporating AI systems in electric vehicles. The methods used were literature review, legal and policy analysis, and compilation of publicly available secondary data.

The first section investigates environmental and sovereignty imbalances caused by massive AI deployment. The second section briefly analyzes the concerns related to environmental justice in current AI regulatory efforts in two countries, United States and Brazil. Lastly, a brief exploratory attempt at considering environmental justice when regulating and governing AI that considers the global majority is presented.

21.1 The Geopolitics of AI: Sovereignty and the Environment

AI models, even the ones designed to mitigate the climate crisis, have a significant carbon footprint. They require an extensive chain of development which includes extraction of natural resources, manufacturing of materials and equipment, model training and deployment, and disposal. Despite this environmental cost, and emerging criticisms of Big Tech companies, leading AI developers continue to have revenues higher than the GDP of some countries (Patterson et al., 2021; Luccioni et al., 2022; Erdenesanaa, 2023; Vries, 2023; Microsoft, 2023).

The concept of embodied carbon, which accounts for "emissions associated with upstream — extraction, production, transport, and manufacturing — stages of a product's life," is useful to understand the full carbon footprint of AI (U.S. Environmental Protection Agency, 2024). Unlike other industries such as construction, standards governing AI do not "provide the methods to measure the embodied carbon within technology systems from a holistic systems-thinking perspective", ultimately leading to many upstream harms being overlooked and underregulated (Mulligan & Elaluf-Calderwood, 2022). Consequently, the proliferation of AI has impacted several "water-stressed regions, draining lakes and rivers while accelerating the displacement of vulnerable populations" (Hogan & Richer, 2024). For instance, data centers are often located in Latin America due to "lower environmental regulations than the U.S. and Europe" and cheaper access to resources despite the high risk of intensifying

climate change-induced drought" in the area (McGovern & Branford, 2024). AI also often uses rare minerals extracted from "zones with lax labour regulations, using methods that ravage landscapes, contaminate groundwater, and destroy natural habitats" (Hogan & Richer, 2024).

Along with upstream effects, the rapid and increasing energy use of AI model training and deployment creates additional environmental disparities. Specifically, "since 2012, the amount of computing power required to train cutting-edge AI models has doubled every 3.4 months" (Kanungo, 2023). As such, currently "training some popular AI models can produce about 626,000 pounds of carbon dioxide, the rough equivalent of 300 cross-country flights in the U.S." while "a single data center can require enough electricity to power 50,000 homes" (Rakhmanov, 2024). This rapid and expanding energy use affects "the world's most marginalized communities first" (Bender et al, 2021) as climate change destroys infrastructure (U.S. Global Leadership Coalition, 2021).

These environmental disparities are deeply connected to geopolitical and sovereignty issues. "Energy sovereignty" advocates and scholars work to shift understandings of who has a right to make decisions about energy from the current empowerment of large, almost entirely U.S.-based companies, which often work against the "self-determination and non-domination" of countries whose resources and energy they are extracting, and towards understanding energy "as a natural commons" respecting "the right of particular communities to decide on energy matters without the demand to increase profits" (Timmermann and Noboa, 2022; Stash Team, 2024; Del Bene, Soler, & Roa, 2019; Castro et al., 2024). Access to energy is a crucial issue of justice across many countries — in Brazil, a study led by Instituto Pólis found that "36 per cent of families spend at least half of their monthly income with power used for cooking and electricity, compromising" leading food insecurity and other economic issues (Instituto Pólis, 2024). So, as AI models consume extensive amounts of energy, countries must be able to understand and regulate these systems and their externalities (Belli et al., 2024).

21.2 Environmental Justice in AI Regulation and Policymaking

Environmental concerns are generally part of the AI agenda globally, but environmental justice concerns do not appear clearly in current regulatory efforts. Below we investigate how policymakers in the U.S. and Brazil are considering environmental issues in AI regulatory discussions.

21.2.1 United States

In 2023, the Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence set the federal agenda for AI regulation in the U.S. This Order explicitly operates under the assumption that AI is a tool with potential to "enable the provision of clean" electric power, without examining the environmental issues raised by the technology itself (Exec. Order No. 14110, 2023). In doing so, this EO, along with subsequent Executive actions, reinforces the idea that AI systems are a solution to, rather than a component of, unsustainable energy use and environmental harms. Ultimately, these assumptions contribute to the sidelining of environmental justice in AI regulation, policymaking, and funding in the United States.

Federal guidance on AI funding and investment encourages the rapid development of AI systems without due consideration of environmental justice. For instance, The National AI R&D Strategic Plan, which outlines priorities for Federal AI investment, suggests only technical solutions to the environmental harms of AI, such as designing "resource-aware AI algorithms" without mentioning the disproportionate impact of these harms (U.S. Select Committee on Artificial Intelligence, 2023). Other relevant Executive guidance, such as the National Institute of Standards and Technology (NIST) AI Risk Management Framework on Generative AI and the Office of Management and Budget (OBM) Policy on Government Use of AI encourage the application of AI systems to "address the climate crisis" without due consideration of their disproportionate environmental impact (The National Institute of Standards and Technology, 2024; U.S. Office of Management and Budget, 2024). Even the voluntary agreement from Big Tech companies to "Manage the Risks Posed

by AI" only mentions AI as a possible solution to climate problems, not as a contributor to them (The White House, 2023).

One of the only places environmental justice concerns are discussed is by the National Artificial Intelligence Advisory Committee (NAIAC), a committee under the Department of Commerce to advise the President on AI. In early 2023, before the Executive action discussed above, they cite research on the considerable energy use, water use, and extraction underpinning current AI systems; yet these considerations are not extended to more impactful documents within this group, nor to other regulatory and legislative efforts regarding AI (National AI Advisory Committee, 2023; National AI Advisory Committee, 2024).

Along with Executive action, the U.S. Senate created a Bipartisan AI Working Group to gather information on AI conducting forums largely made up of industry representatives. Their recommendations do not mention environment, climate, or sustainability, and addressing the "rising energy demand" of AI is mentioned only to "ensure the U.S. can remain competitive with the CCP and keep energy costs down" (The Bipartisan Senate AI Working Group, 2024). Earlier in 2024, Senator Markey introduced the US AI Environmental Impacts Act, "the first legislation to explicitly refer to the environmental impacts of AI" (Adams, 2024). However, no action has been taken on this bill since it was referred to the committee in February.

21.2.2 Brazil

The South American country has a prominent role in international multistakeholder spaces of internet governance and data protection, previously paving the way for AI regulation with the Brazilian Civil Rights Framework for the Internet (2014) and the General Personal Data Protection Law (2020). Attempts to regulate AI in Brazil started with the launch of the Brazilian Strategy for Artificial Intelligence (EBIA) (Brazilian Ministry of Science, Technology, and Innovations, 2021; Belli et al., 2023), even though legislators tried to propose draft bills ("*projetos de lei* — PL") between 2019 and 2021 (PL 5051/2019, 5691/2019, 21/2020, 872/2021), unsuccessfully. Regardless of the efforts to improve civic engagement with these bills and the EBIA, "[...] researchers and civil society advocates have been pointing

out the lack of consideration given by public authorities to the suggestions of participants in consultative processes" (Belli et al., 2023, p. 2).

In 2022, the Brazilian Senate appointed a Commission of 18 Legal Practitioners to help draft the AI regulatory framework for Brazil. Quickly, civil society organizations and scholars criticized the lack of diversity, pointing out that all members were white and did not represent the majority of the population, who would most likely be affected by bias, algorithmic discrimination (Kremer et al., 2023) and climate change. Experts called for interdisciplinary contributions and broader social participation (Rená, 2022). In response, the Commission scheduled twelve public hearings and launched a public consultation process, resulting in the current draft bill, PL 2338/2023, which is under discussion.

PL 2338/2023 covers various issues, mentioning the protection of the environment and sustainable development as principle, and states that AI agents must report to the competent authority the occurrence of serious security incidents, including severe damage to the environment. In an open letter released by a coalition of civil society organizations working toward digital rights and public interest technology suggested a list of possible improvements to PL 2338/2023 including "minimum rules to safeguard the rights of affected individuals, obligations for AI agents, governance measures, and the definition of a regulatory framework for oversight and transparency" (Bernar, 2024). There is a brief mention of the role of AI in exacerbating climate change, but no further recommendations.

Despite Brazil's strong environmental frameworks, including the National Policy on Climate Change (12187/2009), AI regulation has paid little attention to the environmental impact of models like LLMs and genAI. On July 30, 2024, following the G20 summit, the Brazilian government launched the Brazilian Plan for Artificial Intelligence, with BRL 4 billion in investments to address extreme weather, develop renewable energy, and build a supercomputer. While there are hopes that PL 2338/2023 will incorporate climate action, industry associations oppose stricter regulations, arguing it would hinder innovation and make Brazil less attractive for data centers (Viana, 2024). With its vast water resources and the Amazon, addressing climate action in AI regulation is urgent.

Table 1 Preliminary summary of the status of AI and environmental impact regulation in the U.S. and Brazil

COUNTRY	STATUS OF AI REGULATION	GUIDELINES & PRINCIPLES	IS THE ENVIRONMENTAL IMPACT OF AI ADDRESSED?	RELEVANT ENVIRONMENTAL REGULATION	RELEVANT STANDARDS	OVERSIGHT AGENCIES, BOARDS & REGULATORS
UNITED STATES	No comprehensive federal legislation or regulations	White House Executive Order 14110; White House Blueprint for an AI Bill of Rights, with contributions from panels including "industry stakeholders to technology developers to other experts across fields and sectors" and received 130 responses to the RFI	Partially, as seen in government, industry, and academic groups, AI is largely understood as a beneficial tool for climate and sustainability issues, rather than a contributor to them	National Environmental Policy Act (42 USC 4321); Clean Air Act (42 USC 7401); Clean Water Act (33 USC 1251); Endongered Species Act (16 USC 1531); Resource Conservation and Recovery Act (42 USC 6901); Comprehensive Environmental Response, Compensation, and Liability Act (42 USC 9601); Toxic Substances Control Act (15 USC 2601)	Yes, but mostly limited to particular sectors. Examples: NIST Special Publication 1270; NIST Special Publication 800-53; NIST Special Publication 800-37; OMB Guidance on AI; FDA Guidance for AI and Machine Learning in Medical Devices; SEC and FINRA Guidelines on AI in Financial Markets	No AI-specific federal regulator in the US. Federal Trade Commission; Federal Communications Commission; Equal Employment Opportunity Commission; Consumer Financial Protection Bureau and Department of Justice authority applies to "software and algorithmic processes, including AI." National Institute of Standards and Technology
BRAZIL	PL 2338/2023 is currently being discussed in the Senate, following public hearings, public consultation and the support of a Commission of Legal Practitioners	Brazilian Strategy for Artificial Intelligence (EBIA) launched in 2022	Yes, the protection of the environment and sustainable development is a principle in the proposed regulation. And AI agents must report to the competent authority the occurrence of serious security incidents, including severe damage to the environment	Brazilian Environmental Policy (6938/1981); Forest Code (12651/2012); National Policy on Climate Change (12187/2009); Biodiversity Law (13123/2015); National Solid Waste Policy (12305/2010); Environmental Crimes Law (9605/1998); National System of Conservation Units (SNUC) (9985/2000); New Sanitation Framework (14026/2020)	Yes. Examples: Norma ABNT NBR ISO/IEC 42001:2024; ABNT NBR ISO/IEC 38507:2023; ABNT NBR ISO/IEC 23894-2023; ABNT NBR ISO/IEC 22989:2023	The Federal government is expected to designate a competent authority, which will be the agency or entity of the Federal Public Administration responsible for implementing and overseeing Brazil's proposed AI Regulation (new or existing)

ABNT = Associação Brasileira de Normas técnicas; ISO = International Organization for Standardization; PL = draft bill (*projeto de lei*); RFI = Request for Information
FDA = Food and Drug Administration; NIST = National Institute of Standards & Technology.
Source: created by the authors based on existing policies, guidelines and legislation.

21.3 Final remarks: Addressing AI's hidden Environmental and Social costs Transnationally

As countries and jurisdictions regulate AI, including transnational perspectives about climate action and environmental justice is crucial. It is possible to reimagine a comprehensive regulatory ecosystem that includes hidden costs, especially for historically marginalized communities and global majority countries. Below are five preliminary considerations to move this process forward.

Transnational multistakeholder engagement: Connect discussions taking place within spaces like the UN Internet Governance Forum and COP, building bridges and opportunities for multistakeholder collaboration that includes the interests of the countries mostly affected by extreme weather and/or that have faced historical extraction of resources. The efforts should no longer be siloed.

Consider data infrastructure and embodied carbon: When considering environmental and social costs of AI, it is crucial to consider the resources that go into developing and training AI models. From the extraction of raw materials used in hardware to the disposal and recycling of outdated technologies, the AI lifecycle includes energy-intensive processes like data training and storage, which significantly contribute to carbon emissions. Furthermore, resources like cobalt, a critical component in the batteries that power AI hardware, have been linked to severe environmental degradation and human rights abuse, including child labor, hazardous working conditions, and the displacement of local communities.

Learn from other sectors and industries: Regulatory efforts in industries like medical, construction, and civil aviation offer valuable lessons for "big data ecologies" (Hogan, 2018), both in terms of best practices and mistakes to avoid. Existing tools and guidelines can help incorporate risk assessments, accountability mechanisms, and life cycle planning for technology, all grounded in a human rights-based approach that considers both big data and environmental justice.

Listen to people: Advocacy efforts and coalitions can play a big role in increasing awareness about the hidden environmental and social costs of AI. Meaningful community engagement is essential in this process, going beyond public hearings and online public consultations that, despite

their importance, might not reach the communities most affected by harmful technologies. To build trust in technologies and governments, moving away from opaque concepts and into actual efforts to incorporate people's needs and perspectives is a good starting point.

Changes in higher education: Data science is growing as an interdisciplinary field, with more undergraduate and graduate programs offered in higher education (Academic Data Science Alliance, 2024). As universities invest in powerful computing and data centers, it is crucial to incorporate environmental and data justice into curricula (see data collected from syllabi in Appendix I). While other STEM fields, like Engineering and Computer Science, established ethical standards in the 1970s (Hoffmann & Cross, 2021), teaching data ethics remains challenging, and the environmental impact of AI is often ignored, with courses focusing more on AI's role in addressing environmental issues rather than its consequences.

21.4 References

Academic Data Science Alliance. (2024). Member Institutions. https://academicdatascience.org/members-and-communities/current-members/.

Adams, C. (2024). Why we endorsed the US AI Environmental Impacts Act. *Green Web Foundation*. https://www.thegreenwebfoundation.org/news/why-we-endorsed-the-us-ai-environmental-impacts-act/.

AlgorithmWatch. (2022). *Ensure minimum transparency on the ecological sustainability parameters for all AI systems in the AI Act.* https://algorithmwatch.org/en/wp-content/uploads/2022/04/Sustainability-issue-paper-April2022.pdf.

Belli, L. (2023). *To Get Its AI Foothold, Brazil Needs to Apply the Key AI Sovereignty Enablers (KASE)* (SSRN Scholarly Paper 4465501). https://doi.org/10.2139/ssrn.4465501.

Belli, L., Curzi, Y., & Gaspar, W. B. (2023). AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience. *Computer Law & Security Review*, *48*, 105767. https://doi.org/10.1016/j.clsr.2022.105767.

Belli, L., & Hadzic, S. (Eds.). (2023). *Community Networks: Building Digital Sovereignty and Environmental Sustainability*. Publicações Direito Rio. https://www.defindia.org/wp-content/uploads/2024/04/Community-Networks_Building-Digital-Sovereignty-and-Environmental-Sustainability_E-book-.pdf.

Belli, L., Britto Gaspar, W., & Singh Jaswant, S. (2024). *Data Sovereignty and Data Transfers as Fundamental Elements of Digital Transformation: Lessons from the BRICS Countries* (SSRN Scholarly Paper 4903196). https://doi.org/10.1016/j.clsr.2024.106017.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922.

Bernar, L. (2024). Open Letter: Advocating for Brazilian AI regulation that protects human rights. *Coalizão Direitos na Rede*. https://direitosnarede.org.br/2024/07/08/open-letter-advocating-for-brazilian-ai-regulation-that-protects-human-rights/.

Brazilian Ministry of Science, Technology, and Innovations. (2021). *Estrategia Brasileira de Inteligencia Artificial.* https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-diagramacao_4-979_2021.pdf.

Burrell, J., & Metcalf, J. (2024). Introduction for the special issue of "Ideologies of AI and the consolidation of power": Naming power. *First Monday*. https://doi.org/10.5210/fm.v29i4.13643.

Castro, A., Ponce de León, A., Cantera, A. L., Olofsson, V., & Reina-Rozo, J. D. (2024). Energy sovereignty storytelling: Art practices, community-led transitions, and territorial futures in Latin America. *Tapuya: Latin American Science, Technology and Society*, *7*(1), 2309046. https://doi.org/10.1080/25729861.2024.2309046.

COP28 UAE. (2023). *COP28 Presidency calls for global effort to leverage the rise of AI, the energy transition and the growth of the Global South to accelerate sustainable development for all*. (2023). https://www.cop28.com/en/news/2024/06/COP28-Presidency-calls-for-global-effort-to-leverage-the-rise-of-AI.

Del Bene, Soler, & Roa (2019). Energy Sovereignty. In A. Kothari, A. Salleh, A. Escobar, F. Demaria, & A. Acosta (Eds.), Pluriverse: A Post-Development Dictionary (pp. 178-181). Tulika Books.

Erdenesanaa, D. (2023, October 10). A.I. Could Soon Need as Much Electricity as an Entire Country. *The New York Times*. https://www.nytimes.com/2023/10/10/climate/ai-could-soon-need-as-much-electricity-as-an-entire-country.html.

Exec. Order No. 14110, 3 C.F.R. 75191-75226 (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Hoffmann, A. L. & Cross, K. A. (2021). Teaching Data Ethics: Foundations and Possibilities from Engineering and Computer Science Ethics Education [Report]. University of Washington Research Works Archive. Retrieved from http://hdl.handle.net/1773/46921.

Hogan, M. (2018). Big Data Ecologies. Ephemera: theory & politics in organization, 18(3), p. 631-657. https://ephemerajournal.org/sites/default/files/pdfs/contribution/18-3hogan.pdf.

Hogan, M., & Richer, T. (2024). *Extractive AI*. Centre for Media, Technology and Democracy. https://www.mediatechdemocracy.com/climatetechhoganlepagericher.

Instituto Pólis. (2024). *JUSTIÇA ENERGÉTICA Pesquisa de Opinião Pública*. https://polis.org.br/wp-content/uploads/2024/06/justica-energetica.pdf.

Iyer, N. (2022). *Neema Iyer: Digital Extractivism in Africa Mirrors Colonial Practices* (P. Kannan, Interviewer) [Interview]. https://hai.stanford.edu/news/neema-iyer-digital-extractivism-africa-mirrors-colonial-practices.

Kanungo, A. (2023). *The Real Environmental Impact of AI*. Earth.Org. https://earth.org/the-green-dilemma-can-ai-fulfil-its-potential-without-harming-the-environment/.

Kremer, B., Nunes, P. & Lima, T. G. L. (2023). *Racismo algorítmico*. Rio de Janeiro: CESeC.

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model* (arXiv:2211.02001). arXiv. http://arxiv.org/abs/2211.02001.

McGovern, G., & Branford, S. (2024). *Critics fear catastrophic energy crisis as AI is outsourced to Latin America*. Mongabay Environmental News. https://news.mongabay.com/2024/03/critics-fear-catastrophic-energy-crisis-as-ai-is-outsourced-to-latin-america/.

Microsoft. (2023). 2024 Environmental Sustainability Report. *Global Sustainability*. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lMjE#page=1.

Mulligan, C., Elaluf-Calderwood, S. AI ethics: A framework for measuring embodied carbon in AI systems. AI Ethics 2, 363–375 (2022). https://doi.org/10.1007/s43681-021-00071-2.

National AI Advisory Committee. (2023). *Rationales, Mechanisms, and Challenges to Regulating AI.* https://ai.gov/wp-content/uploads/2023/07/Rationales-Mechanisms-Challenges-Regulating-AI-NAIAC-Non-Decisional.pdf.

National AI Advisory Committee. (2024). *Reports*. https://ai.gov/naiac/.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon Emissions and Large Neural Network Training* (arXiv:2104.10350). arXiv. https://doi.org/10.48550/arXiv.2104.10350.

Rakhmanov, I. (2024). *Council Post: Navigating AI's Eco Impact* [Nordgren]. Forbes. https://www.forbes.com/councils/forbestechcouncil/2024/03/14/navigating-ais-eco-impact/.

Rená, P. (2022, May 17). *Inteligência artificial no Brasil: O episódio da Comissão de Juristas*. IRIS-BH. https://irisbh.com.br/inteligencia-artificial-no-brasil-o-episodio-da-comissao-de-juristas/.

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. *Television & New Media*, *20*(4), 350–365. https://doi.org/10.1177/1527476419831640.

Stash Team. (2024). *15 Largest AI companies in 2024*. Stash Learn. https://www.stash.com/learn/top-ai-companies/.

The Bipartisan Senate AI Working Group. (2024). *Driving U.S. Innovation in Artificial Intelligence.* https://www.young.senate.gov/wp-content/uploads/Roadmap_Electronic1.32pm.pdf.

Timmermann, C., & Noboa, E. (2022). Energy Sovereignty: A Values-Based Conceptual Analysis. *Science and Engineering Ethics*, *28*(6), 54. https://doi.org/10.1007/s11948-022-00409-x.

The White House. (2023). Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI [Fact Sheet]. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

U.S. Environmental Protection Agency. (2024). What is Embodied Carbon? https://www.epa.gov/greenerproducts/what-embodied-carbon.

U.S. Global Leadership Coalition. (2021). *Climate Change and the Developing World: A Disproportionate Impact*. U.S. Global Leadership Coalition. https://www.usglc.org/blog/climate-change-and-the-developing-world-a-disproportionate-impact/.

U.S. National Science Foundation. (2024). *Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot.* https://new.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai.

U.S. National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

U.S. Office of Management and Budget. (2024). *Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence*. https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.

U.S. Office of Science and Technology Policy. (2022). *Blueprint for an AI Bill of Rights.* https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

U.S. Select Committee on Artificial Intelligence. (2023). *National Artificial Intelligence Research and Development Strategic Plan 2023 Update*. https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.

Viana, N. (2024). 'A inteligência artificial tem sede — e está de olho no Brasil,' *Agência Pública*, 30 Jul 2024. https://apublica.org/2024/07/a-inteligencia-artificial-tem-sede-e-esta-de-olho-no-brasil/.

Vinuesa, R., Azizpour, H., Leite, I. et al. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233. https://doi.org/10.1038/s41467-019-14108-y.

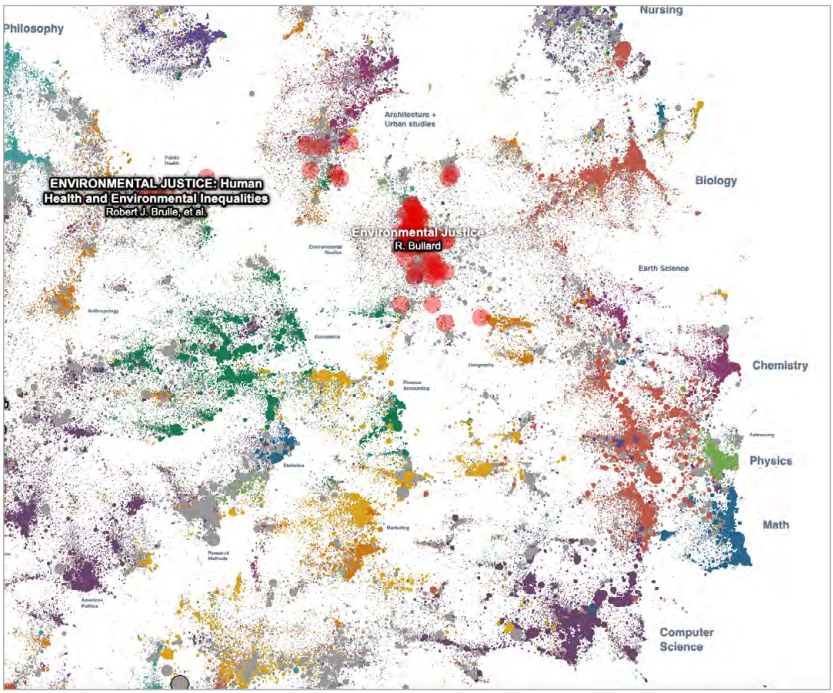
Vries, A. de. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004.

Wang, Q., Li, Y. & Li, R. (2024). Ecological footprints, carbon emissions, and energy transitions: the impact of artificial intelligence (AI). Humanit Soc Sci Commun 11, 1043. https://doi.org/10.1057/s41599-024-03520-5.

Appendix

The data in Fig. 1 represents the most frequently assigned texts in the Open Syllabus corpus, focusing on environmental justice citation. The graph is formed by connecting syllabi with the books and articles assigned in the course, with some areas overlapping or keeping distance. Environmental studies rarely converse with fields like computer science, indicating that students often do not have much contact with environmental studies and environmental justice scholarship.

Figure 1 Most frequently assigned texts in syllabi according to different disciplines, based on environmental justice citations.



Source: Open Syllabus Galaxy (https://galaxy.opensyllabus.org/)[71]

71 More information: https://blog.opensyllabus.org/galaxy-v2.

PART 5

FORESIGHTED SOLUTIONS FOR PRESENT PROBLEMS

22 Rewriting the Rules of the Game: Epistemological and Ontological Challenges at the Intersection of Legal Science and Data Science

Matheus Alles

Abstract

This study examines the epistemological and ontological challenges emerging from the intersection of data science and law. Through a multidisciplinary lens, it analyzes how predictive and language models in the legal domain challenge traditional legal theory and redefine key legal categories. The research highlights the tension between data-driven approaches and conventional legal reasoning, emphasizing the need for a reflexive legal rationality. This framework aims to critically integrate data science insights while maintaining the nuanced interpretative nature of legal thought. The article explores the implications of algorithmic decision-making in law, addressing issues of transparency, accountability, and the changing nature of legal knowledge. The methodology employs a multidisciplinary approach, combining conceptual analysis and literature review. By proposing a balanced approach that harnesses the power of data science without compromising legal principles, this study contributes to the ongoing dialogue on responsible integration of technology in legal practice and theory.

Keywords: Data science, legal epistemology, legal ontology, reflexive rationality, legal theory.

Introduction

The intersection between data science and law has become increasingly prominent, bringing not only opportunities for innovation and efficiency in legal practice but also profound challenges to the theoretical and conceptual foundations of law. This article proposes a critical reflection on how data science not only transforms legal practice but also disrupts the epistemological and ontological bases of law, demanding a radical reconstruction of legal theory.

This study adopts a global perspective, recognizing the diversity of legal systems and cultural contexts in which the intersection of legal science and data science manifests. In doing so, it seeks not only a comprehensive analysis but also an understanding of the nuances and specific challenges faced by different jurisdictions in the digital era

The central problem addressed in this study is the epistemological and ontological tension that arises when traditional legal reasoning, rooted in a hermeneutic and argumentative tradition, confronts the new forms of rationality and knowledge introduced by data science. This tension manifests itself in various dimensions, from the application of predictive models in the legal context to the reconfiguration of fundamental legal categories mediated by algorithms.

Faced with this scenario, the objective of this article is to investigate the epistemological and ontological implications of applying data science in law, identifying the challenges and opportunities that this intersection presents for legal theory and practice. It seeks to contribute to the development of a reflexive legal rationality, capable of critically integrating insights from data science without relinquishing the interpretive richness and contextual sensitivity of legal reasoning.

To achieve this objective, the study adopts an interdisciplinary methodological approach, combining conceptual analysis and literature review. Starting from an exploration of the epistemological and ontological foundations of law, the article critically examines concrete examples of the application of data science in the legal context, such as the use of predictive models to anticipate judicial decisions.

Through this analysis, the study aims to identify the points of tension and the possibilities of reconciliation between traditional legal rationality and the new forms of knowledge introduced by data science. In doing so, the article seeks to contribute to the development of theoretical and methodological frameworks that enable a responsible and transparent integration of data science into law, promoting justice, equity, and social well-being in the digital age.

The methodology adopted in this study is rooted in a multidisciplinary approach, reflecting the complex nature of the intersection between

legal science and data science. The research methodology comprises two main components.

First, a conceptual analysis, in which the study conducts a thorough examination of the epistemological and ontological foundations of law, drawing from established legal theory and philosophy. This provides the theoretical framework for understanding the traditional bases of legal reasoning and knowledge.

Second, research analyses recent scholarly works, case studies, and practical applications of data science in the legal domain. This includes examining predictive models for judicial decisions, the use of language models in legal drafting, and the application of machine learning algorithms in legal analysis. Through this dual approach, the study aims to identify the points of tension and potential reconciliation between traditional legal rationality and the new forms of knowledge and analysis introduced by data science.

The article is also structured in two main parts. The first part explores the epistemological disruption caused by data science in law, examining how new forms of rationality and knowledge challenge the traditional foundations of legal theory. The second part, in turn, addresses the ontological disturbance generated by the application of data science in law, investigating how fundamental legal categories are reconfigured and re-signified in this process.

Throughout these two parts, the article develops an argument in favour of the need for a reflexive legal rationality, capable of critically integrating insights from data science without relinquishing the interpretive richness and contextual sensitivity of legal reasoning. It is through this critical and constructive engagement, it is suggested, that it will be possible to face the epistemological and ontological challenges imposed by the intersection between law and data science, harnessing its transformative potential to promote justice and social well-being in the digital age.

22.1 Discussion

Law, as a discipline of study, is rooted in an epistemological tradition that values hermeneutic interpretation, logical argumentation, and the construction of coherent narratives (Dworkin, 2014). This

tradition, which dates back to Roman jurisprudence and medieval exegesis, views law as a system of norms and principles that must be interpreted and applied to concrete situations through a process of legal reasoning (Berman, 1983).

Legal reasoning, in turn, is an intellectual process, not only methodological, but also from the intersection of this methodological tradition that involves an interpretive process in which there is a dialogue between the norm and social facts and the subsumption of one to the other, in a system of network connections, with legal rationality being the guiding thread and also the foundation that fills a certain gap in this system.

However, this legal reasoning faces a new scientific challenge that promotes a dialogue through a distinct epistemological paradigm — data science.

The latter, unlike legal science, is extracted through a quantitative and qualitative analysis of large volumes of information and the identification of patterns and correlations (Hey, Tansley & Tolle, 2009) that promotes, in the combined application with legal science, a new challenge to the hermeneutic interpretation based on the understanding of phenomena with legal repercussions.

At first glance, it is assumed that this approach between legal science and data science is similar to a system of evidence in common law as a source of law, beyond what is exclusively posited, but in the face of what is practiced at the time of the application of the law, however with a maximization of the capacity of processed, stored and provided information.

However, data science goes beyond the empirical decision-making capacity, reflecting on the axis of the information used until entering the result of the decision and its comparison with a specific concrete situation.

Elements originated from automation condition correlations of segregation, influence on decision-making, and reliability, which ends up generating an epistemological tension between these two sciences.

An example is the predictability of judicial decisions that extend beyond a debate that is merely empirical from the perspective

of precedents, but from a study of techniques for learning the predictability of decisions. In 2017, three authors wrote the article "A general approach for predicting the behavior of the Supreme Court of the United States" where a machine learning model was presented to predict the behavior of the Supreme Court of the United States of America. The model sought to encompass both the individual votes of Supreme Court justices and the overall outcomes of cases between 1816 and 2015 (Katz, Bommarito & Blackman, 2017).

The method used was called *Random Forest*, which evolves over time, taking advantage of feature engineering techniques that comprised more than 240,000 judge votes and 28,000 case outcomes. This interpretation was based on three principles: generality, consistency, and applicability (Katz, Bommarito & Blackman, 2017). The three principles aimed at general application, stable performance, and the possibility of repercussion outside the analyzed samples.

For this, the measurement of common and diverse elements between identification criteria, disagreement of the courts of origin, procedural aspects, and variables of historical behavior of decisions that encompass political directions, rate of disagreement, and reversal were used. During the analyzed period, the system obtained predictability criteria of 70.2% at the case levels and 71.9% at the individual vote levels (Katz, Bommarito & Blackman, 2017).

The system, which is beneficial in the face of an adaptation of precedents in the Supreme Court, runs the risk of generating a boomerang effect — which, when thrown, manifests the articulation of the petitioners in the face of knowledge about the vote of the decision-maker and how to adapt their petition for analysis of agreement, disagreement, or overcoming (distinguishing and overruling). On the other hand, the return of the throw is to the detriment of critical thinking, considering that the agenda of discussion moves away from the legal repercussion of the case discussed to the line of the decision-maker's vote.

The link between claim and decision, in this context, transcends the mere resolution of social controversies under legal protection, shifting to an adaptation of the petitioners' reasons to the decision-maker's decision-making history. A phenomenon that raises a critical

reflection on the core of legal discussion, which momentarily moves away from the reason of the law itself to orbit around the reason of the decision-maker, potentially compromising the integrity of the hermeneutic process.

From the perspective of Luhmann's Theory of Social Systems (2011), there is a reconfiguration of the dynamics between subsystems, where the guiding thread between fact and law is replaced by a connection between fact and precedent, the latter influenced by the decision-maker's history of motivations. In parallel, Dworkin's Theory of Law as Integrity (2014), originally conceived to strengthen legal certainty through a coherent system of precedents, is challenged by this new reality.

This intersection highlights the complexity of legal reasoning, especially in the face of technological advances and the proliferation of data, demanding a continuous re-evaluation of the epistemological foundations of legal science. In this scenario of transformation, the need to rethink legal epistemology emerges, seeking new approaches that can reconcile the hermeneutic tradition with the potentialities offered by data science.

The disruption caused by data science requires a critical re-evaluation of traditional legal epistemology, with the development of new theoretical frameworks that recognize the contribution of quantitative analysis without relinquishing the interpretive richness of legal reasoning.

However, the mere epistemological approach proves insufficient to face the emerging challenges. The intersection between data science and law raises a deeper disturbance that reaches the core of legal ontology. This disturbance transcends questions about methods and the nature of legal knowledge, reaching fundamental inquiries about the very nature and structure of legal reality.

This is because the introduction of data science in the legal field not only challenges the understanding of how law is known but questions what constitutes legal reality itself — which generates an ontological disturbance that emerges when traditional legal categories, constructed over centuries of legal thought, confront new forms of representation and analysis provided by data science (Hildebrandt, 2018).

Therefore, in rethinking legal epistemology, there is an inevitable conduction to reconsider the ontological bases of law, initiating a profound reflection on traditional legal categories and their adequacy to the contemporary technological context. As Mireille Hildebrandt (2018) observes, the integration of data technologies into law not only alters legal practices but challenges fundamental conceptions about what constitutes law and its entities.

Traditional legal ontology is built around abstract categories, such as "contract," "civil liability," and "crime," which emerge from interpretive and argumentative processes (Schauer, 2009). These categories are treated as real and objective entities that exist independently of the social and discursive practices that constitute them (Zheng, Jiang, Ding & Zaheer, 2022).

Data science, on the other hand, operates at a sub-symbolic level, identifying patterns and correlations that may not correspond to existing legal categories (Bengio, Lecun, Hinton, 2015).

Machine learning algorithms, for example, can identify clusters of cases or behaviors that do not fit into traditional legal taxonomies, revealing an alternative ontology based on statistical regularities (Mackenzie, 2015).

This ontological tension is aggravated by the problem of algorithmic opacity. Many machine learning algorithms operate as black boxes, producing results without providing a clear explanation of how those results were achieved (Pasquale, 2015). This lack of transparency raises questions about the accountability and legitimacy of algorithm-based decisions, especially when those decisions have significant legal consequences (Selbst & Barrocas, 2018).

The ontological tension and algorithmic opacity mentioned above find a pertinent illustration in the analogous and contemporary scenario of investments in Artificial Intelligence (AI).

As reported by Futurism (2023), Silicon Valley investors and Wall Street analysts are expressing growing concerns about the ability of technology companies to effectively monetize their AI initiatives. This case exemplifies how the introduction of new technologies, such as AI, can challenge not only established practices but also fundamental conceptual categories in a specific field.

The article highlights that, despite the enormous investments in AI — with Google, for example, projecting capital expenditures in excess of $49 billion in 2023 — there is growing uncertainty about the financial return on these technologies. This situation reflects the tension between traditional expectations of return on investment and the emerging reality of technologies whose value and impact are difficult to quantify in conventional ways.

The opacity mentioned in the original text finds a parallel in the difficulty investors have in understanding how these AI technologies will generate significant revenues. As noted by Jim Covello, senior analyst at Goldman Sachs, "Despite its expensive price tag, the technology is nowhere near where it needs to be in order to be useful" (Futurism, 2023).

The situation described in the article also resonates with the idea of an alternative ontology based on statistical regularities. AI models, by processing vast volumes of data and identifying patterns that may not correspond to traditional categories of business analysis, are effectively creating a new ontology of value and utility that challenges established conceptions in the world of investments.

The example of Silicon Valley and the world of AI investments demonstrates how the introduction of new technologies can disrupt not only operational practices but the fundamental ontological structures in fields such as finance, technology, and, by extension, law.

In parallel, there is an overreliance on AI as a science of analysis for other sciences, where the rupture of this reliance is already apparent when there is unverified use.

This situation of overreliance on AI, followed by growing concern about its real effectiveness, reflects what Luciano Floridi (2014) calls the infosphere — an increasingly complex and interconnected informational environment.

In the legal context, the infosphere challenges not only traditional epistemology but ontology, as there is an expectation that AI can autonomously and profitably solve complex legal problems, echoing a still limited understanding of the nature of legal knowledge and legal reality itself.

In law, this translates into a need to rethink fundamental legal categories and the very processes of legal reasoning and decision-making, especially considering the ethical and social implications of information technology.

The ontological disturbance raised by data science in the legal sphere is not an isolated event but mirrors a broader propensity for technological disruption in various spheres. The case of AI investments in Silicon Valley exemplifies how the implementation of new technologies can challenge established practices and basic conceptual categories.

This situation echoes the notion of an increasingly intricate and interconnected infosphere (Floridi, 2014), in which excessive trust in AI is followed by growing apprehension about its effectiveness.

In the legal field, this reality implies the need to rethink elementary legal categories and the very processes of reasoning and decision-making, seeking a balance between the efficiency promised by AI and principles such as justice, equity, and transparency.

Thus, the disturbance of legal ontology lies in its need to develop to encompass not only new technological entities but also new values and ethical precepts that emerge from the interaction between law and AI.

The ontological disruption caused by data science in the legal field demands a profound rethinking of the very nature of legal entities and categories. It is not merely a matter of adapting existing legal concepts to new technological realities but of recognizing that these technologies may fundamentally alter the ontological landscape of law.

This recognition requires a shift from a static, essentialist view of legal categories to a more dynamic, relational understanding of legal ontology. Legal entities and concepts must be seen not as fixed, pre-given realities but as emergent, context-dependent constructs that are shaped by the socio-technical practices in which they are embedded (Hildebrandt, 2018).

Such a relational ontology would acknowledge the constitutive role of data science in shaping legal reality while also preserving the normative and interpretive dimensions of law. It would require a dialogical approach that brings together legal expertise, technical

knowledge, and ethical reflection to navigate the complex terrain of law in the age of data (Danaher, 2016).

Moreover, this ontological shift must be accompanied by a commitment to transparency, accountability, and public engagement. The opaque and proprietary nature of many data science tools and methods poses significant challenges to the rule of law and democratic governance (Pasquale, 2015). Ensuring that these technologies are developed and deployed in a transparent, accountable, and inclusive manner is crucial for maintaining the legitimacy and integrity of the legal system.

In conclusion, the ontological disturbance generated by the intersection of data science and law represents a profound challenge to the foundations of legal thought and practice. Addressing this challenge requires not only new theoretical frameworks and methodological approaches but also a fundamental rethinking of the nature of legal reality and the role of law in the digital age.

By engaging critically and constructively with the ontological implications of data science, the legal community can develop a more adaptive, responsive, and ethically grounded approach to the challenges and opportunities of the 21st century. This engagement is essential not only for the future of law but for the future of society as a whole, as we navigate the complex terrain of the infosphere and the emerging realities of the data-driven world.

The epistemology and ontology previously analyzed necessitate the exercise of legal rationality, traditionally based on a series of assumptions about the nature of legal reasoning, including the belief in the logical coherence of the legal system, the possibility of arriving at correct answers through rational argumentation, and the autonomy of law in relation to other forms of knowledge.

However, data science introduces a new form of rationality in law, an algorithmic rationality that operates differently from traditional legal rationality, privileging correlation over causality, probability over certainty, and efficiency over coherence.

It is this algorithmic rationality that challenges the autonomy of law, suggesting that legal decisions can be influenced by patterns and trends identified in data external to the legal system.

But how can this conflict between rationalities be overcome?

First, it is essential to recognize that both forms of rationality have valuable contributions to offer to the legal field.

Traditional rationality, with its emphasis on hermeneutic interpretation, logical argumentation, and the construction of coherent narratives, is essential to maintain the integrity and legitimacy of the legal system. On the other hand, algorithmic rationality, with its ability to identify patterns and correlations in large volumes of data, can provide valuable insights and support more efficient and evidence-based decision-making.

To reconcile these two forms of rationality, it is necessary to develop theoretical and methodological frameworks that allow for the responsible and transparent integration of data science into law. This implies not only establishing clear guidelines for the collection, processing, and use of data in the legal context, ensuring privacy protection, fairness, and non-discrimination.

The core lies in the cognizable demonstration of the algorithms used in a transparent manner, allowing data-based decisions to be understandable and appropriate to the ontological and epistemological context in which legal hermeneutics operates, echoing the connections and ruptures that relate to social dynamism.

This translates into the demonstration of legal reasoning associated with a demonstration of data reasoning, which tends to mitigate the risks associated with algorithmic opacity and strengthen confidence in the use of data science in law.

This context is reflected in the exercise of human activity associated with the exercise of activity developed by AI.

For example, the use of advanced language models to assist in drafting more equitable and inclusive contracts and policies can be seen as a positive application of data science in the legal domain. These models have the potential to identify linguistic biases, suggest more neutral alternatives, and promote more accessible and understandable

language. In this sense, they can contribute to the promotion of justice and equality, aligning with fundamental legal principles.

However, it is crucial to consider the epistemological and ontological implications of this approach. From an epistemological point of view, it is necessary to question the extent to which language models can adequately capture and represent the complexity and subtlety of legal reasoning from a human perspective. The drafting of contracts and policies involves not only the choice of words but also the interpretation of legal concepts, the weighing of principles, and the consideration of specific contexts. Can language models, however advanced they may be, encompass this epistemological depth?

Furthermore, there is an ontological concern about how these models can influence the very nature and meaning of legal concepts. If the drafting of contracts and policies becomes mediated by algorithms, this may lead to a reconfiguration of traditional legal categories. Notions such as equity, inclusion, and justice may acquire new meanings and interpretations, shaped by the logic and limitations of language models. This ontological disruption can have profound implications for legal theory and practice.

The core of the problem lies precisely in the exercise of human rationality so that it is not replaced by the predictive activity of AI or the application of concepts obtained through its rationality. It is not a matter of a resolutive method, but of constant and interdisciplinary exercise, as a mechanism for preserving the integrity and autonomy of legal reasoning.

This means the emerging convergence between data science and law through the development of a reflexive legal rationality, capable of critically questioning its own assumptions and adapting to new forms of knowledge and rationality introduced by data analysis. This reflexive rationality implies a commitment to transparency, comprehensibility, and accountability of algorithmic systems used in law, as well as an openness to interdisciplinary and collaborative forms of legal knowledge production.

22.2 Conclusion

Data science presents a new epistemological paradigm that challenges traditional legal reasoning. To reconcile these two forms of rationality,

it is necessary to develop theoretical and methodological frameworks that allow for the responsible and transparent integration of data science into law.

This implies establishing clear guidelines for the collection, processing, and use of data in the legal context, ensuring privacy protection, fairness, and non-discrimination. The core lies in the cognizable and transparent demonstration of the algorithms used, allowing data-based decisions to be understandable and appropriate to the ontological and epistemological context of legal hermeneutics.

Data science goes beyond the empirical decision-making capacity, reflecting from the axis of the information used to the result of the decision and its comparison with the concrete situation. Elements such as automation, segregation, decision influence, and reliability generate an epistemological tension with legal science.

Recent studies, such as the machine learning model to predict the behavior of the U.S. Supreme Court, exemplify the disruptive potential of data science in law. However, to fully harness this potential responsibly, an interdisciplinary effort involving jurists, data scientists, and public policy makers is needed.

Furthermore, it is essential to consider the ontological implications of applying data science in law. The use of advanced language models to assist in drafting more equitable and inclusive contracts and policies, for example, raises questions about the ability of these models to capture the complexity and subtlety of legal reasoning. It is necessary to critically assess the extent to which these models can encompass the epistemological depth involved in interpreting legal concepts, weighing principles, and considering specific contexts.

Another concern is the influence these models can have on the very nature and meaning of legal concepts. Algorithmic mediation in the drafting of contracts and policies may lead to a reconfiguration of traditional legal categories, with notions such as equity, inclusion, and justice acquiring new meanings shaped by the logic and limitations of language models. This ontological disruption requires a deep reflection on the implications for legal theory and practice.

The implications of this study for future research are significant. There is a pressing need for empirical investigations into how different legal systems are adapting to the integration of data science. Furthermore, the development of ethical and regulatory frameworks for the application of AI in law emerges as a critical area for future research. In practice, the results of this study can inform the development of legal curricula, public policies, and organizational strategies for a more effective and ethical integration of data science in the legal field.

Faced with this complex scenario, the need emerges for the legal community to develop a reflexive rationality, capable of critically questioning its own assumptions and adapting to new forms of knowledge and rationality introduced by data analysis. This reflexive rationality implies a commitment to transparency, comprehensibility, and accountability of algorithmic systems used in law, as well as an openness to interdisciplinary and collaborative forms of legal knowledge production.

22.3 References

Alexy, R. (1989). A theory of legal argumentation: The theory of rational discourse as theory of legal justification. Oxford: Clarendon Press.

Bengio, Y., LeCun, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Berman, H. J. (1983). Law and revolution: The formation of the Western legal tradition. Cambridge, MA: Harvard University Press.

Dworkin, R. (2014). O império do direito (3rd ed.) (J. L. Camargo, Trans.). São Paulo: Martins Fontes.

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. Oxford: Oxford University Press.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Hildebrandt, M. (2018). Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. University of Toronto Law Journal, 68(supplement 1), 12-35.

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE, 12(4), e0174698.

Luhmann, N. (2011). Introdução à teoria dos sistemas (3rd ed.) (A. C. A. Nasser, Trans.). Petrópolis: Vozes.

Mackenzie, A. (2015). The production of prediction: What does machine learning want? European Journal of Cultural Studies, 18(4-5), 429-445.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Cambridge, MA: Harvard University Press.

Rouvroy, A., & Berns, T. (2013). Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation through relationships? Réseaux, 1(177), 163-196.

Schauer, F. (2009). Thinking like a lawyer: A new introduction to legal reasoning. Cambridge, MA: Harvard University Press.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham Law Review, 87(3), 1085-1139.

Zheng, Y., Jiang, S., Ding, W., & Zaheer, A. (2022). Ontology-based knowledge representation and semantic topic modeling for intelligent trademark legal precedent research. World Patent Information, 68, 102098.

23 People-Centered Justice AI: Data Dimensions for Embracing a Responsible Digital Transformation

Julio Gabriel Mercado

Abstract

The digital transformation of justice, driven by AI, must be guided by a people-centered approach to ensure its responsible and effective implementation. Simply digitizing the system will not close existing access gaps. Instead, adopting Open Justice principles, emphasizing transparency, accountability, and public participation, is crucial to driving the necessary cultural and organizational shifts. Particularly, Open Justice promotes the publication of judicial data in open, reusable formats, which is key to fostering innovation and inclusivity in AI-driven systems. The quality of the available data will largely determine whether AI's benefits are distributed equitably. To achieve this, five critical dimensions for the publication of data, i.e., standardization, accessibility, completeness, cybersecurity, and privacy, must be addressed. Tackling these issues requires coordinated efforts at national and global levels to ensure that AI advancements serve the public interest and uphold human rights.

Keywords: People-centered Justice, Artificial Intelligence, Open Justice, Open Judicial Data, Access to Justice.

23.1 Introduction

The integration of artificial intelligence (AI) into the justice sector represents a transformative opportunity, but one that must be grounded in a people-centered approach to ensure a responsible implementation. As highlighted by the OECD (2023), achieving this approach requires collaboration across the entire justice system, including courts, prosecutors, police, and correctional institutions, in order to establish robust regulatory frameworks and the necessary institutional strategies.

While digital transformation has progressed in some areas, fundamental change in justice systems has been slow, with core functions remaining largely unchanged for centuries (Muller et al., 2013). Meanwhile, a significant justice gap persists, with 1.5 billion people worldwide unable to resolve their justice problems and two-thirds of the global population lacking meaningful access to justice (Task Force on Justice, 2019). This gap disproportionately affects vulnerable groups such as women, low-income individuals, and ethnic minorities, exacerbating structural inequalities.

In this context, AI offers the potential to expedite judicial processes, reduce case backlogs, and assist in decision-making. However, a simple reliance on technology without addressing deeper cultural and organizational shifts may lead to failure, and even to increased digital exclusion (Addo et al., 2024). The principles of Open Justice, which emphasize transparency, accountability, and public engagement, provide a framework for addressing these challenges (Elena et al., 2019).

In particular, by encouraging the publication of judicial data in open, reusable formats, Open Justice ultimately supports the development of responsible AI systems that respect human rights and help address bias. The governance of data not only ultimately shapes the governance of AI but also largely determines the extent to which its benefits might be distributed equitably and the risks associated with its implementation can be mitigated (Datasphere Initiative, 2024). In this sense, addressing some critical dimensions of judicial data publication is key to ensure AI's eventual success in assisting the delivery of justice in a responsible manner.

This paper aims to define the key dimensions of judicial data (i.e., standardization, accessibility, completeness, cybersecurity, and privacy) that must be addressed to embrace a responsible deployment of AI in the justice sector. It explores the challenges of balancing these dimensions and the importance of coordinated national and global efforts to align AI's use in justice with public interest, fundamental rights, and fairness. Ultimately, it advocates for a people-centered approach that emphasizes inclusivity, transparency, and accountability, ensuring that the benefits of

digital transformation are distributed equitably, contributing to closing the global justice gap.

23.2 Digital Transformation, Open Justice and AI

The digital transformation of justice, particularly in the context of the use of AI, should aim at the adoption of a people-centered approach (OECD, 2023). A people-centered approach to justice can be defined as one that prioritizes a unified vision and purpose aimed at making the justice system more responsive to peoples' needs. This can be done through designing and delivering services based on the justice journey of different groups, with a focus on those populations that face the greatest barriers to accessing justice.

According to recent figures, a total of 1.5 billion people worldwide experience justice problems that they cannot resolve. They are victims of unreported violence or crime, or they have civil or administrative justice problems that they cannot resolve. Meanwhile, a total of 5.1 billion people, representing two-thirds of the world's population, are currently considered to lack meaningful access to justice. This justice gap is both a reflection of and a contributor to structural inequalities, which most often affect individuals and collectives in disadvantaged situations, such as women, low-income persons, gender-diverse persons, or persons belonging to ethnic minorities (Task Force on Justice, 2019).

The persistence of this justice gap, which can be defined as a person's inability to obtain an effective, legally sound and actionable resolution to a dispute, makes the use of AI and its promise to expedite judicial processes, reduce persistent case backlogs, and assist judicial decision-making, a critical area of focus. However, transforming justice requires more than just deploying digital tools. The adoption of Open Justice principles by judicial institutions can support the cultural shift that they need to advance digital transformation in a people-centered and inclusive manner.

Open Justice is a vision that calls for transparency and accountability, making the workings of the justice system clear and accessible (Elena et al., 2019). It also promotes the publication of justice data in open formats, fostering informed public engagement and

enabling evidence-based innovation. Open Justice also calls for collaboration between justice institutions and stakeholders to drive digital transformation in a way that focuses on creating social value through more responsive and tailored processes that meet people's justice needs.

Open Justice provides tools for the development of responsible AI in justice, understood as the creation and deployment of AI systems that seek to ensure positive social impact, respect for people's rights, and minimization of bias and error, based on compliance with current legal standards and shared ethical principles (Adams, 2024).

23.3 The role of judicial data for delivering people-centered AI

Open Justice provides judicial institutions with a starting point for adopting mechanisms for transparency, accountability, and public oversight of the quality of the data they publish. This can include establishing publication priorities, participatory and collaborative mechanisms to protect the rights and interests of the people they serve throughout the data publication cycle, particularly by ensuring that the published data reflects the experience with justice of individuals and groups in vulnerable situations.

The availability and quality of judicial data are crucial for developing AI systems. Open Justice fosters a robust data ecosystem that promotes data publication and reuse, supporting advocacy, innovation, and the redesign of processes to better meet people's legal needs (World Justice Project, 2023). However, merely making data available is insufficient; publication policies must include safeguards to ensure data quality and integrity, as well as protect the privacy of individuals involved in judicial proceedings, particularly when AI systems rely on these data.

The importance of data in the development of AI systems is not new, but it becomes increasingly critical as their use becomes more ubiquitous (UNESCO, 2024a), while institutions strive to understand and regulate a technology whose evolution and scope have yet to be fully grasped. In 2018, the European Commission for the Efficiency of Justice (CEPEJ) presented ethical principles for

the use of AI in justice through its European Ethical Charter. This charter emphasized the importance of using high-quality, certified data to train AI systems, while maintaining the traceability of this data to prevent changes that could influence judicial decisions. It also underlined the need to address privacy concerns related to data used in the development of AI systems (CEPEJ, 2018). In line with their original position, in a more recent informative note the CEPEJ emphasizes the fact that any existing gaps or biases in judicial data can significantly impact the overall validity of AI-generated results, thus reducing the effectiveness and fairness of AI systems in the justice sector (CEPEJ, 2024).

The recently approved European AI Regulation is a significant step towards the establishment of a common global framework for the development of responsible AI, which has clear implications for its use in the field of justice (Regulation (EU) 2024/1689). This regulation emphasizes the clear impact that AI can, and most likely will, have on democracy, the rule of law, and individual rights. In particular, as a result of this regulation, justice AI enters a category considered as high-risk, which is therefore subject to stringent transparency, documentation, and oversight requirements.

The European AI regulation also highlights the importance of high-quality data and access to it as a means of structuring and ensuring the safe operation of AI systems, while avoiding becoming an additional source of social discrimination. To this end, it urges the establishment of appropriate management and governance practices for the data used in the context of AI systems, in order to achieve high quality datasets for training, validation and testing. In this regard, attention is drawn to biases that are considered to be inherent in the datasets used and, as such, may affect the outcomes of AI systems, thereby perpetuating and amplifying existing discrimination against certain vulnerable groups. To this end, it is established as a requirement that datasets be as complete and error-free as possible, which should not affect the use of techniques to protect the privacy of individuals.

While this regulation provides a starting point for global regulation on this topic, there have been significant advancements, particularly from Global South countries like Brazil, in addressing the availability and

quality of data for the development of AI. The National Justice Council (Conselho Nacional de Justiça, CNJ), the institution responsible for overseeing the administration of justice in the country, has been working for several years on establishing common standards for the publication of cases and documents, particularly aimed at facilitating their reuse in AI systems. As this paper is being written, the CNJ is proposing an enhanced regulation that reflects the data needs and risks arising from the increasing use of generative AI tools.

Meanwhile, from a universal standpoint, UNESCO is currently developing Guidelines for the use of AI systems in courts and tribunals. These guidelines will emphasize key principles for AI training data, such as transparency, quality, integrity, and data governance. Key aspects addressed by these Guidelines will include the need for robust data governance frameworks and infrastructures to protect personal data and promote responsible data-sharing practices, enhanced privacy protections, enhancing transparency regarding training data, and empowering deployers and users to effectively evaluate the quality and integrity of data (UNESCO, 2024b).

23.4 Data needs for a people-centered justice AI

Bias in legal data significantly influences the development of AI applications, potentially leading to unfairness or errors in prediction tasks, or biased information generation in question-answer tasks (Sargeant et al., 2024). For AI to work well, it needs a large volume of diverse and accurate data. Biases in AI systems can result from incomplete or unrepresentative data, leading to unfair outcomes and perpetuating existing disparities.

To address these challenges, it is critical to identify key needs for the availability and publication of judicial data, recognizing the transformative impact that AI systems can have on the delivery of justice, while also addressing various existing frameworks that come into play in the development of these systems, particularly when it comes to aligning them with the protection of fundamental rights, compliance with legal requirements, promotion of sustainability, and the maximization of the public interest (Belli et al., 2024). In this regard, it is essential to understand that the generation, classification, and use of these data must be conducted in a responsible and

inclusive manner, through transparent, accountable, and participatory mechanisms that emphasize achieving more people-centered justice by ensuring that people's rights and needs are respected and represented throughout the data lifecycle.

There are five main dimensions that are critical to the effective publication and use of justice data in the context of AI. These require action by justice institutions that seek to promote the genuine inclusion of all communities in shaping the development of a people-centered justice AI. The first aspect is **standardization**. It is vital to establish unified standards for data publication to ensure consistency and quality. Currently, judicial data is often published in an *ad-hoc* manner, which leads to inconsistencies and difficulties in using that data to inform system-wide innovations. Standardizing data formats and protocols can enhance the interoperability and reliability of judicial data, facilitating its use in AI systems, as well as in other applications.

The second aspect is **accessibility**. To be effectively used, judicial data must be easily accessible. This requires making data available through open, reusable formats, while ensuring that it can be integrated from multiple sources. Open judicial data portals allow for direct access, verification, and reuse of data. This aspect supports transparency and, therefore, improves the quality of AI systems by providing a reliable and traceable resource. However, balancing open access with privacy concerns remains a challenge that publication policies need to address.

Thirdly, the **completeness** of data is crucial for ensuring that AI systems can offer fair and equitable responses. In many justice systems, data often lacks representation of diverse groups, such as women, ethnic minorities, low-income persons, gender-diverse persons, or persons with disabilities. This underrepresentation can limit the effectiveness of AI systems and perpetuate existing inequalities. Efforts must be made to improve data collection and representation, guided by principles such as data equity, which emphasizes the need for inclusive data practices that respect human rights and promote fairness.

The fourth aspect is **privacy**, whose balance with data accessibility and completeness is often complex but necessary. As AI systems

increasingly rely on large datasets, it is essential to protect personal information while allowing for a meaningful and effective use of data. To achieve this balance, publishing institutions can resort to various measures, which should encompass the whole data publication cycle. These include conducting risk assessments to identify and mitigate privacy risks, applying data minimization principles to ensure that only the necessary data is collected and used, and maintaining ongoing human oversight to ensure that privacy concerns are continuously addressed and managed.

Finally, the aspect of **cybersecurity** clearly impacts the necessity to protect both the data and the systems used, as well as to address ethical and privacy issues related to data handling. While this dimension encompasses a broader context within the field of AI (i.e., focusing on the security of the systems themselves) the approach to data usage requires a holistic perspective. This perspective should encompass not only the protection of data in its initial dimensions (capture, storage, and management) but also risk controls surrounding the information security involved.

These five dimensions generate tensions and balances that must be addressed and discussed by judiciaries at two levels. One is the intra-systemic or primary level of justice institutions or systems. This first level can be dealt with through national or sectoral AI strategies or public policies, whereby the institution or system takes a position and combines its interests with those of the stakeholders in its direct sphere of influence. These initiatives should aim to guide the ethical and responsible development of AI systems in the judiciary through measures and approaches that could range from soft measures, such as the establishment of ethical guidelines or standards, to new regulations or legislation (OECD, 2024).

On the other hand, the traditional approach to judicial data governance, which focuses solely on the legal requirements within each specific jurisdiction (i.e., on the scope within which each institution carries out its jurisdictional work), is not compatible with the nature of AI system development. This development is inherently polycentric (Xue, 2024), and therefore transcends the boundaries of national jurisdictions. Consequently, it is essential for the above-mentioned four dimensions to also be discussed at a secondary, inter-systemic level involving

various actors, including legislative bodies, data-publishing institutions, companies that develop and use AI systems, and justice system users, at a global scale. In this regard, multi-stakeholder forums, such as the Internet Governance Forum's Dynamic Coalition for Artificial Intelligence, are working to connect and empower all populations, ensuring that AI systems are developed from a people-centered perspective and can therefore help them reap the benefits from digital transformation, in terms of closing the justice gap that prevent them from meaningfully accessing their rights (Belli et al., 2024).

23.5 Conclusion

To be people-centered, digital transformation processes in justice should not be limited to the adoption of new technologies. They must be supported by Open Justice policies that guide the necessary cultural and organizational shifts to mitigate digital exclusion and ensure that the benefits of digitalization reach all individuals equitably.

The role of data in developing responsible AI systems for justice must be addressed, as there is a key interrelation between these data, how their governance is conducted, and the need to mitigate biases and errors that can affect equity and justice in the application of AI tools within the judicial process.

Therefore, implementing AI systems in the justice sector requires a rigorous and well-structured approach to data publication and management. Based on these considerations, five fundamental dimensions should be addressed, each presenting challenges and opportunities that must be tackled jointly, both at the primary level by the institutions publishing the data and at the secondary level, given the polycentric nature of the AI value chain, by the various actors involved in the process.

Standardization, accessibility, completeness, privacy and cybersecurity are the main five dimensions identified that should be taken into consideration to govern judicial data, as a prerequisite to ensure a successful integration of AI in the provision of justice that focuses on peoples' needs and on closing the justice gap effectively. Addressing these five dimensions in a collaborative and open manner will be key to facilitate a digital transformation of justice that serves the justice needs of society as a whole.

23.6 References

Adams, R. (2024, May 27th). Responsible AI practices for business leaders (Episode 13). [Video podcast episode] *Unpacked.* Lab 45. https://lab45thinktank.com/podcast/responsible-ai-practices-for-business-leaders-with-dr-rachel-adams-ceo-gcg/.

Addo, P. M., Baumann, D., McMurren, J., Verhulst, S.G., Young, A. & Zahuranec, A.J. (2021). *Usages émergents des technologies au service du développement : un nouveau paradigme des intelligences.* Policy Paper. AFD. https://www.afd.fr/fr/ressources/technologies-developpement.

Belli, L. & Gaspar, W. (2024). AI Transparency, AI Accountability, and AI Sovereignty: An Overview. En L. Belli & W. Gaspar (eds.). *The Quest for AI Sovereignty, Transparency and Accountability.* Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition. FGV, 21-28. https://direitorio.fgv.br/en/publication/quest-ai-sovereignty-transparency-and-accountability.

CEPEJ (2024). *Use of Generative Artificial Intelligence (AI) by judicial professionals in a work-related context.* CEPEJ Working group on Cyberjustice and Artificial Intelligence. https://www.coe.int/en/web/cepej/-/information-note-on-the-use-of-generative-artificial-intelligence-ai-by-judicial-professionals-in-a-work-related-context.

CEPEJ (2018). *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment.* https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Datasphere Initiative (2024, June 25th). 6 reasons why Data matters for AI. *The Datasphere*. https://www.thedatasphere.org/news/6-reasons-why-data-matters-for-ai/.

Elena, S. & Mercado, J.G. (2019). A Theoretical Approach to Open Justice. In Elena, S. (coord). *Open Justice: An Innovation-Driven Agenda for Inclusive Societies.* SAIJ, 17-40. http://www.bibliotecadigital.gob.ar/items/show/2569.

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Muller, S. & Barendrecht, M. (2013). *The Justice Innovation Approach: How Justice Sector Leaders in Development Contexts Can Promote Innovation.* The World Bank Legal Review: Legal Innovation and Empowerment for Development, Vol. 4, 17-30. https://doi.org/10.1596/9780821395066_CH02.

OECD (2024). *Governing with Artificial Intelligence: Are Governments Ready?* OECD Artificial Intelligence Papers, No. 20. OECD Publishing. https://doi.org/10.1787/26324bc2-en.

OECD (2023). *Recommendation of the Council on Access to Justice and People-Centered Justice Systems* (OECD/LEGAL/0498). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0498.

World Justice Project (2023). *Disparities, Vulnerability, and Harnessing Data for People-Centered Justice.* WJP Justice Data Graphical Report II. https://worldjusticeproject.org/our-work/research-and-data/wjp-justice-data-graphical-report-ii.

Task Force on Justice (2019). *Justice for All. Final Report.* Center on International Cooperation. https://cic.nyu.edu/wp-content/uploads/2023/02/english_task_force_report_27jun19-min_compressed.pdf.

Sargeant, H. & Magnusson, M. (2024). *Bias in Legal Data for Generative AI.* Workshop on Generative AI and Law (GenLaw) at International Conference on Machine Learning 2024. https://blog.genlaw.org/pdfs/genlaw_icml2024/9.pdf.

UNESCO (2024a). *Survey on the Use of AI Systems by Judicial Operators.* UNESCO Global.

Judges' Initiative. https://unesdoc.unesco.org/ark:/48223/pf0000389786.

UNESCO (2024b). *Draft UNESCO Guidelines for the Use of AI Systems in Courts and Tribunals.* UNESCO Global Judges' Initiative. https://unesdoc.unesco.org/ark:/48223/pf0000390781.

Xue, J.H. (2024). Polycentric Theory Diffusion and AI Governance. In C. Aguerre, M. Campbell-Verduyn & J.A. Scholte (eds.), *Global Digital Data Governance. Polycentric Perspectives.* London: Routledge, 223-237. https://www.researchgate.net/publication/377135755_Polycentric_Theory_Diffusion_and_AI_Governance.

24 Fostering AI Research And Development: Towards A Trustworthy LLM. Mitigating Compliance Risks Illustrated via Scenarios

Liisa Janssens, Saskia Lensink and Laura Middeldorp

Abstract

The rapid growth of Large Language Models (LLMs) challenges the Rule of Law, necessitating a thorough examination of their disruptive potential. This paper highlights the importance of adhering to these principles for responsible LLM deployment. Using a scenario-based approach, we show how specific design choices can lead to unintended consequences. We present a hypothetical case of developing an LLM, focusing on the inclusion of an opt-out option for personal data removal. Two scenarios are explored: one with and one without this option, illustrating how this decision impacts compliance with the Rule of Law. The paper emphasizes anticipating regulatory requirements and linking design choices to legal principles during research and development. By addressing these considerations early, stakeholders can better prepare for legislative changes and mitigate compliance risks. This paper aims to guide end-users, policymakers, researchers, and industry participants on mitigating risks and ensuring responsible LLM deployment.

Keywords: Large Language Models, Design-choices, Opt-out option, Global Majority, Compliance, Rule of Law, Research and Development, Deployment.

Introduction

The European Commission published the AI Act in the Official Journal on the 12th of July 2024[72]: a legal framework guiding the development and deployment of AI systems. The AI Act aims to uphold the values of the European Union and at the same time leverage the capabilities of AI. Further developments on the AI

72 "Today, on July 12, 2024, EU Regulation No. 1689/2024 laying down harmonized rules on Artificial Intelligence ("Regulation" or "AI Act") was finally published in the EU Official Journal and will enter into force on August 1, 2024.".

Act will make compliance a moving target. This is in the nature of (new) laws and regulations, since these need to be understandable, transparent and trustworthy but are not supposed to be set in solid bedrock. Interpretations of the meaning of the AI Act via case law (jurisprudence) is yet to commence, and this can lead to questions how to innovate with AI with the aim to deploy these new innovations. The extra complicating factor lies in the fact that AI is a moving target as well. This combination creates one of the biggest challenges for all parties who want to deploy AI aligned with laws and regulations. The nature of law complicates early-stage research initiatives which strive for, or promise, alignment with the AI Act when it is time for deployment of these models. The aim of the AI Act is not to stifle innovations, it asks for a forward-looking eye: the ability to foresee what it takes to become compliant when the time has come to deploy, in the legal reality, what has been made.

The question that becomes relevant to all (from developers to end-users) is how to deal with the moving target of compliance in the research and development process of AI models? Mitigating future compliance issues with all the ins and outs of the AI Act is difficult, but this does not mean that future compliance issues cannot be scoped, addressed and tried to be mitigated.

First, the AI Act can be seen as a protection of the Rule of Law: *"It aims to protect fundamental rights, democracy, the rule of law and environmental sustainability from high-risk AI, while boosting innovation and establishing Europe as a leader in the field. The regulation establishes obligations for AI based on its potential risks and level of impact."* (European Parliament 2024). In this paper the lens of the Rule of Law will be presented as a lens to scope future compliance issues with the AI Act. The Rule of Law also allows for the review of the legal effect of specific design choices in LLMs from a more fundamental perspective.

An informed viewpoint about how future compliance risks could manifest can be provided via a scenario-based approach (Janssens, Lucassen, Middeldorp, Lobbezoo, & Schoenmakers, July 2024). Via this approach insightful perspectives are given which can inform decision-makers about what is at stake and what could be the best design choices in order to comply with the ambition of deploying

an LLM for the public good. Part of the scenario-based approach is the lens of the Rule of Law (Stein, 2019).[73] In this paper we use this approach to analyse potential norm violations of the Rule of Law via a hypothetical use case: an LLM which is built from scratch. This use case takes as point of departure that it is necessary to gain control, as much as possible, over the data used to train an LLM. We will refer to this hypothetical LLM as 'LLM-from-scratch'.

A design choice that can be made by the development team of the LLM-from-scratch is whether an opt-out option before training needs to be included. Typically, training data of LLMs are curated beforehand by using algorithms that remove personal identifiable information. The performance of these algorithms, however, is not perfect and personal information can still reside within the training dataset after curation. Therefore, in addition to algorithmic means to remove personal identifiable information, one could opt for including an 'opt-out' option that allows the public to check the curated dataset for the presence of any of their personal information and request that this information is taken out of the dataset before it is used to train the LLM.

This paper investigates how the design choice of the implementation of an opt-out option within the development phase of an LLM is related to the protection of the Rule of Law and therewith fostering the well-being of the global majority. Two types of scenarios are investigated: one where an opt-out option is implemented within the development phase and one where the opt-out option is not present. For both scenarios, we will evaluate the hypothetical legal effect of the inclusion/exclusion of an opt-out option and how this affects the tenets of the Rule of Law. For instance, how does the design choice of presenting an opt-out option prior to training the model impact norms like respecting justice, legitimacy and transparency?

73 "The Rule of Law refers to a principle of governance in which all persons, institutions and entities, public and private, including the State itself, are accountable to laws that are publicly promulgated, equally enforced and independently adjudicated, and which are consistent with international human rights norms and standards. It requires, as well, measures to ensure adherence to the principles of supremacy of law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness and procedural and legal transparency." Robert Stein, What Exactly Is the Rule of Law? 2019, p. 188.

In addition, what could be possible consequences if an LLM built from scratch does not implement an opt-out option before training?

This paper is structured as follows: Section 2 gives an overview of the opt-out option, the Rule of Law and explains the scenario-based method. Section 3 describes the analysis performed on the two scenarios, one where an opt-out option is implemented and one where an opt-out option is not implemented and investigates how the decision of including or excluding an opt-out option affects the tenets of the Rule of Law. We conclude this paper with Section 4 by providing recommendations which can be inspiring for everyone who is planning to develop and deploy LLMs and wants to deal with future compliance issues. The recommendations in this paper can be used by decision-makers to make an informed decision on incorporating an opt-out option in the development phase of an LLM.

24.1 Development

In this section the relation between the design choice made during development of the opt-out option and the Rule of Law is explained. Although the focus in this article is set on the Rule of Law as one of the foundational principles of the European Union (Article 2 Treaty on the European Union), the Rule of Law is also one of the foundational principles of the United Nations. The Rule of Law is meant to foster well-being of all human beings, amongst who the global majority is part, as can be read in the clarification of the United Nations.[74]

The United Nations clarifies the Rule of Law as follows:

> *"It requires measures to ensure adherence to the principles of supremacy of the law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness, and procedural and legal transparency." (United Nations, 2024)*

74 "Rule of law issues includes emerging and critical issues such as the proliferation of hate speech and incitement to violence; preventing radicalization/violent extremism; climate change and the environment impacting on the security and livelihoods of people; and the complexities of artificial intelligence and cybercrime." (United Nations, 2024).

Therefore, our assessment of an LLM in the EU context can be an example how the tenets of the Rule of Law (accountability, transparency, liability and contestability) can be globally applicable to foster the well-being of the global majority. Societies all over the world can learn from this hypothetical use case in shaping their own rules, regulations and policies around the development of LLMs to foster and ensure that the well-being of people, i.e. the global majority, is maintained. In case an opt-out option is implemented, it is of importance that the datasets used to train the LLM are open and can be searched by everyone in the world that may be present in the datasets. Access to the internet is a prerequisite to make this possible. However, since not everyone has access to the internet (Bradshaw, 2001) it is desirable -when the datasets contain personal information of people from countries who have no or limited access to the internet- to take auxiliary precautions. For example, independent institutions may be asked to perform a check for persons who are unable to do this.

24.1.1 The Opt-Out Option and the Rule of Law

Good governance is about accountability, transparency, (addressing) liability and contestability. The aim of the mechanisms of the Rule of Law is to produce a government that is legitimate and effective. Good governance is about legitimate, accountable and effective ways of obtaining and using public power and resources in the pursuit of legitimate goals.

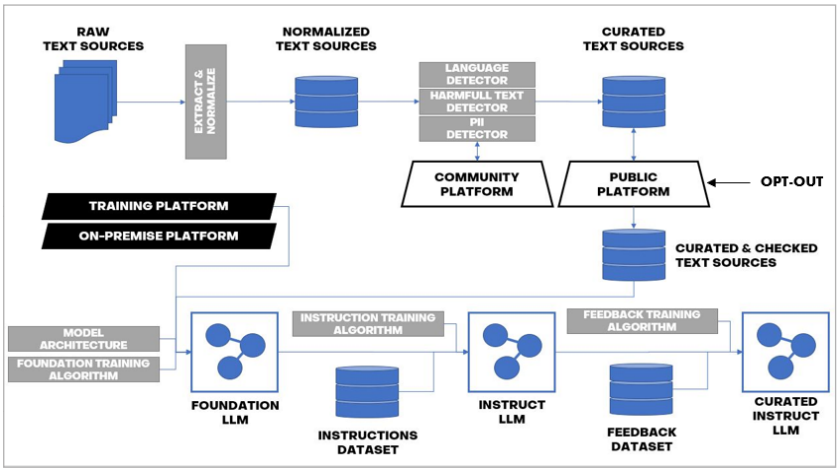
When a government has the ambition to use an LLM for the public good this model needs to foster good governance. The opt-out option is an example of a design choice which underpins this ambition.

24.1.2 The Opt-Out Option

Figure 1 illustrates the pipeline of the LLM-from-scratch. The opt-out option takes place via a public platform prior to training the model. Due to privacy legislations and agreements made with the contributors of the data sources for LLM-from-scratch, the entire database cannot be made accessible to the general public. Instead, the database can be searched by the public to examine whether their personal information is present in the database. In other words, the database is not made available but is available to be searched.

The documents checked for opt-out will be removed from the database.

Figure 1 Pipeline of the LLM-from-scratch



24.1.3 Scenario-Based Method

LLMs can be beneficial but at the same time bring about (unintended) drawbacks that can challenge the Rule of Law. There is a need for a method that can be used to identify the tension between the Rule of Law and the consequences of design choices in the development of LLMs.

We have developed a method that identifies the tensions between the Rule of Law and emerging and disruptive technologies, of which an LLM is an example, by means of a scenario analysis. Scenarios are a useful instrument to simulate a specific environment through which an LLM can be deployed. By mapping the events in the scenario to (tenets of) the Rule of Law, advice, which is informed by European norms and values, can be shaped regarding design choices of an LLM on both technical and functional level. In addition, a scenario provides a contextualization of how the LLM will operate in practice to identify possible norm violations which need to be mitigated.

The opt-out option within the LLM-from-scratch pipeline can be regarded as a technical requirement. Is it necessary that an opt-out option is incorporated before training the model? And

what are possible consequences of (not) incorporating an opt-out option? To investigate this, two scenarios will be analysed, one without and one with an opt-out option implemented. In each scenario, we will map the tenets of the Rule of Law to the events and consequences happening in the operational context of the LLM-from-scratch. Furthermore, both scenarios highlight the potential benefits and risks to maintaining the Rule of Law.

24.2 Discussion

24.2.1 Scenario Analysis: Opt-Out Option in two scenarios

This section analyses the role of an opt-out option and the effect it has on the tenets of the Rule of Law by means of scenarios. Before the scenarios are introduced, remarks on the data curation and the relationship with an opt-out option are presented.

LLMs are typically trained on a large amount of textual data, where the data are curated before training using curation algorithms. These algorithms remove harmful texts and personal identifiable information as displayed in **Figure 1.** The accuracy rate of curation algorithms is rather good, between 90-95% (Dasgupta, Ganesan, Kannan, Reinwald, & Kumar, 2018), however given a population of, say, X million, a personal detection rate of 5-10% can still be substantive. Even when only one case ‘slips through’ dangers can already arise. This one case can erode legal certainty on individual level, as well as -when this is widely spread via media- on a societal level. A case like that can have a big impact on the trust of potential end users in the LLM.

The next two sub-sections present the scenarios and showcase how the tenets of the Rule of Law may be challenged.

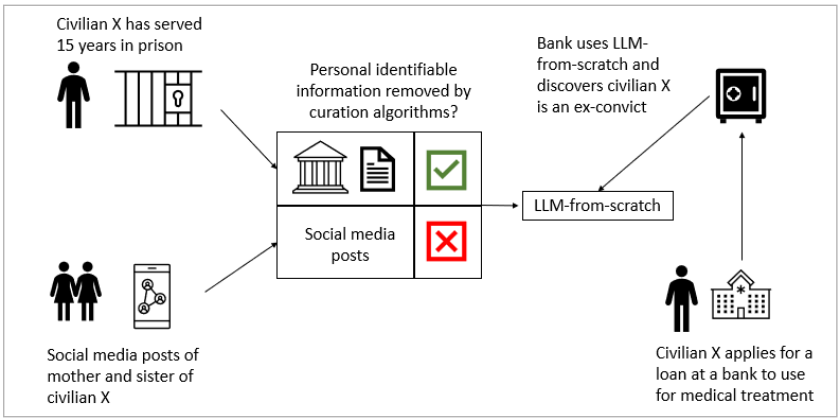
24.2.1.1 Scenario One: No Opt-Out Option Implemented

Figure 2 visualizes the scenario where no opt-out option has been implemented. Civilian X has served 15 years in prison. The LLM-from-scratch uses case law describing the timeline of the crime and the verdict of civilian X to train the model. The mother and sister of civilian X post on social media they are happy he has been released from prison. These social media posts are also included

in the training dataset. The curation algorithms manage to remove personal information of civilian X from case law, but not from the social media posts of the mother and sister.

Civilian X has recently been diagnosed with a rare form of cancer which needs expensive treatment. Civilian X applies for a loan for the treatment at the bank which conducts a background check on civilian X using the LLM-from-scratch. The bank discovers civilian X is a convicted murderer who has been in prison for 15 years. The bank is doubting: should they accept the loan or refuse the loan because he/she has been a person who was convicted for a crime? Is the person eligible to obtain a loan?

Figure 2 Visualisation of scenario where no opt-out option has been implemented



In this scenario, the presence of the personal data of civilian X in the LLM-from-scratch has a negative effect on acquiring a loan.[75] **Figure 3** analyses how the events in the scenario are challenging the tenets of the Rule of Law.

75 Other areas where a negative impact may be found are being rejected for a job application since you are being a family member of a (ex)-criminal or not getting a rental apartment/mortgage because you are connected to a (ex)-convicted killer.

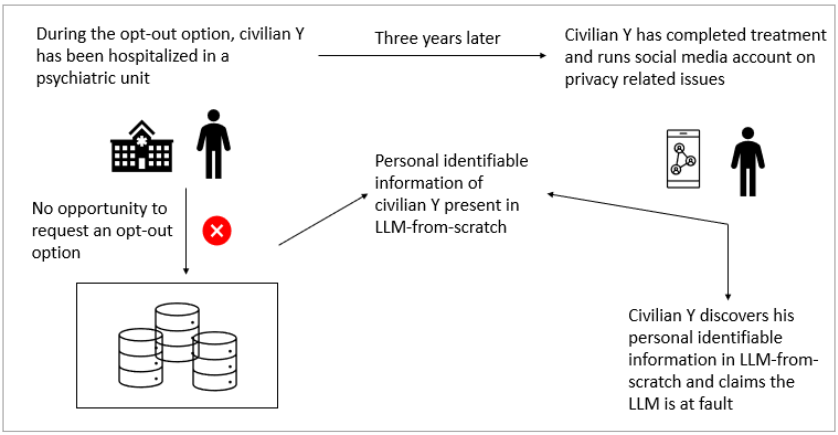
Figure 3 Analysis Of Scenario One: No Opt-Out Option Implemented

Tenet	
Accountability	When it comes to the tenet **accountability**, the bank has to make decisions which may sort legal effects that are based on (juridical) facts. When this decision, as stated in the scenario description, is made without a legal basis this can lead to unjust situations such as not approving a loan which -if there is no other ground on which this disapproval stands- is unjust. The bank's decision to not grant a loan to the ex-convicted must be based on a legal basis. If this is not the case, it is illegal which is contradictory to the tenet of legitimacy.
Transparency	It can be questioned if the output given by the LLM-from-scratch provides factual information and whether it is **transparent** how the data is collected, curated and trained. The decision to make use of data of court cases is an important decision in itself whose result should be clearly communicated to the general public to mitigate risks of people being unaware that their court cases could be included within the model.
Contestability	**Contestability** implies that the ex-convicted should be given the possibility to contest the decision which is made by the bank, also in front of court. Therefore, it is of importance that, if the bank uses the LLM-from-scratch to base their decision upon, the bank is transparent about the fact it is using a LLM and also provides context on how it has been used. For example, what is the prompt that they used?
Liability	If the LLM-from-scratch is not transparent about the fact that it has used court cases to train the model and the bank is unaware that this type of data has been incorporated, the tenet **liability** is challenged. If the bank does not know that court cases have been included, the question raises if they can be held liable for errors, mistakes or damages that are consequences of their (wrongful) decisions. Or is the LLM-from-scratch liable?

24.2.1.2 Scenario Two: The Opt-Out Option Is Implemented

Figure 4 visualizes the scenario where an opt-out option has been implemented. We focus on civilian Y, who has been hospitalized in a psychiatric unit during the opt-out option and therefore did not have the opportunity to request an opt-out option. Three years later, civilian Y has completed treatment and uses social media to create awareness on privacy issues. Civilian Y discovers that his personal identifiable information is present in the LLM and decides to claim the LLM team. The question raises if the team developing the LLM is responsible for removing personal information of persons who did not have had the opportunity (e.g. due to mental health issues, age or disabilities) to make use of the opt-out option.

Figure 4 Visualisation of scenario where an opt-out option has been implemented



Regardless of the answer to the question whether the LLM team is responsible, the tenets of the Rule of Law are challenged as reflected in **Figure 5**.

Figure 5 Analysis Of Scenario Two: Opt-Out Option Is Implemented

Tenet	
Transparency	It is important to create awareness and **transparency** amongst both civilians and family members or caretakers of incapacitated persons about the possibilities of the opt-out option. A possible solution to deal with the situation of incapacitated persons could be to provide an indirect opt-out option to family members or caretakers of the incapacitated persons. The family member or caretaker can access the database for the incapacitated person and request an opt-out option in his/her name. The opt-out option is thus requested indirectly via relatives of the incapacitated person. However, how can it be ensured that the incapacitated person gives consent to the family members or caretakers to perform such an action? If an opt-out option is implemented, the team building the LLM-from-scratch needs to carefully consider how the opt-out option should be implemented and presented to the public. **Transparency** plays an important role in this. It is important that LLM-from-scratch is transparent about all the architectural choices made, including the decisions made on the implementation of the opt-out option. Furthermore, policies and communication strategies can be used to inform the public about the opt-out option. The logging of the architectural choices including reasoning why each choice is made gives LLM-from-scratch the opportunity to inform the public and meet their expectations.
Accountability and liability	Transparency in the architectural choices also contributes to the other tenets of the Rule of Law: a transparent way of working is necessary to give the powers (legislative, executive and judicial power), who have legitimatized decision power the ability to check if the architectural decisions are **legitimate**. This is also important when you take **accountability** questions into account: without the possibility to address responsibility about the architectural choices made, the accountability cannot be addressed. This can lead to problems when errors occur which raises **liability** issues and it becomes important to contest if the architectural choice made have led to these errors and therewith liabilities. Another important question is linked to being future proof: how can it be ensured that the opt-out option is made future proof as in e.g. compliance with the AI Act and related legislation such as the GDPR and connected jurisprudence? It is currently not a legal obligation to provide an opt-out option to the public before training a LLM. Moreover, if one decides to implement an opt-out option, no guidelines are given on how to do this. Beyond current questions connected to how compliance is conceived at this moment and time, it can be helpful to take the tenets of the Rule of Law when shaping the architecture choices, such as the opt-out option, to strive for future proof compliancy. It is therefore important to closely monitor the AI Act and related legislation for possible changes.

24.3 Conclusion

In this paper we have investigated how the design choice of incorporating an opt-out option during the development of a hypothetical LLM, LLM-from-scratch, can contribute to a fair deployment of the LLM and how the tenets of the Rule of Law may be challenged by this decision. The scenario analysis aids in making an informed decision whether an opt-out option should be included to ensure a fair and just deployment of an LLM. The scenario analysis has resulted in the following recommendations:

In the case that an opt-out option has not been implemented:

- If court cases or other sensitive data are included in the training phase of an LLM, this should be made transparent and clearly communicated to the users of that LLM. As the public did not receive an opt-out option, the curated data still (potentially) contains personal identifiable information which may cause negative outcomes for individuals in society.
- The developers of the LLM should be aware that claims could follow from individuals who feel discriminated or disadvantaged by the LLM.

In the case that an opt-out option has been implemented:

- Providing an opt-out option to the public before training the model can reduce the possible chance of injustice in legal effects. This option can also provide a form of human oversight via the check of the public on the dataset before training. It can even be seen as a form of citizen participation.
- Access to the internet is an important enabler to make a check on the data and opt-out option possible. If the data contains personal information of people who have less accessibility to the internet, the data could be (double) checked by independent institutions to foster the well-being of the global majority.
- It is important that the LLM developers investigate how to deal with incapacitated persons in the opt-out option. For example, by exploring the possibility to provide an indirect opt-out option to family members or caretakers of incapacitated persons. The design of such an indirect opt-out option should be carefully

thought through and can be quite challenging since incapacitated persons may be unable to give their consent.

- Developers of LLMs should be aware that, in the case that they do not provide a special opt-out option for incapacitated persons, they can receive claims from persons who, at the time of the opt-out option, were incapacitated.
- The workings of the opt-out option should be transparent. In order to achieve this, the design choices made regarding the opt-out option and the implementation of the opt-out option should be clearly logged and documented.

24.4 References

Bradshaw, A. C. (2001). Internet users worldwide. Educational Technology Research and Development, 111-117.

Dasgupta, R., Ganesan, B., Kannan, A., Reinwald, B., & Kumar, A. (2018). Fine grained classification of personal data entities. arXiv preprint arXiv:1811.09368.

European Parliament, 13th of March 2024 — 12:25, Official Press Release, Artificial Intelligence Act: MEPs adopt landmark law.

Janssens, L., Lucassen, O., Middeldorp, L., Lobbezoo, L., & Schoenmakers, O. (July 2024). Responsible AI and the Rule of Law. TNO.

NATO. Retrieved August 2023, accessed at https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.

Stein, R. A. (2019). What exactly is the rule of law. Hous. L. Rev., 57, 185.

United Nations, (2024), United Nations and the Rule of Law. What is the Rule of Law, Retrieved October 2024, from https://www.un.org/ruleoflaw/what-is-the-rule-of-law/#:~:text=It%20requires%20measures%20to%20ensure,and%20procedural%20and%20legal%20transparency.

Legislation

Article 2 of the Treaty on European Union: *"The Union is founded on the values of respect*

for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."

AI Act Draft Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 — C9 0146/2021 — 2021/0106(COD))

Appendix A

List of Key Definitions

AI	Artificial Intelligence
AI Act	European legislation on harmonised rules on Artificial Intelligence
EDT	Emerging Disruptive Technology
LLM	Large Language Model
NATO	North Atlantic Treaty Organization
OSINT	Open Source Intelligence
Requirements	Technical requirements which can become tools of good governance Rule of Law. Is shaped by various sources, such as: case law; legal doctrine; legal interpretation methods; positive law; rules and regulations; draft rules and regulations and legal theory.
AI Technology	AI technologies and/or systems with applied AI applications
Rule of Law tenets	accountability, transparency, contestability mechanisms, processes of rules and regulations; case law; policies; etc.

25 Addressing Gender Data Gaps in the Global Majority: Opportunities and Challenges of Synthetic Data

Ronald Musizvingoza

Abstract

This paper explores pervasive gender data gaps affecting the global majority, highlighting their negative impact on health, particularly among women and girls. When these data gaps persist, the rapid use of AI can exacerbate existing inequalities by failing to fully incorporate the global majority's experiences. We argue that synthetic data can be a powerful tool for addressing these gaps by generating representative datasets that reflect diverse gender experiences. While acknowledging the risks associated with synthetic data, including potential biases and cybersecurity threats, this paper emphasises the need for robust methodologies, ethical frameworks and guidelines to ensure its responsible use. By integrating real-world data and fostering collaboration with gender experts, we advocate for a multifaceted approach to AI development that prioritises gender equality. Ultimately, we call for policies that promote inclusive research and data practices, ensuring synthetic data contributes to equitable health outcomes for the global majority.

Keywords: AI, Synthetic Data, Gender Data Gaps, Health, Global Majority.

Introduction

Gender data gaps, or the lack of data on diverse gender experiences, are widespread and disproportionately impact people from the global majority(Musizvingoza & Lopes, 2022). Women comprise half the world's population, but data on their status, health, and well-being is lacking. Closing these gaps is crucial to reflect the global majority's experiences and needs, especially within artificial intelligence(AI) (Musizvingoza, 2024). Despite global commitments to Sustainable Development Goal 5, only 48% of the necessary data to assess progress is available. With a 3% annual growth rate, collecting all

required gender-specific data will take 22 years, missing the 2030 deadline by over a decade(Encarnacion, Emandi, & Seck, 2022).

Gender data gaps perpetuate inequalities in education, work, and healthcare by failing to support effective programs and overlooking marginalised groups(Paris21, 2024). AI tools can worsen this issue by embedding and amplifying gender biases if not trained on comprehensive, representative data(O'Connor & Liu, 2023). Since these tools learn from their training data, excluding or misrepresenting the global majority, especially women can have serious consequences especially in healthcare. For example, AI tools for liver disease screening were found to be less accurate for women(Straw & Wu, 2022), and delays in diagnoses for Black patients have been linked to biased datasets(Williams, 2023). Additionally, AI in judicial sentencing were found to be discriminatory towards black offenders(Lippert-Rasmussen, 2022).

One potential solution to address gender data gaps is using artificially generated synthetic data to mimic the original dataset's characteristics or meet predetermined criteria(Deng, 2023; Marwala, Fournier-Tombs, & Stinckwich, 2023). By simulating the properties of original datasets, synthetic data can be pivotal for training AI tools, especially in contexts where data is sensitive, scarce, or biased. This paper examines the risks and opportunities associated with using synthetic data to address gender data gaps in health, particularly from the perspective of global majority populations frequently underrepresented in AI and data governance discussions. It will provide insights into how synthetic data can address gender data gaps and enable gender and health equity.

25.1 Discussion

25.1.1 Gender Data Gaps

Gender data gaps are globally prevalent, particularly in developing countries and regions such as Southeast Asia, Latin America, Sub-Saharan Africa, and the Pacific, impacting millions of vulnerable women and girls(Kathleen Grantham, 2020). Despite commitments from 193 countries to the 2030 Agenda, comprehensive data on gender-specific SDG indicators is still lacking(Encarnacion et al.,

2022). These gaps are especially pronounced in healthcare, with only 21.8% of gender health indicators available in 2023(World Bank, 2024). This lack of data hampers effective public health responses, particularly in the global majority, especially in African countries(Adebisi & Lucero-Prisno, 2022). For example, during the COVID-19 pandemic, 76% of high-income countries reported COVID-19 case data by sex, compared to only 37% of low-income countries (Hawkes et al., 2021).

AI development and decision-making are controlled mainly by the Global North, particularly North America(Anthony, Sharma, & Noor, 2024). Nevertheless, the impact of AI on the global majority is substantial(Norori, Hu, Aellen, Faraci, & Tzovara, 2021). AI models are often trained on data generated online, which excludes experiences from connectivity-limited environments, making the models unrepresentative of the global context. With 2.6 billion people offline, representing 37% of the global population, this issue is further exacerbated in developing countries, where 96% of those offline reside and where, on average, 21% of women have internet access compared to 32% of men(ITU, 2022).

Moreover, AI models trained on biased data amplify gender bias. Medical studies have historically excluded female participants, leading to research data collected from males being generalised to females(Merone, Tsey, Russell, & Nagle, 2022). A significant portion of the datasets used for training AI algorithms in healthcare is derived from such biased research, resulting in persistent gender bias. This gap has far-reaching implications for healthcare, particularly in disease prevention, diagnosis, and treatment(di Lego, 2023; Norori et al., 2021). To address these issues, the World Health Organization (WHO) AI for health guidance(World Health Organization, 2021) highlights the importance of ethical, legal, and human rights considerations, focusing on data governance, algorithmic transparency, inclusiveness, equity, and accountability(Lopes, Saitabau, Rustagi, & Khosla, 2023).

25.1.2 Synthetic Data

Synthetic data can address imbalances and underrepresentation, helping fill gender data gaps in AI model training(Deng, 2023). It helps overcome data scarcity, sensitivity, and bias challenges by providing

a flexible and safe alternative to real-world data(Deng, 2023). For example, in healthcare, synthetic data is used as a proxy for real data to support medical research while ensuring confidentiality(Giuffrè & Shung, 2023; Gonzales, Guruswamy, & Smith, 2023; Kokosi & Harron, 2022; Laderas et al., 2017; Reiner Benaim et al., 2020). Other notable uses of synthetic data include estimating the benefits of healthcare policies and interventions, pre-training models for specific patient populations, and improving public health models for predicting disease outbreaks (Giuffrè & Shung, 2023; Gonzales et al., 2023; Kokosi & Harron, 2022; Laderas et al., 2017; Reiner Benaim et al., 2020). Furthermore, synthetic data supports the creation of digital twins, simulating real-time behaviour, including gender-specific health patterns(Giuffrè & Shung, 2023).

Synthetic data's key features — privacy preservation(Gonzales et al., 2023; James, Harbron, Branson, & Sundler, 2021; Tiwald, Ebert, & Soukup, 2021), scalability(Almirall et al., 2022), realistic generation(Dahmen & Cook, 2019), representativeness(James et al., 2021; Tiwald et al., 2021), and reproducibility — are critical in addressing gender data gaps. Synthetic data helps balance datasets, augment limited data, and generate high-dimensional data, improving reliability for accurate gender analysis(Juwara, El-Hussuna, & El Emam, 2024). The UZIMA-DS project in Kenya exemplifies the use of AI-ready synthetic datasets to create early warning systems, addressing data gaps while promoting open access in health research(Thuku, Baker, Mwigereri, Waljee, & Siwo, 2024). Another example is the World Bank's Synthetic Data for an Imaginary Country, a hierarchical simulation and training dataset covering demographic, education, and health variables(World Bank, 2023). These examples highlight how synthetic data can be used to close gender data gaps in health by enhancing the representation of underrepresented groups, improving access to gender-specific data, enabling more accurate simulations, and facilitating broader sharing of gender-sensitive information, particularly in resource-limited settings, leading to more equitable health interventions and research. While synthetic data offers opportunities for training AI models, it risks oversimplifying complex human experiences, perpetuating biases, and neglecting the realities of the global majority, underscoring the need for ethically grounded approaches that integrate real-world data.

25.1.3 Methods for Generating Synthetic Data

Methods for synthetic data generation can be categorised into statistical and probabilistic approaches, machine learning techniques(ML), and deep learning methodologies(DL) (Hernandez, Epelde, Alberdi, Cilla, & Rankin, 2022). Statistical methods generate synthetic data by sampling from existing datasets(Kaur et al., 2021; Pourshahrokhi, Kouchaki, Kober, Miaskowski, & Barnaghi, 2021). ML approaches use models like decision trees and regression to generate new data points that mimic the statistical properties of the original data, especially when real data is scarce or sensitive(Gonzales et al., 2023; Lu et al., 2023). DL-based methods use neural networks to generate synthetic data(Achuthan et al., 2022; Mohamed & Frank, 2024; Nikolentzos, Vazirgiannis, Xypolopoulos, Lingman, & Brandt, 2023). In healthcare, synthetic data generation techniques produce valuable datasets, including synthetic patient records for research and analysis(Nikolentzos et al., 2023), synthetic time-series health records to capture dynamic patient health trajectories(Li, Cairns, Li, & Zhu, 2023), realistic medical images such as MRI and CT scans for model training(Skandarani, Jodoin, & Lalande, 2023), and simulations of imbalanced clinical variables in HIV antiretroviral therapy datasets(Giuffrè & Shung, 2023; Kuo et al., 2023). These datasets can help address gender data gaps by providing more comprehensive and representative data for gender analysis, enabling more inclusive and effective healthcare solutions.

These synthetic data generation techniques can address gender data gaps by reflecting real-world diversity, even when actual data is limited or biased. They enhance fairness, equity, privacy, and inclusivity — key elements of gender data — while improving representation, intersectionality, and bias mitigation in underrepresented groups, thus fostering more accurate gender analysis. For instance, DL-based methods can generate synthetic data that closely mimics real-world diversity(Ali et al., 2022) and balance gender-skewed datasets for more representative outcomes(Makhlouf, Maayah, Abughanam, & Catal, 2023). Statistical and probabilistic approaches can be used to enhance data privacy — a critical aspect of gender data while maintaining the statistical properties of original datasets and protecting sensitive information(Skandarani et al., 2023). When

gender-specific data is scarce, synthetic data can augment the dataset, providing richer insights and improving analysis for underrepresented genders(Motamed, Rogalla, & Khalvati, 2021; Yu et al., 2020). Additionally, synthetic data can create high-dimensional datasets that capture complex relationships among variables, including gender, enabling sophisticated analyses(Sun, van Soest, & Dumontier, 2023). These techniques can also identify and mitigate biases in real-world datasets by balancing underrepresented gender categories and promoting equitable representation(Paproki, Salvado, & Fookes, 2024; van Breugel, Kyono, Berrevoets, & van der Schaar, 2021). Ensuring privacy, fairness, and equity is crucial for comprehensive gender data. Synthetic data can support these principles by generating diverse datasets that reflect real-world complexities, ensuring demographic parity, and improving the accuracy and fairness of analyses and decisions(Rajabi & Garibay, 2021).

25.1.4 Challenges

Synthetic data in healthcare poses challenges, with risks of bias amplification, low interpretability, and insufficient methods for auditing data quality(Giuffrè & Shung, 2023). Furthermore, synthetic data entails multifaceted risks, including cybersecurity threats, model inaccuracies, data integrity, misuse, intellectual property infringement, and data contamination, which can exacerbate gender data gaps if not adequately addressed(Marwala et al., 2023). For example, cybersecurity threats could lead to the exposure of sensitive data related to underrepresented gender groups. At the same time, model inaccuracies might perpetuate existing biases by generating flawed or incomplete gender data. Data misuse and contamination can distort gender representation in datasets, further skewing analysis. Additionally, intellectual property infringement could limit access to diverse datasets, thus hindering efforts to ensure fairness, inclusivity, and equitable representation of gender data. For instance, a synthetic dataset based on facial images of predominantly men or a specific racial group will reflect this imbalance if not addressed, perpetuating gender biases(Hao et al., 2024).

The success of synthetic data in healthcare depends on ensuring diversity, transparency, bias mitigation, privacy, fairness, and robust

evaluation to advance AI responsibly and represent underrepresented genders(Gonzales et al., 2023). Countries like Singapore have developed guidelines to harness synthetic data's potential while balancing utility and protection risks by defining its purpose, preparing data thoughtfully, following best practices, and managing re-identification risks(Personal Data Protection Comission, 2024). Similarly, the United Nations University recommends ethically using synthetic data in AI, emphasising diverse data sources, various generative models, transparency, and quality metrics(de Wilde et al., 2023). These guidelines help close gender data gaps and promote inclusive AI solutions by ensuring synthetic data is designed to benefit diverse populations, particularly in the global majority.

25.2 Conclusions

In conclusion, while synthetic data offers significant promise for addressing gender data gaps and promoting equitable AI, its application must be cautious due to risks like bias amplification and misuse. The methods for generating synthetic data provide potential solutions by enabling the creation of more representative datasets and mitigating biases that disproportionately affect underrepresented groups, particularly the global majority. Since AI reflects real-world gender biases, addressing these is key for equitable AI. We recommend comprehensive gender data collection, inclusive definitions, well-trained data collectors, collaboration with gender experts, and adopting ethical guidelines and best practices widely recognised in data governance to ensure the responsible use of synthetic data. Furthermore, we recommend supplementing synthetic datasets with real-world gender data to ensure a more accurate and inclusive portrayal of people from the global majority. Prioritising gender equality in AI development, evaluating data for bias, diversifying teams, and enforcing ethical guidelines will mitigate potential biases. The Global Digital Compact is a critical opportunity to embed gender perspectives into digital governance. Without such efforts, AI may widen existing gender gaps. Future research should refine synthetic data techniques, develop auditing frameworks, and address socioeconomic factors to capture gender disparities. Policy recommendations should prioritise the implementation of ethical

guidelines for synthetic data usage, foster transparency in data practices, and promote inclusive research agendas that consider the needs and experiences of the global majority to ensure that synthetic data contributes to meaningful and equitable outcomes.

25.3 References

Achuthan, S., Chatterjee, R., Kotnala, S., Mohanty, A., Bhattacharya, S., Salgia, R., & Kulkarni, P. (2022). Leveraging deep learning algorithms for synthetic data generation to design and analyze biological networks. *J. Biosci.*, *47*.

Adebisi, Y. A., & Lucero-Prisno, D. E. 3rd. (2022). Fixing Data Gaps for Population Health in Africa: An Urgent Need. *Int. J. Public Health*, *67*, 1605418. https://doi.org/10.3389/ijph.2022.1605418.

Ali, H., Biswas, Md. R., Mohsen, F., Shah, U., Alamgir, A., Mousa, O., & Shah, Z. (2022). The role of generative adversarial networks in brain MRI: a scoping review. *Insights Imaging*, *13*(1), 98. https://doi.org/10.1186/s13244-022-01237-0.

Almirall, E., Callegaro, D., Bruins, P., Santamaría, M., Martínez, P., & Cortés, U. (2022). *The use of Synthetic Data to solve the scalability and data availability problems in Smart City Digital Twins*. Retrieved from http://arxiv.org/abs/2207.02953.

Anthony, A., Sharma, L., & Noor, E. (2024). *Advancing a More Global Agenda for Trustworthy Artificial Intelligence*. Carnegie Endowment for International Peace.

Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, *19*(5). https://doi.org/10.3390/s19051181.

de Wilde, P., Arora, P., Buarque, F., Chin, Y. C., Thinyane, M., Stinckwich, S., … Marwala, T. (2023). *Recommendations on the Use of Synthetic Data to Train AI Models*. United Nations University.

Deng, H. (2023). *Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer*. Geneva, Switzerland.: UNIDIR.

di Lego, V. (2023). Uncovering the gender health data gap. *Cad. Saude Publica*, *39*(7), e00065423. https://doi.org/10.1590/0102-311XEN065423.

Encarnacion, J., Emandi, R., & Seck, P. (2022). It will take 22 years to close SDG gender data gaps. Retrieved from https://data.unwomen.org/features/it-will-take-22-years-close-sdg-gender-data-gaps.

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digit. Med.*, *6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3.

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digit. Heal.*, *2*(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082.

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., ... Tang, H. (2024). *Synthetic Data in AI: Challenges, Applications, and Ethical Implications*. Retrieved from http://arxiv.org/abs/2401.01629.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053.

ITU. (2022). *The State of Broadband 2022: Accelerating broadband for new realities*.

James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: Exploring use cases to optimise data utility. *Discov. Artif. Intell.*, *1*(1), 15. https://doi.org/10.1007/s44163-021-00016-y.

Juwara, L., El-Hussuna, A., & El Emam, K. (2024). An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns (New York, N.Y.)*, *5*(4), 100946. https://doi.org/10.1016/j.patter.2024.100946.

Kathleen Grantham. (2020). *Mapping Gender Data Gaps: An SDG Era Update*. (March). Retrieved from https://data2x.org/resource-center/mappinggenderdatagaps/.

Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A., & Markuzon, N. (2021). Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Informatics Assoc.*, *28*(4), 801–811. https://doi.org/10.1093/jamia/ocaa303.

Kokosi, T., & Harron, K. (2022). Synthetic data in medical research. *BMJ Med.*, *1*(1), e000167. https://doi.org/10.1136/bmjmed-2022-000167.

Kuo, N. I.-H., Garcia, F., Sönnerborg, A., Böhm, M., Kaiser, R., Zazzi, M., ... Barbieri, S. (2023). Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. *J. Biomed. Inform.*, *144*, 104436. https://doi.org/10.1016/j.jbi.2023.104436.

Laderas, T., Vasilevsky, N., Pederson, B., Haendel, M., McWeeney, S., & Dorr, D. A. (2017). Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. *BioRxiv*. https://doi.org/10.1101/232611.

Li, J., Cairns, B. J., Li, J., & Zhu, T. (2023). Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digit. Med.*, *6*(1), 98. https://doi.org/10.1038/s41746-023-00834-7.

Lippert-Rasmussen, K. (2022). Algorithm-Based Sentencing and Discrimination. *Sentencing Artif. Intell.*, 0. https://doi.org/10.1093/oso/9780197539538.003.0005.

Lopes, C. A., Saitabau, A., Rustagi, N., & Khosla, R. (2023). A digital health governance agenda for sexual and reproductive health and rights. *Sex. Reprod. Heal. Matters*, *31*(4), 1–6. https://doi.org/10.1080/26410397.2024.2372865.

Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2023). *Machine Learning for Synthetic Data Generation: A Review*. *14*(8), 1–19.

Makhlouf, A., Maayah, M., Abughanam, N., & Catal, C. (2023). The use of generative adversarial networks in medical image augmentation. *Neural Comput. Appl.*, *35*(34), 24055–24068. https://doi.org/10.1007/s00521-023-09100-z.

Marwala, T., Fournier-Tombs, E., & Stinckwich, S. (2023). *The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development*. (1), 1–11.

Merone, L., Tsey, K., Russell, D., & Nagle, C. (2022). Sex Inequalities in Medical Research: A Systematic Scoping Review of the Literature. *Women's Heal. Reports (New Rochelle, N.Y.)*, *3*(1), 49–59. https://doi.org/10.1089/whr.2021.0083.

Mohamed, S., & Frank, L. (2024). *Generative Adversarial Networks (GANs) for Synthetic Test Data*. (June).

Motamed, S., Rogalla, P., & Khalvati, F. (2021). Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Informatics Med. Unlocked*, *27*, 100779. https://doi.org/10.1016/j.imu.2021.100779.

Musizvingoza, R. (2024). Bridging the Gender Data Gap: Harnessing Synthetic Data for Inclusive AI. Retrieved from https://unu.edu/macau/blog-post/bridging-gender-data-gap-harnessing-synthetic-data-inclusive-ai.

Musizvingoza, R., & Lopes, C. A. (2022). *Limited gender data deepens inequalities*. https://doi.org/10.54377/916A-61EC.

Nikolentzos, G., Vazirgiannis, M., Xypolopoulos, C., Lingman, M., & Brandt, E. G. (2023). Synthetic electronic health records generated with variational graph autoencoders. *Npj Digit. Med.*, *6*(1), 83. https://doi.org/10.1038/s41746-023-00822-x.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347.

O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI Soc.* https://doi.org/10.1007/s00146-023-01675-4.

Paproki, A., Salvado, O., & Fookes, C. (2024). Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias. *ACM Comput. Surv.*, *56*(11). https://doi.org/10.1145/3663759.

Paris21. (2024). *Improving co-ordination to move the gender equality agenda forward in Africa*. Retrieved from https://www.paris21.org/news/improving-co-ordination-move-gender-equality-agenda-forward-africa.

Personal Data Protection Comission. (2024). *Privacy Enhancing Technology (PET): Proposed Guide On Synthetic Data* (No. 1; pp. 1–42). Singapore: Personal Data Protection Comission-Singapore.

Pourshahrokhi, N., Kouchaki, S., Kober, K. M., Miaskowski, C., & Barnaghi, P. (2021). *A Hamiltonian Monte Carlo Model for Imputation and Augmentation of Healthcare Data*. 1–9.

Rajabi, A., & Garibay, O. O. (2021). Towards Fairness in AI: Addressing Bias in Data Using GANs. In C. Stephanidis, M. Kurosu, J. Y. C. Chen, G. Fragomeni, N. Streitz, S. Konomi, … S. Ntoa (Eds.), *HCI Int. 2021 — Late Break. Pap. Multimodality, Ext. Reality, Artif. Intell.* (pp. 509–518). Cham: Springer International Publishing.

Reiner Benaim, A., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., … Beyar, R. (2020). Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med Inf.*, *8*(2), e16492. https://doi.org/10.2196/16492.

Skandarani, Y., Jodoin, P.-M., & Lalande, A. (2023). GANs for Medical Image Synthesis: An Empirical Study. *J. Imaging*, *9*(3). https://doi.org/10.3390/jimaging9030069.

Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Heal. Care Informatics*, *29*(1). https://doi.org/10.1136/bmjhci-2021-100457.

Sun, C., van Soest, J., & Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *J. Biomed. Inform.*, *143*, 104404. https://doi.org/10.1016/j.jbi.2023.104404.

Thuku, N., Baker, J. A., Mwigereri, D. G., Waljee, A. K., & Siwo, G. (2024). UZIMA-DS AI-Ready Synthetic Data.

Tiwald, P., Ebert, A., & Soukup, D. T. (2021). *Representative & Fair Synthetic Data*. (2002), 1–5.

van Breugel, B., Kyono, T., Berrevoets, J., & van der Schaar, M. (2021). DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Adv. Neural Inf. Process. Syst.* (Vol. 34, pp. 22221–22233). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper{_}files/paper/2021/file/ba9fab001f67381e56e410575874d967-Paper.pdf.

Williams, P. (2023). Retaining Race in Chronic Kidney Disease Diagnosis and Treatment. *Cureus*, *15*(9), e45054. https://doi.org/10.7759/cureus.45054.

World Bank. (2023). World — Synthetic Data for an Imaginary Country, Sample, 2023 A synthetic hierarchical dataset for simulation and training purposes. World Bank.

World Bank. (2024). *Data Availability | World Bank Gender Data Portal*. Retrieved from https://genderdata.worldbank.org/en/data-availability?indicator=SH.DTH.NCOM.ZS{\&}year-bucket=0{\&}country=LUX{\&}topic-year-bucket=0.

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance* (pp. 3–22). https://doi.org/10.1142/9789811238819_0001.

Yu, N., Li, K., Zhou, P., Malik, J., Davis, L., & Fritz, M. (2020). Inclusive GAN: Improving Data and Minority Coverage in Generative Models. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), Comput. Vis. — ECCV 2020 (pp. 377–393). Cham: Springer International Publishing.

The **authors** of this book are Luca Belli, Walter Britto Gaspar, Alice Rangel Teixeira, Amrita Sengupta, Andrea Bauling, Anuti Shah, Avantika Tewari, Bárbara Lazarotto, Chinasa T. Okolo, Dennis Ramphomane, Catherine Bielick, Ekaterina Martynova, Elise Racine, Faizo Elmi, Guangyu Qiao-Franco, Hellina Hailu Nigatu, Isha Suri, Jess Reia, Julio Gabriel Mercado, Laura Middeldorp, Leo Celi, Liisa Janssens, Liu Shaoyu, Liu Zijing, Mahmoud Javadi, Masego Morige, Matheus Alles. Mbali Nzimande, Nils Brinker, Pablo Trigo Kramcsák, Rachel Leach, Richard Ngamita, Richard Skalt, Rocco Saverino, Rodrigo Gameiro, Ronald Musizvingoza, Saskia Lensink, Shiva Kanwar, Shweta Mohandas, Sizwe Snail Ka Mtuze, Ying Lin, Yonah Welker, Zeerak Talat.

This volume is the 2024 outcome report of the Data and Artificial Intelligence Governance (DAIG) Coalition of the United Nations Internet Governance Forum (IGF). The Coalition is a multistakeholder group aimed at fostering critical discussions on data and AI governance from diverse perspectives, emphasizing the experiences, challenges, and contributions of the Global Majority in shaping sustainable and effective frameworks.

The DAIG Coalition aims to promote studies and stakeholder engagement to gather and evaluate evidence regarding the impact of data-driven and AI systems, critically analysing existing frameworks and institutional arrangements that regulate data and AI, and proposing policy updates that reflect the realities and aspirations of stakeholders globally, with a strong focus on inclusivity and equity.

This volume on "**AI from the Global Majority**" aspires to contribute to the evolving dialogue on how AI systems can align with diverse social, economic, and democratic ambitions. Particularly, the study seeks to address pressing questions regarding how data-driven AI systems affect key policy issues such as the full enjoyment of human rights, the fight against discrimination, biases and disinformation, and the preservation of cybersecurity, safety, democracy and the rule of law.

These all-important topics are discussed focusing primarily on the Majority World, drawing unique insights and lessons from Global South countries, which are typically underrepresented in dominant narratives.

IGF Internet Governance Forum

FGV DIREITO RIO