The gears of the machine. Power and inequalities in artificial intelligence

Collectively authored

Working Paper
December 2024

Directorate for Global Justice and International Cooperation Services of Barcelona City Council

Ajuntament de Barcelona

Commissioned by:
Oxfam Intermón

Publication Date:
December 2024

Coordination:
Carlos Bajo Erro

Authorship: Afef Abrougui, Sarah Chander, Ervin Félix, Pablo Jimenez Arandia, Pelonomi Moiloa, No Tech For Apartheid, Salvatore Romano, Sofia Scasserra, Sara Suárez-Gonzalo, Ana Valdivia, Paola Villareal and Thomas Wright.

Review: Liliana Arroyo Moliner, Carlos Bajo Erro, Aina Gallego, Thai Jungpanich, Judith Membrives i Llorens, Paz Peña, Ismael Peña-López, Natalia Pereira Martín, Sebastián Ruiz Cabrera, Susana Ruiz Rodríguez and Hernán Saenz Cortés.

Translation in Spanish version: Arantxa Albiol Benito, Belén Carneiro y Camino Villanueva.

Translation in English version: Kim Causier y Teri Jones-Villeneuve.

Layout: Jimena Zuazo.

Published by Directorate for Global Justice and International Cooperation Services of Barcelona City Council

Content

Methodology Note

In recent years, artificial intelligence (AI) tools and algorithmic systems have become ubiquitous in everyday life for a growing part of the planet's population. Meanwhile, the life cycle of those systems along with their design, development and implementation have been extended and become more complicated, leading to one of the most global and interconnected supply chains of the global economy. As a result, the expansion of AI and algorithmic systems has had an influence that goes well beyond their users or direct audiences. The territories and population groups that are the furthest removed from the global process of digitalization – those who are disconnected – also feel the consequences of the rising adoption of this technology.

The Directorate for Global Justice of the Barcelona City Council suggested putting together a book that explores the different layers of these impacts on the lives of these people to contribute to the social understanding of this multifaceted phenomenon and take a critical look at the intersection of digitalization, rights, democracy and inequality. The City Council commissioned Oxfam Intermón to produce this book as part of the third phase of the Global Digital Justice project being carried out jointly by the two entities.

This book was designed to provide a place where different voices could come together – experts in various areas, each with their own perspectives based on their approaches and experiences – to share the findings of their research in specific domains. The text, intentionally diverse and heterogeneous, aims to convey the complexity of the phenomenon and its various ramifications, interactions and even interpretations.

To handle a phenomenon with so many different aspects and the impacts of AI in today's world in a systematic way, a framework was needed that could not only take into account the various facets of this prism but also its ties, connections and relationships. To do so, Oxfam Intermón turned to the Multidimensional Inequality Framework (MIF), a conceptual and methodological framework to measure inequality using a multidimensional approach. The MIF was created by the Centre for Analysis of Social Exclusion (CASE) at the London School of Economics and Political Science (LSE), the School of Oriental and African Studies (SOAS) and Oxfam.

The MIF is based on Amartya Sen's Capability Approach, which establishes the vital aspects that influence people's ability to achieve well-being. The MIF draws from this theoretical and philosophical basis to build a framework to systematically, rigorously and comprehensively measure inequality. It identifies a series of measurable indicators that can be associated with an objective to assess inequality in people's quality of life. This attempt at measurement goes beyond the book's scope on the impact of AI on inequality, but clearly identifying the dimensions that condition access to well-being is a useful starting point to establish the key areas that need the most attention. The MIF is a foundational tool and a conceptual framework to estimate the various impacts of artificial intelligence.

The book commissioned by the Directorate for Global Justice of the Barcelona City Council uses this framework, which has been widely accepted, compared and used by various international institutions and centres for study with a global reach. The publication features seven "domains" identified by the MIF as the main areas that matter for human life to determine the impact of inequality: life and health; physical and legal security; education and learning; financial security and dignified work; adequate living conditions; participation, influence and voice; and individual, family and social life. Finally, to completely adapt this framework to the specific needs of this publication, two additional aspects with a global perspective were added, although they were not included in the MIF, given its focus on individual well-being: geo-strategic influence from an economic viewpoint and eco-social impact.

Because the MIF is a general framework, it had to be adapted to meet specific needs. For example, although the classification establishes independent aspects, efforts were made to determine the points of intersectionality, i.e. the ways some of these layers are interrelated. Similarly, the MIF is built upon a specific approach to inequality. This can be seen in how it treats the different aspects of phenomenon and its reflection in various disparities or the way it avoids an exclusively socioeconomic perspective and views inequalities through other lenses. It also goes beyond the specific dimensions to address other symbolic and subjective issues involved in creating meaning.

The new areas that were identified by adapting the MIF shape the book's structure. Each area could be approached in any number of ways, and the choice was made for this exploratory publication to take one approach that would be representative or interesting for readers and be appropriate for tackling the multifaceted perspective. To make this choice, Oxfam Intermón oversaw a collective consultation process. To provide background, the consideration and analysis of each of the areas was handled by a specialized organization or expert that had carried out relevant research in the field. Neither the approach nor the study process or conclusions drawn by each of the experts had to be shared by the Directorate for Global Justice of the Barcelona City Council or Oxfam, but their contributions are valuable in promoting and framing the discussion.

Since the goal and purpose of the publication is to add to the debate on a controversial subject, the authors of the various articles have shared the results of their research without having to meet specific content guidelines aside from rigorous analysis. This means that each piece of the puzzle does not reflect the opinions of the other authors or organizations involved in the publication or the individual positions of the other experts who participated in or the entities commissioning the project.

1. Introduction: Artificial Intelligence and Inequalities

Sara Suárez-Gonzalo,
Universitat Oberta de Catalunya - Communication Networks and Social Change / Internet Interdisciplinary Institute.
Translation: Kim Causer

Recent data show that inequality – far from declining – is a growing problem. Inequality is profound, and the barriers to living a fulfilled life with equal opportunities are increasing. Aspects like a person's gender identity, place of origin or residence, physical features, age, education or income level, and in particular, combinations of several of these factors, still clearly determine their social status, capacities and opportunities. Both as a cause and consequence of this, wealth is becoming more concentrated in the hands of the owners of large companies, particularly technology corporations. In such a context, it is not surprising that technological advancements – particularly artificial intelligence, which progressively affects everyday life – contribute to deepening inequalities rather than improving living conditions for all.

People often say that artificial intelligence (AI) 'is here to stay'. But how did it get here? *Someone brought it.* So, who, how, where, why and what (did they bring it) for become pivotal questions. Not to mention who does it all affect and how.

This volume, promoted by Barcelona City Council's Global Justice Directorate and curated by Oxfam Intermón, aims to shed light on these and other questions, offering a framework to better understand how artificial intelligence and inequality are connected. It also provides insights into how to move towards a freer, fairer and more equal future.

Rising Inequalities

We can only really understand the present and future implications of AI once we have a grasp of the context in which it takes place and is situated – one characterised by wide, ever-growing socio-economic disparities. Let's therefore start with what is possibly the least common question: where? Or in other words: in what context is artificial intelligence used?

Data show that **inequalities are a pressing problem on a global scale**. Spain is no exception. The recent Oxfam report *Inequality Inc. How Corporate Power Divides Our World and the Need for a New Era of Public Action* (Riddell et al., 2024) reveals that '4.8 billion people are poorer than they were in 2019'. Women, racialized peoples and other marginalized groups are the most affected by this situation, which furthermore translates into a widening gap between the Global North and the Global South, from which Global North countries benefit considerably despite a small percentage of humanity living in them (only 21%, according to the report). Meanwhile, it highlights a growing global concentration of wealth in the hands of a few – generally, owners of large companies that dominate markets, particularly in the technology industry. In fact, in the race to establish a monopoly in the technology sector, owners of big tech companies, like Meta, Alphabet and Amazon, are among the world's richest people (Fortune, 2023; Murphy and Schifrin, 2024; Statista, 2024).

The causes and dimensions of inequalities are multiple, complex and deeply interrelated. Researching and condemning inequalities are part of Oxfam Intermón's long history. The **Multidimensional Inequality Framework** (MIF) is a contribution based on Amartya Sen's Capability Approach, developed as part of a collaboration between academics in the Centre for Analysis of Social Exclusion (CASE) at the London School of Economics (LSE) and the School of Oriental and African Studies (SOAS), led by Abigail McKnight and Oxfam practitioners. The framework measures certain indicators based on several domains in people's lives in which inequalities can be observed: individual, family and social life; life and health; participation, influence and voice; adequate living conditions; financial security and dignified work; education and learning; and physical and legal security. Furthermore, the *VI Informe sobre la desigualdad en España 2024* (sixth report on inequality in Spain 2024) published by Fundación Alternativas, includes current data and relevant observations on this matter. The chapter written by García López (2024) interprets relevant data from a survey by Oxfam Intermón and 40dB on social perceptions and inequalities in Spain, which was conducted in the second half of 2023. According to this study, **eight out of ten people in Spain believe that inequalities exist**. The perception is heightened among women, people over 65 years of age or with middle or middle-upper income levels and those who identify as being white or Caucasian. Considering some of the dimensions of social inequality (which are all interconnected), Spaniards most strongly perceive economic inequality (the difference between rich and poor). This is followed by migration inequality (especially among irregular migrants) and territorial inequality (the disparity between different neighbourhoods within urban areas). However, most Spaniards believe that these inequalities can be eradicated, mainly if the central government, the European Union, autono-

mous communities, local councils, the media, social movements and companies (in this order of perceived importance) were to pay attention and provide the necessary resources to do so. The digitalization of society, accelerated by measures implemented to manage the COVID-19 pandemic (Ayala Cañón et al., 2022) also contributes, in different ways, to reproducing and exacerbating these and other inequalities. On the one hand, the investments made into digitalization are highly inconsistent. In Spain, proof of such are the differences in how autonomous communities obtain funding, such as European Next Generation funds for research, development and innovation (R&D+I) and for digitalization. In both cases, the Community of Madrid has benefited the most. On the other hand, there is a sharp socio-digital divide, meaning that digitalization affects people in very unequal ways (Pons and Gordo, 2024). This is particularly the case but not limited to services and products that are essential for everyday life, i.e. those used to mediate with public administrations, banks, medical services, but also those that help us communicate and maintain relationships with each other and with public and private agents. Recent reports highlight this, such as those published by Fundació Ferrer i Guàrdia (*La brecha digital en España. Conocimiento clave para la promoción de la inclusión digital* [The digital divide in Spain. Key knowledge to promote digital inclusion], 2023) or autonomous community-level reports by the consultancy firm KPMG and the Generalitat de Catalunya (Acebo Pérez, 2022) and another by the Consell Assessor del Parlament sobre Ciència i Tecnologia (Fernández-Ardèvol et al., 2024). In general, **these studies show that the digital divide is, essentially, another socio-economic divide**, that is intersectional, particularly affecting people over 65 years old (with those over 75 being more adversely affected); those with low levels of education and income; people with disabilities; rural residents and women.

Artificial Intelligence in a Context of Inequalities

Before continuing with the rest of the questions posed at the beginning of this chapter (who [brought artificial intelligence] how, what for, who does it affect and how), it is essential to begin with a simple definition of a concept that is central to this volume, yet particularly complex: artificial intelligence.

Artificial intelligence is a **knowledge discipline** that generates agents that, based on processing massive sets of data, produce results very similar to those of humans, with a certain level of autonomy and ability to adapt to new contexts and cases. These are the characteristics of an 'artificially intelligent' system. But how do artificial intelligence systems 'learn'? Machine learning is the current artificial intelligence paradigm: a set of techniques designed to create and train systems based on processing huge amounts of data on the 'problems' to be addressed and 'solutions' to be found. Today, artificial intelligence is more present than ever before in the collective imaginary and also in people's lives.

Even though the term 'artificial intelligence' was coined in 1959 (albeit with some controversy), studies in the field had already been under way for years. Since then, and particularly in recent years, **progress in the field of artificial intelligence has been significant and notable, and its uses have proliferated across myriad sectors**. Systems have been developed to personalize the medical treatment of complex illnesses, decide who should and should not receive social benefits, improve the quality of urban transport, predict the risk of criminals reoffending or speed up hiring new staff, among other examples. However, it was not until the advent of 'generative' artificial intelligence systems – like ChatGPT, a sub-field of the discipline designed to autonomously produce texts, images and other content including audiovisual formats – at the end of 2022, that most people had realized that artificial intelligence had become part of their daily lives.

These advances, however, have not affected all layers of the population equally. In recent years, the number of studies, actions and appeals on its negative consequences are manifold, especially regarding the unequal and often discriminatory effects of the development and use of artificial intelligence systems and applications. In a context of marked inequalities like today's, it is hardly surprising that technological development – particularly artificial intelligence development – contributes to deepening disparities in power, wealth and wellbeing, rather than improving people's living conditions (Eubanks, 2018).

But beyond the importance of the context in which it is produced, understanding **who can influence how and why artificial intelligence is developed and used** is essential. Fundamentally, only large technology corporations are capable of developing artificial intelligence solutions and implementing them in a market that they already monopolize. They can do so thanks to their long-standing exclusive control over three key elements: first, the data required to create and train the systems; second, the technical standards that enable interoperability between systems and promote the use of certain products and services over others; and third, the physical infrastructure that supports their operation. And their pre-eminence is further upheld thanks to a general geopolitical imbalance of power and wealth (McChesney, 2013; Tufekci, 2017; Zuboff, 2019).

Consequently, artificial intelligence development has been (and is) heavily influenced by the economic interests of these large companies, which define how these technologies are, why they are used and who has access to them. Furthermore, they operate under conditions of opacity and secrecy, which contribute to their success. In this scenario, public institutions, civil society organizations and citizens are in a position of weakness facing considerable barriers to exercising any kind of influence. To this end, democratic principles that should guide the digital transformation are compromised, inequalities are further perpetuated, and **marginalized peoples and groups are, once again, most negatively affected** through a new form of replicating factors that contribute to social divides.

As a result, in recent years, **several artificial intelligence systems have been released that have proven to be imprecise, useless, undesirable or unfair** for several reasons. Much attention has been drawn to the serious issue that they are frequently being used for objectives that negatively affect the entire population, with particularly severe effects on socio-political movements and marginalized groups and people. Decision-making based on arbitrary and unfair criteria, the limitation of opportunities or discrimination of, for example, migrants or people in precarious working situations, living in poverty or in detention have been scrutinised. Furthermore, more recently, the public exposure of other drawbacks of using artificial intelligence has become more prevalent, such as its environmental impact or its influence on public perceptions of the accompanying self-serving narratives. Civil society organizations, movements defending fundamental rights and independent journalistic and academic research play a central role in raising collective awareness about and condemning this matter. **This publication is therefore a much-needed contribution**.

The Roots and Objectives of this Publication

This collective volume is part of the **third phase of the Global Digital Justice project led by Oxfam Intermón and Barcelona City Council's Global Justice Directorate.**[1] Its aim, more so than previous phases, is to analyse whether artificial intelligence systems are deepening existing inequalities or creating new ones.

Published within the framework of the previous phase, the report *Desplazar los ejes: alternativas tecnológicas, derechos humanos y sociedad civil a principios del siglo XXI* (shifting

1 Project website [Spanish]: https://www.oxfamintermon.org/es/derechos-digitales-justos-igualitarios

the axes: alternative technologies, human rights and civil society in the early 21st century) (Calleja-López et al., 2022), prepared by the technopolitics unit of the Communication Networks and Social Change research group (Internet Interdisciplinary Institute, Open University of Catalonia) already highlighted some of the main issues associated with developing and using artificial intelligence. Among them are the limitations derived from training these systems; the fact that it is a proprietary technology, designed and controlled by large technology companies; the dubious promises constantly made by the industry; the ownership of the physical infrastructures behind these systems; the economic and geopolitical power relations on which it all depends; the oppression logic that these technologies produce and reproduce; and the lack of transparency.

On this basis, **the objective of this collective volume is to examine how the development and current uses of artificial intelligence are connected to inequalities**. Oxfam Intermón asked the 13 authors to address this matter from a situated, decolonial, feminist perspective, based on the following questions: what implications does the advancement of artificial intelligence have on (in)equalities? Do these technologies contribute to general wellbeing and promote the interests of the entire population? Do they favour any particular interests? Are citizens' fundamental rights or basic freedoms being compromised? What are the characteristics of the contexts in which we can address the above questions?

Diverse Perspectives and Approaches: A Collective Work

The publication is extensive. It examines several fundamental aspects for understanding the relationship between artificial intelligence and inequalities. All of the views **offer a comprehensive and in-depth approach to better understand this matter through diverse fields and compelling case studies**. Its objective, however, is not to exhaustively collate all possible related subjects but rather to provide a timely and curated selection of the most relevant ones, based on their social significance, current influence on the public sphere, and the availability of experts in the research area or related activities, who are responsible for writing each chapter.

The content of the volume is the result of a thorough process, made up of four stages. Firstly, Oxfam Intermón prepared a proposal of the subjects considered of interest in line with the works undertaken in the previous phases of the Global Digital Justice project, and more generally, with the organization's experience of research and field work related to this volume's subject matter. This proposal served as a starting point for two debate sessions. The first was held on 4 July 2024 and led by Carlos Bajo as part of the advisory council meetings for the Global Digital Justice project. Renata Ávila, Paz Peña, Cristina Colom, Paola Ricaurte, Eliana Quiroz, Andrea Costafreda, Laura Nathalie Hernández and Marta Peirano attended, while Liliana Arroyo provided subsequent written observations. The second, held on 12 July 2024, was led and moderated by Carlos Bajo representing Oxfam Intermón and Sara Suárez-Gonzalo (researcher and author of this introduction), and five specialists participated: Andrea Rosales Climent (Open University of Catalonia), Marta Galcerán Vercher (Barcelona Centre for International Affairs, CIDOB), Manuel Portela Charnejovsky (Pompeu Fabra University), Núria Vallès Peris (Spanish National Research Council) and Judith Membrives i Llorens (LaFede and AlgoRights). After including this session's Making note of the session's contributions, Sara Suárez-Gonzalo delivered a new proposal to the Oxfam Intermón team of the topics and content of interest. The organization later adapted the proposal and session content to the Multidimensional Inequality Framework until reaching the definitive structure of this text. The content for each of the sections was finally defined through conversations with the potential authors.

As a collaborative piece, **each chapter draws on the views, knowledge and experiences shared by the authors and the organizations they represent**. The reader should therefore not expect a uniform approach or overall coherence throughout the volume, particularly regarding the use of terms, the definition or interpretation of the contexts in which the research topic is situated, or the implications of the elements discussed. More specifically, with this collective body of work, Oxfam Intermón has intentionally sought to showcase a plurality of voices to better understand the complexity of a phenomenon for which no clear consensus exists, as it is highly contextual and its implications are still unfolding. Nevertheless, the organization team's attentive, exhaustive work of coordinating and reviewing the contributions, ensures textual cohesion and complementarity.

Structure and Content

After this **introduction**, the collective volume is structured into two main blocks, followed by a summary and conclusions.

The **first part addresses structural issues** and focuses on the context of this publication. It offers contributions that are key to understanding the current impact that artificial intelligence has from the perspective of socio-political, cultural, economic structures and the environmental issues surrounding it.

The first chapter, written by Sofia Scasserra (associate researcher at the Transnational Institute), reflects on the reproduction of the colonial, extractive logic that, throughout history, has shaped the development of humanity, and which now takes on a new neo-extractive dimension related to the deployment of new information and communication technologies like artificial intelligence. More specifically, it examines the inequalities and power imbalances caused by the new deployment of this logic between the Global North and the Global South, as well as between regions considered central and those at the social peripheries.

Ana Valdivia (lecturer and researcher in artificial intelligence, government and policy at Oxford Internet Institute, University of Oxford) covers a fundamental and often neglected aspect: the environmental impact of deploying the material infrastructure that supports artificial intelligence system development and implementation. Increased harmful emissions, the exploitation of natural resources and the disproportionate consumption of energy and water are the focal points of this contribution, which debunks the market-driven narrative that promises that artificial intelligence will solve climate change.

The **second block** explores the implications that this technological development has on different areas of people's lives, health and wellbeing.

The chapter by No Tech For Apartheid examines how artificial intelligence is related to the preservation (or destruction) of human health and life in the context of modern wars and armed conflicts. The authors explain how artificial intelligence is being used to improve the lethal efficiency of arms in wars, particularly the one ravaging Gaza. The chapter also focuses on the role that big tech companies play in this, driven by economic interests, making war a perverse laboratory for trialling technologies.

Sarah Chander (expert in European Union fundamental rights and equality policy, director and co-founder of Equinox – Racial Justice Initiative) continues with a chapter on race, ethnicity and origin-based discrimination arising from the use of artificial intelligence. The chapter dissects the main reasons for concern related to the technological advances in the area of individual, family and social life. Chander explores race and origin-based discrimination of people in different contexts, highlighting cases of structural racism in the development of artificial intelligence.

The following chapter examines the influence that artificial intelligence currently has on creating and distributing knowledge, and how accessible it is for all citizens. Pelonomi Moiloa (founder and director of Lelapa AI) dissects the implications that artificial intelligence has on producing knowledge and examines the nature of its potential influence on learning processes, and therefore, on the way that citizens understand the world. On one hand, she analyses how Global North values are imposed, in relation to the circumstances in which the technology is produced. On the other hand, she studies the standardizing effect these practices have and the impoverishment associated with tailoring knowledge generation. Lastly, she emphasizes the need for and impact of incorporating linguistic diversity in this technological process.

Promoted by Oxfam Mexico, the following chapter discusses issues related to financial security and dignified work in the so-called 'new platform economy' or 'gig economy'. The chapter covers the organization's experience in researching the impact of artificial intelligence on the conditions of people working for technology platforms. Paola Villarreal (independent researcher) and Ervin Félix (researcher at Oxfam México) explain the role that algorithms play in the imposition, loss of negotiation power and management of these conditions, and how they affect workers' lives.

Pablo Jiménez Arandia (journalist and researcher specialising in how technology and society interact) analyses the use of automated decision-making systems by public services, particularly regarding access to benefits and social protection. He challenges discourses related to rationality and improving social services' efficiency, which often accompany the introduction of artificial intelligence in this sector. The author highlights that, in practice, grave issues have been associated with the current use of the technology in this area, related to transparency, governance and being able to audit algorithmic systems. He also notes significant undesired effects, such as people losing access to basic services or the creation of social welfare management architectures that have been proven discriminatory, rooted in suspicion, and disproportionately affect the most vulnerable social groups.

The next chapter focuses on one of the star topics, which has been prominent in the public sphere since scandals like the Cambridge Analytica controversy during the 2016 electoral campaign for Donald Trump (in the United States) and Brexit (in the United Kingdom): how artificial intelligence tools can influence or condition citizens' votes in democratic elections. Salvatore Romano and Thomas Wright (both from AI Forensics) provide examples of how biased political information can influence votes based on several investigations conducted by their organization. They focus on the role that chatbots and generative AI tools play in the future of democratic integrity.

The chapter by Oxfam Middle East and North Africa (Oxfam MENA) concludes the second part of the report. It discusses the intersections of discrimination and gender-based violence with artificial intelligence. More specifically, Nadine Mouawad and Afef Abrougui (Fair Tech researchers) identify how these technologies are used in the region, and the AI strategies that large technology companies deploy, analysing how it all affects activism and gender policies.

Given the multi-faceted nature of the volume, these ten chapters, including this introduction, inevitably address interrelated issues, the overall interpretation of which is essential within the current context. That is why, the document concludes with a **summary and discussion** by Oxfam Intermón, which links the main contributions and provides a framework for interpretation. Based on such, it provides insights that could help address some of the issues highlighted and build a new context for the development and implementation of a technology that, in short, should contribute to advancing and promoting the needs and interests of everyone, not just a few.

References

Acebo Pérez, Mónica. (2022). *Estudi sobre la bretxa digital*. Generalitat de Catalunya. Departament d'Empresa i Treball y KPMG. [Catalan]. https://politiquesdigitals.gencat.cat/web/.content/00-arbre/ciutadania/Pla-de-xoc-contra-la-bretxa-digital/Estudi-sobre-la-bretxa-digital-a-Catalunya_DEF.pdf.

Ayala Cañón, Luis; Laparra Navarro, Miguel; Rodrígez Cabrero, Gregorio (coords.). (2022.) *Evolución de la cohesion social y consecuencias de la Covid-19 en España*. Madrid: Fundación FOESSA and Cáritas Española Editores. [Spanish]. https://www.caritas.es/main-files/uploads/sites/31/2022/01/Informe-FOESSA-2022.pdf.

Calleja-López, Antonio; Cancela, Ekaitz; Cambronero, Marta. (2022). *Desplazar los ejes: alternativas tecnológicas, derechos humanos y sociedad civil a principios del siglo XXI*. Oxfam Intermón. [Spanish] https://www.oxfamintermon.org/es/publicacion/desplazar-los-ejes-xxi-oxfam.

Eubanks, Virginia. (2018). *Automating Inequality. How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Fernández-Ardèvol, Mireia; Suárez-Gonzalo, Sara; Sáenz-Hernández, Isabel. (2024). *Desigualtat digital i vellesa: la bretxa digital que encara cal tanca*. Consell Assessor del Parlament sobre Ciència i Tecnologia. [Catalan]. https://www.parlament.cat/document/intrade/406609846.

Fortune. 'Global 500'. (2023). https://fortune.com/ranking/global500/search/.

Fundació Ferrer i Guàrdia. (2023). *La brecha digital en España. Conocimiento clave para la promoción de la inclusión digital*. Disponible en: https://www.ferrerguardia.org/blog/publicaciones-3/encuesta-brecha-digital-en-espana-2022-73

García López, Ernesto. (2024). 'La percepción social de la desigualdad en España: una aproximación', in Salido Cortés, O. and J. Ruiz-Huerta Carbonell. *VI Informe sobre la Desigualdad en España 2024*. Fundación Alternativas. [Spanish]. https://fundacionalternativas.org/publicaciones/vi-informe-sobre-la-desigualdad-en-espana/.

McChesney, Robert W. (2013). *Digital Disconnect. How Capitalism is Turning the Internet Against Democracy*. New York: The New Press.

Murphy, Andrea; Schifrin, Matt. (2024). *The Global 2000. 2024*. Forbes. https://www.forbes.com/lists/global2000/.

Oxfam Intermón (no date). *Multidimensional Inequality Framework.* https://inequalitytoolkit.org/.

Pons, Aleix; Gordo, Ignacio. (2024). 'Los posibles efectos sobre la desigualdad territorial de la transición digital', in Salido Cortés, O. and J. Ruiz-Huerta Carbonell. (2024). *VI Informe sobre la Desigualdad en España 2024*, Fundación Alternativas. [Spanish]. https://fundacionalternativas.org/publicaciones/vi-informe-sobre-la-desigualdad-en-espana/.

Riddell, Rebecca; Ahmed, Nabil; Maitland, Alex; Lawson, Max; Taneja, Anjela. (2024). *Inequality Inc. How Corporate Power Divides Our World and the Need for a New Era of Public Action*. Oxfam Intermón. https://lac.oxfam.org/en/latest/publications/desigualdad-sa.

Statista. (2024). *Ranking de las empresas líderes en el mundo en 2024, por valor de marca (en millones de dólares)*. [Spanish]. https://es.statista.com/estadisticas/539088/ranking-de-las-25-principales-marcas-en-el-mundo-por-valor-de-marca/.

Tufekci, Zeynep. (2017). *Twitter and the tear gas. The Power and Fragility of Networked Protest.* New Haven: Yale University Press.

Zuboff, Shoshana. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

2. Colonized Code: Unravelling Economic Inequalities in the Age of AI

Sofía Scasserra
Translation: Teri Jones-Villeneuve

At other times in history, the international division of labour has reserved for the countries of the periphery or the Global South the role of suppliers of raw materials. Historically, the countries of Latin America and other regions in the same geopolitical category have been responsible for providing the basic resources that have driven the growth of the world economy: minerals, foodstuffs and a wide variety of low value-added products. Following this logic of international division of labour, these products were processed abroad and returned as finished goods, causing a structural imbalance of payments that has profoundly marked the global distribution of power. Digitalisation and, more recently, the deployment of artificial intelligence has not changed this structure; indeed, it has built on it, adapting it to favour an unequal development model that follows the same parameters.

Artificial intelligence production chains continue to feed on resources from peripheral countries, this time incorporating the extraction of a very particular raw material, data, as well as skilled labour. However, they ensure that profits are kept in the hands of a few through the production capacities that are concentrated in the Global North and that condition the possibilities for developments in this field, and also the distribution of profits.

What Artificial Intelligence Is and Isn't

There is an expression that says, "Don't blame the tool, blame the user"[1]. Fake news, unauthorized videos and photos, discrimination, xenophobia and exclusion are just some of the problems that have emerged with the use of artificial intelligence (AI). But is AI actually the real problem? Or are these just symptoms of more deeply rooted problems that have existed since these tools were invented?

Reducing AI to a mere tool has an inescapable political undercurrent to it. It means comparing a powerful technology that has changed (and will continue to change) the nature of work, education, healthcare, and access to information and democracy – just to name a few things – to a simple bottle of water, chair, fork or other such technologies we use daily. Technology in general, and digital technologies in particular, had and have the power to transform society. The intention and objective of the manufacturer or designer is an inherent part of the design and production of any technology. This intention and objective may be altruistic, meet an occasional need, or, given that we live in a capitalist system, seek to make as much money as possible. And if manufacturing AI is a profit-seeking endeavour, it can ignore the social consequences it has and focus only on "selling". Thus, "automatism, and its utilization in the form of industrial organization, which one calls automation, possesses an economic or social signification more than a technical one." (Simondon, *On the Mode of Existence of Technical Objects*, 2017, p. 17). In this simple sentence, Gilbert Simondon explains that the process of automation (in this case, via AI) entails an intrinsically social problem. Given that the design and manufacture of any tool that automates processes includes a particular world view, the tool defines how things should align with that view. As such, any automated tool is a space for political struggle more than a tool. Indeed, if AI brings with it social problems, and those problems are mainly because the manufacturer – who has the power to produce the technology – produces it in a specific way with specific intentions, regulating this production and challenging this economic power is the new collective fight of the twenty-first century.

Ultimately, viewing AI within the frame of social struggles helps us to understand how to challenge this space to ensure an equitable, fair future where social justice exists.

This document does not intend to address gender, environmental or labour issues, among others that have come into play with AI. Instead, it adds its weight to the historic, centuries-long struggle of people in the Global South: considering AI as a new form of extractivism and technological colonialism that requires renewed efforts to not become a new Potosí.

Where Does the Periphery Fit In?

Historically, the periphery countries have played a well-defined role in the international division of labour. The global economy is organized around the assumption that comparative advantages are relatively stable variables and designated according to the resources that each economy has under its control at the moment borders are established. As such, the periphery countries in general, and Latin America in particular, have supplied the raw materials that have fuelled global growth: minerals, food and various low value-added commodities. These products are processed externally and re-enter our economies as final goods, creating a crisis in the structural balance of payments that has created economic imbalances since time immemorial.

1 "La inteligencia artificial no es peligrosa; nosotros, sí" ("Artificial intelligence is not dangerous; we are"). *La Nación*, January 3, 2015. https://www.lanacion.com.ar/tecnologia/la-inteligencia-artificial-no-es-peligrosa-nosotros-si-nid1757065/

Dependency theory and Raúl Prebisch[2], an Argentinian economist who described these processes and created strategies to avoid the pitfall of weakening trade terms, have historically been extremely popular in organizations such as UN Trade and Development (UNCTAD), which researches development in periphery nations to identify responses to these seemingly entrenched dynamics.

We are blinded by the shadow of technology. This has always been the case. The shadow traps us in the face of new tools that appear to be more magical than real. This has already happened with electricity, telecommunications and so many other inventions that marked certain areas throughout the history of humanity. Technology constantly surprises and surpasses us. No one creates an invention in order to do something worse than how people do it. Who would wear a pair of glasses to see worse? All technology improves our lives in unimaginable ways and surprises us in the process. This shadow makes us believe that another world is possible, that things can change. We also imagine selling technology in an integrated world, able to respond to new, dynamic demands that finally fix the payment crisis and put us on the global stage.

But is that what is happening? Let us take a closer look.

AI Production: Artisanal vs. Industrial

If we think about an industry, we imagine a process where a generally heterogeneous commodity is obtained. This commodity makes its way to a factory, goes through various processes, is changed, gains value and becomes a product to be sold on the market. The product then undergoes quality control and must be homogeneous and meet a defined quality standard. The product is then brought to market to be sold massively in various markets, and its market positioning reflects its level of service, quality and low cost. Broadly speaking, this process describes nearly every industrial process or industrialized product. When countries in the Global North become industrialized, they want to dominate this process across the economy.

This process also applies to AI: data are extracted – millions of pieces of various types of data that are heterogeneous and even of poor quality. The industrialization process starts: these data are selected and improved through content moderation and processing in what I like to call the "algorithmic factory": a web of algorithms that process data and convert them into valuable information. Quality control is then carried out on this information through content moderation and testing on the digital market: is the information provided by a chatbot GPT what consumers are looking for? When we answer this question through our queries, we are collaborating with this quality control process. Finally, a standardized and homogeneous product is then sold massively on the global market at an affordable price.

This describes what I like to call "industrial AI", or large-scale digital industries. The generative AI that was recently launched on the market seeks to dominate industrial AI by offering a platform on which it can build other, smaller-scale AI-based tools. Access to its production capacity is nearly impossible for periphery countries, since not only are inordinate quantities of data extracted, but the cost to train these models in terms of energy and hardware is also out of reach. For example, ChatGPT has been estimated to consume more than half a million kilowatt hours (kWh) of electricity every day, while the average Spanish household consumes around 9 kWh on a daily basis[3]. The necessary hardware for this technology also has a cost. The cost of an NVIDIA A100 GPU is esti-

2 A profile of Raul Prebisch can be found here: https://unctad.org/osg/former-secretaries-general-and-officers-charge/raul-prebisch

3 "ChatGPT consume 55.000 veces más electricidad que la media de los hogares de España" ("ChatGPT consumes 55,000 times more electricity than the average household in Spain"). *Business Insider, March 11, 2024.* https://www.businessinsider.es/chatgpt-consume-55000-veces-electricidad-hogar-medio-1371358

mated at around USD 10,000 to USD 15,000. Given that ChatGPT uses 25,000 of these GPUs, the cost of acquiring the hardware to train the model would exceed USD 250 million[4].

Meanwhile, today there are AI-based solutions that stem from these large models. Smaller digital industries have arisen in niche markets (AI for students, AI for journalists, AI for designers, etc.), but all of them use one of these large models as a basic part of their design. Copilot, ChatGPT, Gemini, Claude and others are attempting to dominate the market of large industrialized models by telling the rest of the world that "this market is already occupied, now see how our model can be used to create something more niche".

This has led to the emergence of tools in smaller yet important markets. Companies in the Global North (Europe, USA and other industrialized economies) are working to create models for entire sectors (medicine, education, etc.). And in the periphery? We are left with what I call "artisanal AI": AI-based solutions for one-off companies and needs in our territories when Global North companies have not designed them. If we look at AI production and research, the figures back up this reality: while the United States accounts for 60% of new investment in AI, countries such as China, the United Kingdom, Israel, Canada, France, Japan and German rank just behind[5]. Latin America is not even included in the statistics. Various organizations have undertaken a joint effort – an AI adoption index[6] in the region – to show how AI is being used in the region and how to improve the nascent ecosystems, but little has been said about large-scale production. If we look at patent registrations with the World Intellectual Property Organization (WIPO)[7], the situation is similar with a slight difference: China leads the world in investment, followed by the same countries mentioned above along with India.

In other words, the AI production market is dominated by large corporations and countries carrying out research on developing AI and large-scale models on the basis of which smaller solutions can be designed and sold in smaller markets. With the Global North dominating the industry, how is the Global South expected to find its place in it? It's the same old story.

But is this the only role we play in the periphery? Absolutely not.

Data Extractivism and Colonialism

When taking apart the links of the AI value chain, as we did above, the first link is the extraction of raw materials – the data. Then there is the improvement and processing, and finally quality control and mass sales. Let's take a look at what happens during each of these stages and how those in the periphery are placed.

The Data

At the start of the new millennium, people already had mobile phones, but these phones were neither smart nor designed for the various uses and apps we see today. Smartphones appeared in the late 2000s, marking the start of data extractivism – extractivism at a global scale. Data were extracted, exported on a massive scale to large technology production centres, without permission, authorization or consideration of

4 *Diez preguntas frecuentes y urgentes sobre Inteligencia Artificial* (Ten Frequently Asked and Urgent Questions about Artificial Intelligence). Fundación Sadosky, 2024. https://program.ar/wp-content/uploads/2024/08/Diez-preguntas-frecuentes-y-urgentes-sobre-Inteligencia-Artificial.pdf

5 Top 10 Countries Leading in AI Research & Technology in 2025. *Techopedia*, 2024. https://www.techopedia.com/top-10-countries-leading-in-ai-research-technology

6 Latin American Artificial Intelligence Index (ILIA) website: https://indicelatam.cl/home-en-2024/

7 Patent Landscape Report - Generative Artificial Intelligence (GenAI). World Intellectual Property Organization (WIPO), 2024. https://www.wipo.int/web-publications/patent-landscape-report-generative-artificial-intelligence-genai/en/key-findings-and-insights.html

the effect it would have on society. And unlike this extractivism, nothing was left for those who produced the data, not even any tax revenue, given that the technology companies responsible for the extractivism were located in other countries, and the data were not subject to taxes when crossing borders.

This is because in 1998, long before we realized that data could have value, the World Trade Organization (WTO) approved something with a very confusing name but with very real consequences: the moratorium on customs duties on electronic transmissions[8]. In other words, data could be extracted and exported to other countries without having to pay customs duties on them. This moratorium has been renewed without interruption every two years since it came into effect. During the last WTO Ministerial Conference in Abu Dhabi in February 2024, eight countries called for this principle to be revised. It is becoming increasingly clear that this standard is unfair.

Data are extracted, not only from the Global South but from the entire global population connected to the internet, without the payment of any kind of tax that would be redistributed back in some way to the population from which the data were extracted.

Improving Data to Train Systems

The raw data must be refined, so now comes the most merciless phase with the Global South. It does not stop with transnational extractivism – the process continues by segmenting and deepening the international division of labour in technological production in favour of the most powerful nations.

A significant part of the AI creation process involves selecting information, "cleaning" it and training the systems. This operation is known as "content moderation", and the working conditions are extremely poor and take place in strategic locations in periphery countries. Low salaries[9], no social protection, zero-hours contracts[10] and serious mental health problems[11] resulting from the hyperproductivity and content imposed on workers are only some of the horrendous consequences of these jobs that serve the AI industry to improve the quality of the raw material used to train systems more efficiently.

This theme is becoming increasingly visible in the tech world and companies are facing a rising number of lawsuits[12]. In Kenya, a first union of content moderators has already been formed and resistance is spreading. Beyond efforts to change the reality of hundreds of thousands of workers around the globe, the truth is that tech companies, through their strategy of global division of labour, seem to have assigned us tasks that they would never do in the Global North without decent working condi-

8 *The Geneva Ministerial Declaration on global electronic commerce*. WTO. 1998: https://www.wto.org/english/tratop_e/ecom_e/mindec1_e.htm

9 *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic.* Time, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

10 *Subterranean Moderators*. Pulitzer Center, June 5, 2024. https://pulitzercenter.org/stories/subterranean-moderators

11 *Mental trauma: African content moderators push Big Tech on rights*. The Economic Times, October 16, 2023. https://economictimes.indiatimes.com/tech/technology/mental-trauma-african-content-moderators-push-big-tech-on-rights/articleshow/104457622.cms

12 The occupational hazards of cleaning the internet. Coda, February 28, 2023. https://www.codastory.com/authoritarian-tech/reddit-content-moderation-lawsuit/

tions, social security and mental health protection, among other conditions. The idea that AI is produced in Silicon Valley in modern offices with ideal working conditions overlooks this part of the production chain, which is based on exploitation and extractivism.

Data Processing

Processing is where the most value is added in the industrial AI manufacturing process. For example, programmers and engineers work to convert these data into useful, saleable information for an array of services – from a generative AI chatbot to an algorithm that suggests music or a system that detects diseases based on images. Having the knowledge to do these tasks successfully is an indisputable source of added value.

But where are these tasks carried out? Who makes the decision? And above all, who are the owners of the design who retain the intellectual property rights of the service that is created? The answer seems obvious, but there are subtleties involved.

While it is true that Silicon Valley and huge corporations full of highly qualified workers are part of the equation, they are not the only ones working for Big Tech. The Global South provides labour at a much lower cost on a per-project basis, without long-term contracts in place. It is common for countries like India, Argentina and Brazil to have huge numbers of workers – engineers who have studied at local universities (and often state-subsidized public universities, as in the case of Argentina and Brazil) – who work to add value to the Big Tech products at wages that are much lower than those paid in the United States. This is good for the workers, who earn salaries in US dollars in countries with more unstable currencies, but is it good for the country?

In Argentina, for example, three problems have been seen:

1. First, these workers add value to a foreign-based industry, which then reimports the added value to the local country. In other words, this illustrates how the theory of economy dependency that Raúl Prebisch developed – where commodities are sold to then be reintegrated into another product with higher added value and thus weakening the terms of trade – is repeating itself in the technology world. Lines of code or an hour of a programmer's time is sold in order to be reintegrated into a software program, an app, or a new AI tool.

2. Second, this programmer puts pressure on salaries in the domestic market[13]. National companies are frequently in desperate need of talent and workers willing to work for salaries in pesos (given that this is what they can and must pay), but they must compete with a giant market that wants labour paid in dollars. Brain drain makes it very difficult for local tech SMEs to compete[14], which complicates national technology development.

13 "¿Cuánto gana un programador en Argentina?" ("How much does a programmer earn in Argentina?") *BAE Negocios*, September 12, 2024. https://www.baenegocios.com/sociedad/Cuanto-gana-un-programador-en-Argentina-20240912-0058.html

14 *Nuevas dinámicas de comportamiento en el sector de software y servicios informáticos (New behavioural dynamics in the software and IT services sector).* Centro Interdisciplinario de Estudios de Ciencia, Tecnología e Innovación (CIECTI), 2023. https://www.ciecti.org.ar/wp-content/uploads/2023/09/RI-Deslocalizacion-RI3.pdf

3. Finally, these workers often want to receive their salaries in dollars outside the country, and in so doing can avoid having to convert them to local currency and being taxed. As the search for workers has become more difficult, many companies have begun paying salaries in cryptocurrency[15].

In other words, the industry that adds value to AI production does not support the infrastructure that is required to leapfrog the much-needed development. The key point is that the intellectual property rights of these developments belong to North American companies, as the workers are mere project implementers. Once again, resources are extracted from the Global South without any long-term value or technology transfer that would help us support our own development. This story related to the Argentinian experience repeats itself in other countries, such as Chile, Costa Rica and Brazil.

Latin America and How Common Goods are Viewed

The region's role in AI production and development is not one of a protagonist. Throughout history, we have been relegated to being providers of resources so that others can develop technology, profit from it, and expand their world view across the planet. Indeed, AI is a tool that is highly influenced by beliefs and preconceptions, not only from the data collected but also in the programming of the tool.

There are multiple examples of how AI discriminates[16], is sexist and racist, and above all, does not take the Global South into account. Needless to say, most of the services offered work better in English[17] than in any other language. And while they do work fairly well in Spanish, this is not the case for less common languages such as Guarani or Quechua.

There are thousands of examples of technologies that were not considered for a local market. WhatsApp emojis are one case in point. For those in the Southern Cone, maté is more than just a beverage. Inviting others for a maté is a daily occurrence, a symbol of friendship, love and camaraderie. The emoji did not exist and a proposal of several dozen pages was required[18] to justify adding it to the list of Unicode characters. If this happens with such ubiquitous objects as this, imagine for a moment what happens in the black box in which the systems operate. Software programs that do not take local customs into account or disease detection systems designed with data from populations with diets, schedules and climates that are very different from where they are used are just two examples.

Latin America has much to contribute if it were given what it needed for its perspectives, customs and struggles to be reflected in the systems that are used every day.

15 "Cobrar el sueldo en Bitcoins, una alternativa que gana espacio en 2022" ("Paying wages in Bitcoins, an alternative that is gaining ground in 2022"). *Forbes Argentina*, January 5, 2022. https://www.forbesargentina.com/innovacion/cobrar-sueldo-bitcoins-una-alternativa-gana-espacio-2022-n11523

16 Algorithmic Justice League website, a US-based organisation that makes discrimination, especially on the basis of race, ethnicity or place of origin, visible in algorithmic tools: https://www.ajl.org/

17 *AI's language gap*. Axios, 2023. https://www.axios.com/2023/09/08/ai-language-gap-chatgpt

18 "Argentina gana la lucha por el emoji del mate, el primer ícono sudamericano" ("Argentina wins the fight for the mate emoji, the first South American icon"). *El País* / *Verne*, July 12, 2019. https://verne.elpais.com/verne/2019/07/12/mexico/1562944754_939569.html

Debate on common goods is raging across the region with regard to its natural resources. Are people really so obtuse as to claim ownership of a river and contaminate it when it is everyone who benefits from the fresh water? The truth is that common goods deserve special protection. They belong and are valuable to all, and we must all take care of them. Data are common goods. First, they are non-rivalrous goods[19]. This means they are not exhausted upon being consumed. For instance, art is a non-rivalrous good: one person looking at a painting or listening to a song does not prevent another person from doing the same. It is consumed, but not exhausted. The same goes for education, safety and public transport – all public services that must not be privatized. Common goods go one step farther: they do not belong to the State, but the State must protect them for everyone's good.

"You can't buy the wind / You can't buy the sun / You can't buy the rain / You can't buy the heat" says the poet in his song dedicated to Latin America[20]. Perhaps "You can't buy my interpersonal relationships, my tastes, my wants, my knowledge" should be the collective cause we should be taking up.

If this debate were not allowed, we would be losing the power of using data and information that can be produced in the interest of the people in order to create non-rivalrous technologies that benefit all of society.

Who Makes the Rules?

These techniques to create dependency have been sustained for decades through legal agreements that leverage an international system designed to ensure the most powerful win and prevent periphery countries from accessing economic development. It is a system that allows the Global North to continuously maintain its stocks of resources through obscene extractivism.

This occurs with all agreements that regulate the trade of goods and services, investments and financial flows, among other things.

But with digital issues, there are no rules. It was the upheaval brought about by AI that made the need for governance and regulation clear. Indeed, we cannot all be subject to the vagaries of the United States or China – we need rules that put things right and require technology to be designed in accordance with existing regulations. This is a pressing issue. However, no serious attempts to regulate the industry or ensure global governance have yet been made.

19 Simple explanation of the concept of 'rivalry' in economics, via Wikipedia: https://en.wikipedia.org/wiki/Rivalry_(economics)

20 *Latinoamérica (Latin America)*. Calle 13, 2011. [Music video from Youtube]. https://www.youtube.com/watch?v=DkFJE8Z-deG8

Considering Governance Based on Free Trade

The first attempt to implement technological governance in the world was drawn up[21] by lawyers for the large North American tech companies. Putting the fox in charge of the hen-house does not seem like such a great idea. The result was a proposed electronic free trade agreement known as the Trade in Services Agreement (TISA)[22]. It was later duplicated in the proposed Trans-Pacific Partnership (TPP)[23], and today is echoed in the WTO Electronic Commerce Agreement[24] and other bilateral or subregional free trade agreements. The aim of the WTO agreement is to liberalize once and for all (given its supranational character) what happens in the digital sphere.

The text (which is copied from agreement to agreement with a few changes) liberalizes the cross-border transfer of data and does not allow States to regulate the transfer, storage, processing or tracking of data. In other words, it reaffirms and enables the pillaging indefinitely, and prohibits the collection of taxes at the border. In other words, it not only maintains the status quo but sets it into international law so that no State can change it.

It also prohibits the disclosure or transfer of an algorithm's source code, which not only complicates oversight but also – and more importantly – blocks any possibility of a State demanding technological transfer as a condition of access to markets, a legal instrument that can be wielded to combat the current environmental disaster and support a just transition to more sustainable technologies. In South Africa, for instance, the water crisis[25] has made waterless textile dyeing a dire need. However, this technology has not been widely adopted across the industry due to intellectual property regulations that have impeded technology transfer. Another example is the advance of environmental regulations and standards, and their effect as barriers to trade. It is increasingly difficult for developing countries to export products that comply with the international environmental standards required by core countries. This is a hotly debated topic within the WTO, where in 2024 the Carbon Border Adjustment Mechanism (CBAM)[26] was still under discussion.

The e-commerce trade agreement absolves digital platforms from any responsibility regarding the content published on those platforms. This creates problems not only with the spread of fake news but also the promotion of violent content online. To train the systems and clean up the data, content moderators are forced to view such images. This exacerbates the precarious working conditions of workers with mental health issues, as previously discussed.

In other words, the entire agreement is structured around liberal (and extractivist) perspectives designed to increase tech companies' bottom lines, without any thought given to the impact on people.

21 EU Digital Trade Rules: Undermining attempts to rein in Big Tech. The Left in the European Parliament, 2023. https://left.eu/app/uploads/2023/03/Summary-Digital-Trade-EN.pdf

22 The Trade in Services Agreement (TISA) was a free trade agreement for which negotiations began in the mid-2010s. The agreement was negotiated in total secrecy by 23 parties. The documents were leaked on WikiLeaks. No agreement was ever reached and efforts fizzled out once Donald Trump became president in the United States.

23 Negotiations for this agreement also took place during the same decade as TISA. Unlike TISA, this agreement was eventually concluded but without the United States, which had been the main driving force. Today, it includes 12 countries from the Pacific Rim region.

24 Negotiations for the WTO Joint Statement Initiative on Electronic Commerce, a free trade agreement for e-commerce, are currently under way. The agreement has not been officially published; a leaked version can be read on the web page https://www.bilaterals.org/?wto-electronic-commerce-agreement&lang=en.

25 *Why does the Cape Town water crisis impact the textile industry?* O Ecotextiles, 2018. https://oecotextiles.blog/2018/03/07/why-does-the-cape-town-water-crisis-impact-the-textile-industry/

26 This is a mechanism proposed by the European Union (EU) for several industrial products that would restrict the entry into the EU of certain products if their carbon emissions were over a specified limit.
This would close off the market to many industries in periphery countries. https://taxation-customs.ec.europa.eu/carbon-border-adjustment-mechanism_en

The Global Digital Compact

"The Global Digital Compact (GDC)[27] is a framework for international digital cooperation that is currently being negotiated as an annex to the Pact for the Future, an intergovernmental agreement that seeks to 'build a multilateral system that delivers for everyone, everywhere' with concrete actions towards ensuring a better future for 'all of humanity'"[28]. In other words, the aim is to create a global governance framework for digital issues. The GDC follows the three pillars of the United Nations (UN) system: development, peace and security, and human rights. At first glance, this would seem to be the right direction. But there have been some criticisms from civil society[29]:

1. It does not address corporate control of the digital infrastructure: It "reduces the idea of digital public infrastructure to 'shared digital systems' without critically interrogating how the essential publicness of key aspects of digital infrastructure will be transferred into public hands"[30].

2. There is no commitment to corporate regulation: "It does not contain concrete commitments for state parties to regulate business enterprises for digital human rights compliance"[31].

3. It does not address long-term development of local capacities: "While pragmatism for the short run may need localization of dominant AI models, . . . contextual and culturally appropriate digital innovation demands policy strategies for building longer term capacity for local AI models"[32].

4. Mechanisms are needed to ensure inclusive innovation: "Digital public goods cannot be construed as automatic enablers of inclusive innovation; their functional merit lies solely in public interest governance"[33].

The GDC does not address the problem of the concentration of power in the digital economy, and it puts forward superficial solutions that do not tackle the deeper causes of digital inequality. A more robust approach is needed that promotes local capacity-building, public governance of digital infrastructure and effective regulation of companies to ensure a more just and equal digital future that leaves extractivist and colonialist practices behind.

What About Regionalism? Lack of Capacity for Capacity-Building

This dismaying reality with regard to the global governance instruments begs the question: What if we implement regional governance to have AI that reflects our values and principles?

Theoretically, this would be a smart approach in a world that not only seems unable to address long-standing structural inequalities, but appears to be trying to make them worse. In Latin America, we have integration forums such as Mercosur and the Pacific Alliance, among others. But debates on AI seem to revolve around superficial topics about what AI is expected to do: protect privacy, ensure access, not discriminate, etc. While these are important issues, they do not lead to deeper conversations on inequality, the concentration of power or digital industrialization.

27 Global Digital Compact website: https://www.un.org/digital-emerging-technologies/es/global-digital-compact
28 "The Global Digital Compact We Need for People and the Planet", IT for Change, 2024: https://itforchange.net/index.php/global-digital-compact-we-need-for-people-and-planet
29 Global Digital Justice Forum website: https://globaldigitaljusticeforum.net/
30 "The Global Digital Compact We Need for People and the Planet", IT for Change, 2024: https://itforchange.net/index.php/global-digital-compact-we-need-for-people-and-planet
31 Ídem.
32 Ídem.
33 Ídem.

The question then is how to ensure that the voice of civil society gets heard in a governmental bureaucracy that seems to have forgotten those whom it is meant to serve. Forums for participation are limited and the agendas seem to have been co-opted by the corporate lobbyists.

Furthermore, the development of sovereign alternatives is both necessary and urgent: viewing technology based on our own perspectives and no longer being dependent also means producing technology that serves our people. This could be achieved by creating regional public digital agencies that offer services and act as alternatives to Big Tech.

We must ask ourselves: Do we have this capacity? The answer is unclear and complex. There are ideas, there are human resources, and there are researchers who are exploring alternatives and bringing the Latin American perspective to the development of alternative digital technologies. The capacity to imagine a better future seems to be there. But can capacity materialize into opportunities? Can this capacity exist if the brain drain continues? If there is no financing? If there is a structural crisis in the balance of payments? If there are signed trade agreements that prevent governments from taking action to promote policies for national digital development?

We believe that despite all the adversity, it is possible. The region has identified solutions to difficult problems in the past, and productive South–South integration could be the solution now. Building sovereign technologies is only possible if we join forces and sell products to our own markets designed by and for our own people.

Conclusion: Working Towards Economic and Digital Justice

Throughout this paper, we have examined from various angles how the Global South has long been framed within an extractivist and colonialist perspective that is now being repeated with the manufacture of AI at the global level.

In the political and economic struggle for power, it seems that we have lost the battle to use new technology to change perspectives, escape structural dependency and win new comparative advantages with added value.

But all is not lost. There is an ongoing geopolitical battle between China and the United States for domination of the tech market, and if we maintain our regulatory sovereignty and are able to build smart economic strategies for productive integration, we can position ourselves in the global markets with intermediate inputs for technology manufacturing. In the economic fight, if we align ourselves with productive integration, we have the capacity to bring innovative, alternative technologies to market. The question is whether we will have governments that allow the sale of development through public–private participation, where transnational capital is the new proprietor of our intellectual property, or if they will protect industrial sovereignty and value creation.

Economic and digital justice is urgently needed. It is possible to do technology differently. But to make it a reality, we need for every part of the planet to be able to create it, developing competitive, innovative markets that benefit from diversity.

Our destiny is in our hands. Humanity can build something better if we do not allow power to be concentrated in the hands of so few, as in the case of AI. The debate is growing at the global level. We must rise to the challenge. If we fight together, the Global South would be an unstoppable force. We cannot allow the pillaging of our resources for another century. We cannot be another Potosí.

3. The Environmental Impact of Artificial Intelligence: It's Not a Cloud, It's an Industrial Warehouse

Ana Valdivia
Oxford Internet Institute (University of Oxford)
Centre for Capitalism Studies (University College London)
Translation: Teri Jones-Villeneuve

The current climate emergency has created a pressing need to consider the technologies that will be useful and effective in caring for human and non-human lives. Artificial Intelligence has been put forward as a solution because it can analyse patterns in data. The major tech companies are now promoting algorithmic solutions to reduce the impacts of climate change. However, the rising need for energy to fuel this technology has drawn the curtain back on Artificial Intelligence to show that it is part of the problem, rather than the solution. Various academic studies and investigative journalism reports have shown that the recent generative Artificial Intelligence consumes vast amounts of natural resources such as minerals, water and land, in turn creating political and environmental conflicts. This chapter critically reviews how Artificial Intelligence and its infrastructure contributes to climate change, and in so doing sparks resistance from local communities facing the extractivism that supports the ongoing training of Artificial Intelligence models. Ways to change the course of this technology and imagine a world where algorithms truly benefit everyone – if this is possible – are also considered.

Introduction

Climática, an independent media outlet specialized in climate and biodiversity, announced that 2024 was the hottest year on record and the first to exceed the 1.5°C limit of global warming (Robaina 2024). The average global temperature was 1.59°C above pre-industrial levels, which climate scientists have warned has serious ramifications for life on Earth. Valencia (Spain) experienced first-hand the consequences of all the years of burning fossil fuels: the Mediterranean Sea became too warm for the capitalist activity that fed unprecedented storms and floods. And yet, scientists have been warning us since 1856, when Eunice Foote proved that some gases create a greenhouse effect in the atmosphere (Foote, 1856). Another scientist, Guy Callendar, found that land temperatures had risen 0.3°C in just 50 years. Since then, various physical models have been developed to better understand the effect of fossil fuels on the Earth and to show how science and technology can be used to improve our understanding of the impact of burning fossil fuels.

Recent technological advances have also improved our scientific knowledge of the climate emergency and its connection to fossil fuels. More sophisticated computer models and sensors that can make precise measurements have led to the development of more accurate climate models. Within this technological context, artificial intelligence (AI) has been put forward as a key tool to combat climate change. This technology can analyse patterns in huge amounts of data. Cowls et al. (2023) explained the opportunities offered by this technology to mitigate climate change in various sectors such as transport, energy, urban planning and climate prediction. For example, algorithms have been developed based on machine learning techniques that improve a building's energy efficiency by analysing energy-use patterns and optimising energy use (Zekić-Sušac, Mitrović and Has, 2021).

Given this technological optimism that AI could be used to tackle the climate emergency, the major tech corporations have wasted no time in developing and funding their own solutions through their philanthropic commitments. One example is Jeff Bezos's company, Amazon. This tech firm, which according to Wikipedia has an operating income of US$36.85 billion, founded the Besos Earth Fund in 2020 with US$10 billion to combat climate change. The foundation announced on 16 April 2024 that it was going to invest US$100 million in AI projects to "protect our planet". The foundation's vice chair presented the initiative on YouTube, saying:

> "When you hear AI, what comes to mind? For me, it's the potential for humanity to solve problems – big problems, like climate change and nature loss. The question is, How? That is what we're exploring at the Bezos Earth Fund, and I am so excited to announce our AI for Climate and Nature Grand Challenge – a commitment of up to $100 million to harness the power of AI to protect our planet."
>
> Lauren Sánchez (2024)

The Bezos Earth Fund plans to focus its climate change mitigation efforts on sustainable proteins, energy grids and biodiversity conservation. Some of the research questions mentioned in the promotional video include, Can AI can analyse protein combinations to produce meat alternatives? and Can AI bring renewable energy to communities that do not have access to electricity? However, it is interesting to note that the foundation does not ask whether AI can help Amazon calculate or mitigate its own carbon footprint. In fact, Amazon was accused in 2020 of drastically underestimating its greenhouse gas emissions.

Journalist Will Evans investigated a story in which he reported that Amazon did not want to publicly disclose its carbon footprint to independent and non-profit organizations such as the Carbon Disclosure Project so they could make the information public and compare it with other companies. Moreover, in its counts, Amazon only included the carbon emissions of products the company makes itself and excluded those from products they did not manufacture, which make up an estimated 60% of the company's sales. As a result, some of Amazon's employees have pressured the company to implement genuinely sustainable practices. Recently, a group of Amazon employees issued a report, called *Burns Trust: The Amazon Unsustainability* Report, which criticizes the tech company for contributing to the climate emergency by selling AI tools to fossil fuel companies (Amazon Employees for Climate Justice, 2024).

Moreover, little is known about the environmental impact of Amazon's digital infrastructure. The company provides cloud-based services, such as storing data and running algorithms, to other companies. However, these "cloud" services are not being provided in the sky, but rather from an industrial warehouse on terra firma. These Amazon services are provided through data centres, also known as server farms, which require energy and water to run and cool the servers that store and process the data or train AI tools such as ChatGPT. Because data centres run 24/7, 365 days a year, their environmental impact is colossal and has caught the attention of scholars, journalists and activists in recent years. For example, local communities near Amazon data centres in the United States – which has the most data centres in the world – have protested against the noise they make and the amount of water they use. In Virginia, protesters opposed the Amazon data centre because it would require installing over 40 km of power lines in a rural area (Smolaks, 2015). This case, as well as others we will discuss in this chapter, show how AI and its growing infrastructure are problematic for the climate emergency because they exacerbate inequality and violate environmental rights.

This chapter, called "The environmental impact of Artificial Intelligence: it's not a cloud, it's an industrial warehouse" and published in the book *Los engranajes de la máquina. Poder y desigualdades en la inteligencia artificial.* (The gears of the machine. Power and inequalities in artificial intelligence), offers a critical look at the Artificial Intelligence industry and infrastructure. We start by analysing the materiality of Artificial Intelligence through an overview of its infrastructure and supply chains. We then take a closer look at real cases of resistance that have emerged due to the social inequality that Artificial Intelligence infrastructure creates. Finally, we assess the sustainability of the current development model for this technology, as well as the types of instruments that exist to push back against the inequality of the tech industry.

AI Infrastructure and Supply Chains

The physical infrastructure of the digital world and the internet have remained out of sight over the years. The major technology companies have used the cloud metaphor to talk about the place where our conversations, images and other digital materials are stored, thereby rendering the enormous data centres and submarine cables invisible. An icon of a blue cloud is even used on our computers to refer to this place. This cloud metaphor creates a sense of immateriality through an ethereal image of digitalization without any environmental consequence or impact (Jacobson and Hogan, 2019; Wiig, 2015). Although some scholars, activists and journalists warned us years ago of the materiality of the digital world and advised staying aware of the amount of water and electricity it consumes (Hogan, 2015; Velkova, 2016; Brevini, 2021; Peña, 2023), it was only recently that the digital materiality stopped being invisible.

AI was trumpeted as a technology that could solve climate change through algorithms that identify deforestation or predict sea rise (Boston Consulting Group, 2022). However, generative artificial intelligence, which is trained to generate texts or images through other texts or images, has called into question the benefits of this technology for climate change due the amount of resources it requires. Some estimates have found that Llama 3, Meta's large language model, used 22 million litres of water in just 97 days – the amount a resident of Barcelona's Sant Andreu neighbourhood would use in 643 years (Li et al., 2023). And it is not just the amount of water but also electricity and other natural resources needed for training these types of algorithms that has risen over the years: as an algorithm gets larger and more sophisticated, greater computing capacity is needed, which requires even more resources.

But the issue goes beyond the amount of resources needed for training these algorithms. Recent research has shown that AI supply chains can bring to light the resources and labour required to make an AI chip that is later used to train the algorithm (Valdivia, 2024). These supply chains can be divided into three phases: mineral resource extraction, data centres and electronic landfills.

Mines, Factories and Chips for Artificial Intelligence

Environmental activist and political scientist Jorge Riechmann, PhD, recently reported that humanity consumed a greater volume of materials in just eight years (2016–2023) than in the entire twentieth century (Materialflows.net, 2024). Such a volume of materials shows that our societies are consuming more products made of materials that must be extracted, transported and manufactured. The textile and food industries contribute to this extraction, which is necessary to make clothing and food. However, the digital transformation means that the electronics industry is also contributing to this extractivism by mining materials required to make mobile phones, computers and chips.

Within this digital transformation, the mobile telephone and laptop industry has received substantial attention, particularly with regard to the types of materials that go into making our phones and their environmental impact (Rodriguez, 2024). Meanwhile, the AI chip industry has gone largely unnoticed. In fact, companies manufacturing electronic products for this technology have become the most powerful based on variable market capitalization, an economic indicator that reflects the total value of a company's shares. On 15 November 2024, the American corporation NVIDIA – the tech company that sells 80% of the chips used in AI (also known as GPUs, or graphics processing units) – surpassed other companies including Apple, Microsoft and Amazon to become the global leader in terms of market capitalization, with US$3.6 trillion (Statista, 2024).

Although NVIDIA distributes the chips, Taiwan-based TSMC actually manufactures the GPUs. TSMC uses specialized equipment from the Netherlands to manufacture AI electronics, which also has serious environmental consequences that challenge claims that this technology can save us from climate change. For example, TSMC's latest sustainability report estimated that the amount of water used at its plants has increased year after year, reaching 260,000 cubic metres of water per day in 2022 (S&P Global Ratings, 2024). Due to the amount of water used by TSMC, the Taiwanese government got caught up in a scandal in 2021 when it announced to the country's farmers that priority would be given to chip manufacturing over rice crops (Zhong and Chang Chien, 2021). But water is not the only issue. GPUs are made of 99.9% silicon, i.e. sand, the extraction of which also causes serious environmental harm related to mineral deposit mining (Missouri Coalition for the Environment, 2023). Other materials such as tantalum, gold and silver are also used in smaller amounts, but their extraction also has negative impacts (Valdivia, 2024).

Data Centres

AI algorithms have become increasingly sophisticated, and to achieve this, a considerable amount of data is required to train the algorithms. If we analyse algorithm size based on the number of parameters – the configuration variables required to guide the algorithm as it is trained – we can see that it has grown in recent years. For example, while BERT, a language model released by Google in 2018, has 340 million parameters, GPT-3, the ChatGPT algorithm designed by OpenAI in 2020, has 175 billion. This higher number of parameters has changed the way the algorithms are programmed, since they cannot be trained "locally", i.e., on your own computer. These algorithms must be trained in "the cloud", i.e., in the data centre, not only due to the algorithm's size but also because of the quantity of data required. This is reason for the increased attention on the environmental impact of AI in recent years, since the cloud, as mentioned in the previous section, is not in fact a cloud but rather an industrial warehouse that uses enormous quantities of water and energy, not to mention land.

It has been estimated that data centres consume 1% of global electricity (Spencer and Singh 2024). Although not all data centres are used to train AI, this technology has increased the energy use of those centres that do. For example, Microsoft and Google noted in their 2024 sustainability report that their carbon emissions have risen due to the energy they need to train their AI algorithms. What's more, many data centres use fossil fuels to produce energy. Because these companies have made commitments to achieving zero emissions by 2030, Google, Amazon and Microsoft have announced investments in nuclear energy. In the future, they would be able to say they do not produce carbon emissions, but they would be consuming uranium (Penn and Weise, 2024).

Data centres also use water. The server rooms in this infrastructure must be cooled, and water is the most efficient way to do this. Data centres operate 365 days a year and so consume water constantly. As a result, the environmental impact of data centres is also related to the amount of water they use. But it is difficult to know how much water they use given the lack of transparency around their consumption and because the companies hide behind the veil of corporate secrecy to avoid disclosing these data. However, there have been scandals related to data centre water use. Microsoft became embroiled in one in the Netherlands in 2022. While the tech company founded by Bill Gates promised that it would need only 12 to 20 million litres of water to build its data centre, a local newspaper uncovered – in the middle of a summer drought advisory – that it had actually consumed 84 million litres a year during the centre's construction, or four times more than initially announced (Vuijk, 2022). Although Microsoft said that 36 million litres were put back into the water supply system, this claim was rejected by the local community, which had opposed the project prior to construction. As a result, the Netherlands issued a nine-month moratorium on data centres, but with two exceptions, one of which was the Microsoft data centre.

Electronic Waste

AI's environmental impact does not end in a data centre, but in a landfill. Although much has been said about the amount of electronic waste the Global North sends to the Global South and the environmental harm this waste causes, very little analysis of how AI and its infrastructure contribute to this waste has been documented. Recent research into the environmental impact of AI and its supply chain has shown that GPUs have a life span of around 3 to 5 years (Valdivia, 2024). This means that within 5 years, data centres dispose of their chips and install new ones, thus creating more electronic waste. A recent study from Nature estimated that if no measures are taken, the electronic waste produced by the latest AI models would increase by 1.2 to 5 million tonnes between 2020 and 2030 (Wang et al., 2024).

Electronic waste can contaminate soil and water due to the chemicals that leach out of the electronic circuits. For example, higher levels of mercury have been found in soil and ground and surface water near two e-waste recycling sites in Ghana (Amponsah et al., 2022).

The Impact on Local Communities: the Case of Meta in Talavera de la Reina (Spain)

Talavera de la Reina is a city in Castilla-La Mancha (central Spain), a region well known for its rural heritage. However, this area has suffered from major population decline due to rural flight, a phenomenon referred to in Spain as España vaciada, or "empty Spain" (Taibo 2021). Talavera de la Reina has also become known recently for hosting the first Meta hyperscale data centre in Spain. The project is part of what the Castilla-La Mancha regional government declared a "Project of Singular Interest", together with a casino, airport and golf course, a label that facilitated "urbanization through land rezoning, neglect of existing ecosystems and permission to build in protected areas" (Escudero-Gómez 2023). The data centre project covers 191 hectares to build 130,000 square metres to house servers. This ambitious project promises a sustainable plan with actions such as "reversing biodiversity loss" and "returning more water than is consumed" (Meta and Zarza Networks 2023).

The environmental impact assessment to which we had access estimated that the data centre would have an installed electrical capacity of 248 MW, the equivalent of what 71 Spanish households consume in one year. In terms of water supply, total consumption was estimated at 327 million litres a year, equal to the annual consumption of 7,000 households. Additionally, the environmental impact assessment stated that the project was located in a zone covered by a conservation plan for the Spanish imperial eagle and the cinereous vulture, but the land occupied by the data centre accounted for only 0.004 and 0.009% of this zone. The environmental impact on these protected species was estimated by counting the number of birds observed during six field visits between December 2021 and June 2022.

However, Meta's environmental impact assessment was criticized by environmental, academic and news organizations. As in the Microsoft case in the Netherlands, one of the main criticisms was about water consumption, a sensitive topic given that Spain faces extremely severe droughts that are worsening with the climate emergency. One important objection, formally raised by SEO/Birdlife, the Spanish Ornithological Society, was about this infrastructure's water use. However, the allegation was dismissed. The Meta hyperscale data centre announced that its drinking water consumption would be reduced from 327 million to 40 million litres annually, and reassessed its maximum water demand from 37 to 10 litres a second. Even with these revised estimates, Meta will consume 8% of the total water used in Talavera de la Reina. Meta plans to dramatically reduce its water consumption by using dry air coolers, which are cold air flow systems that do not rely on water. But the literature on data centre cooling suggests that this method is effective in cool regions, which is not the case for Talavera de la Reina in summer.

Moreover, SEO/Birdlife claimed that the bird monitoring method needed to be improved to better account for the impact of the data centre on local fauna. SEO/Birdlife noted that the study should take place over at least a full year to more accurately assess the data centre's impact in the region. In addition, the environmental impact assessment did not consider how the data centre would affect the loss of food production and forage areas. The report also did not mention that the zone would be located just 3 km from the critical zone for the Spanish imperial eagle. Once again, these allegations about the impact of data centres on fauna were dismissed.

In response, the first grassroots organization against data centres in Spain was created: Tu Nube Seca Mi Río ("Your cloud is drying up my river"). This organization defines itself as a techno-environmental group, set up in April 2023 with the aim of

> "[R]aising awareness about the environmental impact of data centres, especially with regard to water use. We have united as a result of the Meta/Facebook initiative to create a large data centre in Talavera de la Reina (Madrid, Spain)."
>
> *Tu Nube Seca Mi Río*

During an event organized at the University of Cambridge (United Kingdom), which brought together various local communities across Europe opposed to the impacts of data centres, an activist from Tu Nube Seca Mi Río stated that Mark Zuckerberg "drove [her] to become an activist against data centres." Her family had previously opposed the construction of the airport, another of the Singular Projects supported by the local government, and was now looking into how this data centre could also negatively impact the area. Local farmers were already struggling to irrigate their fields due the severe droughts affecting the region. Although to date we do not know if the data centre in Talavera de la Reina is directly related to AI, we do know that Mark Zuckerberg is planning to virtually host the metaverse and related products, some of which are designed with AI (Nix 2023). Given the current climate emergency, a pressing question arises: Do we prioritize water use to support life, or to support Silicon Valley's dreams?

A Proposal to the Possible Development of Ai in a Climate Emergency Context

Technological solutions do exist that can prevent and mitigate climate impacts. Physical models and differential equations that can help predict the path of a hurricane, a heat wave or a cut-off low have been effective in providing advanced warnings on climate risks (if and when politicians are aware of the environmental risks associated with the climate emergency). We know that these impacts, which are increasingly severe and pose risks to human and non-human life, are a consequence of a capitalist, neoliberal and extractivist economic system that burns fossil fuels to accumulate capital. AI, if not applied with a critical approach, simply continues feeding this economic system that is destroying our oceans, lands and mountains. We must consider the benefit an algorithm can offer our society, comparing and critically analysing the limits, risks and negative consequences associated with the algorithm, as well as the volumes of natural and financial resources required to operate it.

Although much as been said about the risks of algorithms, such as greater discriminatory bias (Valdivia and Sánchez Monedero 2022), today this infrastructure has also become a source of oppression. This infrastructure includes data centres that hide their environmental impacts or are not transparent about their water use, supply chains for the chips used to train AI algorithms that extract minerals and create serious environmental harm to the land, and the electronic waste created by the chips at the end of their life that continues to contaminate soil, rivers and aquifers (Taffel 2021). These processes of extractivism and dispossession echo the critical theories of data colonialism developed by scholars Ulises A. Mejias and Nick Couldry, who trace the link between the intrinsic colonialism in data extractivism and the resources required to manufacture digital devices that take from the Global Majority for the profit of the Global North (Mejias and Couldry, 2024). Given the current climate emergency, we must remain alert, not only to whether the technological solution will be truly useful, but also to the emerging social and environmental injustices due to the growing digital infrastructure industry as more data centres are being built, and which thus contribute to more electronics factories. All of this growth has an impact on the soil, insects, birds and local communities where data centres are built, on the mines where minerals are extracted to make the chips, or on the electronics landfills where chips are disposed of at the end of their life.

The negative consequences and discriminatory impact of various technological solutions have led government institutions to take action, such as when the European Commission established a legal framework to regulate this technology, known as the Artificial Intelligence Act. Although digital rights organizations are satisfied with the final text, this regulation does not address the environmental damage that AI causes. In fact, only the United States

has published a text that does so. In 2024, the Senate drafted a bill under President Biden's administration that would require the US Environmental Protection Agency to study the environmental impacts of AI and set up a system to report impacts. Although the European AI legal framework does not take AI's environmental impacts into account, there are other legal tools that could be useful to shed light on the resources this technology consumes. For example, the European Parliament revised the 2012 Energy Efficiency Directive, which now requires data centres with a capacity of more than 50 MW to report their usage metrics, such as for energy and water. However, there is a regulatory opportunity to protect local communities against the growing AI infrastructure, which includes data centres as well as mines, chip factories and electronics landfills.

The concept of the black box algorithm has been discussed for years, but we should now also be talking about the black box algorithmic and digital infrastructure. We need transparency and accountability with regard to AI supply chains and impacts. For example, it is impossible to trace all the suppliers of corporations such as NVIDIA or Microsoft, which means it is impossible to calculate their actual carbon emissions. Although there are regulatory frameworks that require more transparency about supply chains, many of the sustainability reports published by the tech industry players do not take their suppliers' emissions into account. This means that we do not know with any certainty what the real carbon footprint of the tech and digital industry is. We need a regulation that promotes the common good and takes a truly democratic decision about digital infrastructure – in other words, a regulation that lets communities decide if they want to allow a silica mine to be created or a data centre built, based on transparent communication about the consequences of that decision. Communities should be informed about the environmental damage caused across the technological supply chain, and know which of these private actors are damaging our ecosystems.

But we need more than regulation. We can also reappropriate the technology, and especially AI, and redirect it towards the common good and move away from capital accumulation. Yásnaya Elena Aguilar Gil, a Mixe linguist and activist, introduced the concept of *tequiología*, a view that attempts to distance technology from its usual purpose of serving the market, skills and private interests and instead use it for the common good and cooperation. Doing so involves sharing data, making code accessible and building public digital infrastructures that truly serve the people during this time of climate emergency. The idea is to start small and local, drawing from the historic resistance of oppressed people who, organized in small communities, forged a mutual support network to avoid taxes or rebel against the abuses of the Spanish Crown.

Using this framework as a basis, AI could become a powerful tool for the common good and building resistance to injustice if it were applied to less ambitious projects at a local scale. Injustices such as natural resource exploitation or the opaqueness of the data centre industry, which is analysing vast quantities of public information – something that AI is very good at – could be made more transparent. The aim is to reappropriate AI to put it to use for the people and the collective interest. As the Ayuujkjä'äy linguist herself says:

> "If the world adopted this vision of *tequiología*, maybe we could save the creative potential of new technologies and divert them away from a system that devours them and threatens human life."

References

Amazon Employees for Climate Justice (2024). *Burns Trust: The Amazon Unsustainability Report.* Available at: https://static1.squarespace.com/static/65681f099d7c3d48feb86a5f/t/668ebf702516716ca72bbf98/1720631157044/unsustainability-report.pdf (Accessed 10 November2024).

Amponsah, Lydia Otoo, et al. (2023). Mercury contamination of two e-waste recycling sites in Ghana: an investigation into mercury pollution at Dagomba Line (Kumasi) and Agbogbloshie (Accra). *Environ Geochem Health* 45, 1723–1737. https://doi.org/10.1007/s10653-022-01295-9

Brevini, Benedetta (2021). *Is AI Good for the Planet?* Polity, Cambridge (UK)

Boston Consulting Group (2022). *How AI Can Be a Powerful Tool in the Fight Against Climate Change.* Available at: https://web-assets.bcg.com/ff/d7/90b70d9f405fa2b67c8498ed39f3/ai-for-the-planet-bcg-report-july-2022.pdf (Accessed 14 November 2024).

Callendar, Guy Stewart (1938). The artificial production of carbon dioxide and its influence on temperature. *Quarterly Journal of the Royal Meteorological Society*, 64(275), 223–240. https://doi.org/10.1002/qj.49706427503

Cowls, Josh, Tsamados, Andreas, Taddeo, Mariarosaria and Floridi, Luciano (2023). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *AI & Society,* 1–25. https://doi.org/10.1007/s00146-021-01294-x

Escudero-Gómez, Luis Alfonso (2023). Construir cualquier cosa en cualquier lugar: los Proyectos de Singular Interés en la región de Castilla-La Mancha (España). *EURE* (Santiago) 49, no. 147: 1–24. http://dx.doi.org/10.7764/eure.49.147.08

Evans, Will (2023). Private Report Shows How Amazon Drastically Undercounts Its Carbon Footprint. *Reveal News*. Available at: https://revealnews.org/article/private-report-shows-how-amazon-drastically-undercounts-its-carbon-footprint/ (Accessed 10 November 2024).

Foote, Eunice (1856). ART. XXXI. Circumstances affecting the heat of the sun's rays. *American Journal of Science and Arts* (1820–1879), 22(66), 382.

Hogan, Mél (2015). Data flows and water woes: The Utah Data Center. *Big Data & Society*, 2(2). https://doi.org/10.1177/2053951715592429

Jacobson, Kate and Hogan, Mél (2019). Retrofitted data centres: A new world in the shell of the old. *Work Organisation, Labour & Globalisation*, 13(2), 78–94. https://doi.org/10.13169/workorgalaboglob.13.2.0078

Li, Pengfei, et al. (2023). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. https://doi.org/10.48550/arXiv.2304.03271

Materialflows.net (2024). *Global trends of material use*. Available at: https://www.materialflows.net/global-trends-of-material-use/ (Accessed 18 November 2024).

Mejias, Ulises A. and Couldry Nick (2024). *Data Grab: The New Colonialism of Big Tech (and How to Fight Back*). WH Allen, Dublin (Ireland).

Meta and Zarza Networks (2023). *Proyecto de Singular Interés "Meta Data Center Campus"*.

Missouri Coalition for the Environment (2023). *Impacts of Silica Mining*. Available at: https://moenvironment.org/wp-content/uploads/sites/370/2023/01/Impacts-of-Silica-Mining-1.3.2023-1.pdf (Accessed 18 November 2024).

Nix, Naomi (2023). Facebook pivoted to the metaverse. Now it wants to show off its AI. *The Washington Post.* Available at: https://www.washingtonpost.com/technology/2023/05/14/meta-generative-ai-metaverse/ (Accessed 18 November 2024).

Penn, Ivan and Weise, Karen (2024). Hungry for Energy, Amazon, Google and Microsoft Turn to Nuclear Power. *The New York Times*. Available at: https://www.nytimes.com/2024/10/16/business/energy-environment/amazon-google-microsoft-nuclear-energy.html (Accessed 18 November 2024).

Peña, Paz (2023). *Tecnologías para un planeta en llamas*. Paidós, Santiago de Chile (Chile).

Robaina, Eduardo (2024). Un 2024 para la historia: el año más caluroso y el primero por encima de 1,5 °C. *Climática*. Available at: https://climatica.coop/2024-ano-mas-caluroso-y-por-encima-15-oc-copernicus/ (Accessed 10 November 2024).

Rodriguez, Helena (2024). Los dos "Mobiles": radiografía del impacto ecosocial del sector de los 'smartphones'. *Climática*. Available at: https://climatica.coop/mobile-impacto-eco-social-smartphones/ (Accessed 18 November 2024).

S&P Global Ratings (2024). TSMC And Water: A Case Study Of How Climate Is Becoming A Credit-Risk Factor. Available at: https://www.spglobal.com/ratings/en/research/articles/240226-sustainability-insights-tsmc-and-water-a-case-study-of-how-climate-is-becoming-a-credit-risk-factor-12992283

Sánchez, Lauren (2024). AI for Climate and Nature: The Bezos Earth Fund Announces $100M Grand Challenge. Available at: https://www.youtube.com/watch?v=nyP3yU5Qu9Y&ab_channel=BezosEarthFund (Accessed 10 November 2024).

Smolaks, Max (2015). Amazon faces Virginia protest over power lines. Data Center Dynamics. Available at: https://www.datacenterdynamics.com/en/news/amazon-faces-virginia-protest-over-power-lines/ (Accessed 10 November 2024).

Spencer, Thomas and Singh, Siddharth (2024). What the data centre and AI boom could mean for the energy sector. *International Energy Agency*. Available at: https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector (Accessed 18 November 2024).

Statista (2024). Leading tech companies worldwide as of November 15, 2024, by market capitalization (in billion U.S. dollars). *Statista*. Available at: https://www.statista.com/statistics/1350976/leading-tech-companies-worldwide-by-market-cap/ (Accessed 18 November 2024).

Taffel, Sy (2019). *Digital Media Ecologies: Entanglements of Content, Code and Hardware.* Bloomsbury, Ireland (Dublin).

Taibo, Carlos (2021). *Iberia vaciada: Despoblación, decrecimiento, colapso*. Los Libros de la Catarata.

TSMC (2023). TSMC 2022 *Sustainability Report*.

Valdivia, Ana and Sánchez Monedero, Javier (2022). *Una introducción a la IA y la discriminación algorítmica para movimientos sociales*. AlgoRace. Available at: https://www.algorace.org/wp-content/uploads/2024/09/2022-11-informe-algorace.pdf (Accessed 19 November 2024).

Valdivia, Ana (2024). The supply chain capitalism of AI: a call to (re) think algorithmic harms and resistance through environmental lens. *Information, Communication & Society*, 1–17. https://doi.org/10.1080/1369118X.2024.2420021

Velkova, Julia (2021). Data that warms: Waste heat, infrastructural convergence and the computation traffic commodity. *Big Data & Society*, 3(2). https://doi.org/10.1177/2053951716684144.

Vuijk, Bart (2022). Datacenter Microsoft Wieringermeer slurpte vorig jaar 84 miljoen liter drinkwater. *Noordhollands Dagblad*. Available at: https://www.noordhollandsdagblad.nl/regio/noordkop/datacenter-microsoft-wieringermeer-slurpte-vorig-jaar-84-miljoen-liter-drinkwater/11414775.html (Accessed 18 November 2024).

Wang, Peng, et al. (2024). E-waste challenges of generative artificial intelligence. *Nature Computational Science*, p. 1–6. https://doi.org/10.1038/s43588-024-00712-6

Wiig, Alan (2015). The Urban, Infrastructural Geography of 'The Cloud'. Looking at where data moves, where it *lives*. *Medium*. Available at: https://medium.com/vantage/the-urban-infrastructural-geography-of-the-cloud-1b076cf9b06e (Accessed 14 November 2024).

Zekić-Sušac, Marijana, Mitrović, Saša and Has, Adela (2021). Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International Journal of Information Management* 58. https://doi.org/10.1016/j.ijinfomgt.2020.102074.

Zhong, Raymond and Chang Chien, Amy (2021). Drought in Taiwan Pits Chip Makers Against Farmers. *The New York Times*. Available at: https://www.nytimes.com/2021/04/08/technology/taiwan-drought-tsmc-semiconductors.html (Accessed 18 November 2024).

4. Military Accelerationism: Artificial Intelligence, Big Tech, and the Genocide in Gaza

NoTechForApartheid

The intersection between artificial intelligence and the military industry is increasingly intense, as evidenced by the fact that investment in this area is growing at a rapid pace. Much of this investment volume is commonly labelled as investment in the development of artificial intelligence, but it is often overlooked that it is part of defence-related contracts. One of the effects of this alliance between artificial intelligence and the military industry has been the dehumanisation of warfare, with the consequences that this has also had in legal terms. Two circumstances mark the progress of this relationship: on the one hand, the introduction of artificial intelligence in the industrial warfare sector itself; but, on the other, also the incorporation of consumer technology, that is, that which is not expressly developed for military use, into its use by armies.

The transfer of consumer technology to war-related uses is particularly problematic. The use of data of civilian origin in the training of models that end up in defence work, or the participation of consumer technology corporations in the development of tools dedicated to military purposes outline some of these conflicts. Together with issues related to a supposed infallibility that is not so, or a supposed objectivity plagued by biases, they build enormous mechanisms of surveillance, control and repression, for which it is necessary to configure new regulatory and social response frameworks.

Introduction

The genocidal assault on Gaza that began in 2023 is among the most destructive and deadly modern military campaigns. The daily death toll has surpassed all other major 21st century conflicts,[1] including deaths of women and children.[2] Thousands more have lost limbs from Israeli bombings, and been killed indirectly from disease and starvation. At least 90% of Gaza residents have been displaced, and over 80% of Gaza's buildings have been destroyed.[3]

Two developments have been key to this genocide: the weaponization of Artificial Intelligence (AI) throughout military, tactical, and surveillance systems; and the integration of consumer technology capacities into military operations.

The integration of consumer technology into the Israeli military has enabled the military's technical and warfighting operations to scale past what their internal systems could handle. While the Israeli military has long had ties with Big Tech companies, as well as the military and surveillance sectors, the Project Nimbus contract signed in 2021 has expanded these relationships and directly involved Google and Amazon in the mass surveillance and genocide of the Palestinian people.

Project Nimbus provisions Amazon AWS and Google Cloud services to the Israeli military and government. Cloud services have been widely used by the Israeli military.[4] Col. Racheli Dembinsky, commander of the IDF's Center of Computing and Information Systems stated:

> with the onset of the Israeli army's ground invasion of Gaza in late October 2023, […] the internal military systems quickly became overloaded […] cloud services offered by major tech firms allowed the army to purchase unlimited storage and processing servers at the click of a button […]. But the "most important" advantage that the cloud companies provided […] was their advanced capabilities in artificial intelligence.

Israeli military veteran turned activist Ori Givati says Israel's ability to process vast stores of data to surveil Palestinian people "is an integral part of the occupation".[5] This has long included AI in at least some form. "Wolf Pack," a set of surveillance systems, uses facial recognition to register and identify Palestinians, often without knowledge or consent.[6] This and other invasive surveillance systems feed into massive long-term data storage systems that provide the training data for AI-based operational intelligence systems. Since 2023, multiple AI-based systems for target selection, tracking, and combat have come to light:

- Habsora (The Gospel), is an AI system that generates targets for attack, facilitating a "mass assassination factory".[7]
- Lavender is an automated kill list, using AI to analyze data collected on most Gaza residents via mass surveillance, and mark people as targets.[8]
- Where's Daddy is an automated system used to track targets (including targets generated by Lavender), and bomb them when they enter their family home, sometimes killing entire families.[9]

1 https://www.aljazeera.com/news/2024/1/11/gaza-daily-deaths-exceed-all-other-major-conflicts-in-21st-century-oxfam
2 https://www.oxfam.org/en/press-releases/more-women-and-children-killed-gaza-israeli-military-any-other-recent-conflict
3 https://www.pbs.org/newshour/world/90-of-gaza-residents-have-been-displaced-by-israels-evacuation-orders-un-says, https://www.middleeastmonitor.com/20240820-un-over-80-of-gazas-buildings-destroyed/
4 https://www.972mag.com/cloud-israeli-army-gaza-amazon-google-microsoft/
5 https://theintercept.com/2022/07/24/google-israel-artificial-intelligence-project-nimbus/
6 https://www.amnesty.org/en/documents/mde15/6701/2023/en/, https://www.972mag.com/isdef-surveillance-tech-israel-army/
7 https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/
8 https://www.972mag.com/lavender-ai-israeli-army-gaza/
9 ibid.

- Drones and rovers with automated targeting and tracking[10]

These technologies amplify dehumanization, obfuscate accountability, and bring warfare closer to fully automated mass killing. As one officer put it,[11]

> "You're fighting from inside your laptop". In the past, "you would see the whites of your enemy's eyes, look through binoculars and see him explode." Today, however, when a target appears, "you tell [soldiers] through the laptop, 'Shoot with the tank.'"

In this chapter, we explore how AI weapons fail to live up to descriptions such as "precise" and "objective," break down Project Nimbus' true military nature and the lies around it, and illustrate how tech workers have raised concerns and faced retaliation and repression as a result of speaking out. We then critically examine AI kill lists and argue that their use constitutes war crimes.

AI Warfare

The use of AI for warfare has been expanding, particularly in the past decade. In 2015, Siemens estimated that worldwide military spending on robotics has ballooned to $7.5 billion, and projected an accelerating increase to $16.5 billion by 2025.[12] US government spending on AI is overwhelmingly military: nearly 90% of contract values were within the Department of Defense.[13]

The rise in military applications of AI is directly rooted in the expansion of surveillance apparatuses globally, and is enabled by the massive amounts of data generated by the consumer technology sector. AI is the militaries' chosen means to operationalize their surveillance data, for example analyzing drone surveillance images to find individuals, or flagging social media posts by any criteria of their choosing. AI weapons provide operators with invulnerability, are a means of psychological warfare, and appropriate consumer technology to reduce costs.

AI / ML Background

AI fills many roles in the popular imaginary and in public discourse. Fundamentally, however, AI derives from statistics, leveraging vast data sets and involving optimization of parameters from a random initialization. The most common type of model in use is known as a discriminative model, termed because it distinguishes between types of inputs, e.g. guessing whether a flower is an iris or an orchid, or which word one is most likely to type next. Each example is represented to the model as a set of numbers that describes the object. This could be "features," i.e. calculable aspects of an object such as the number of petals on a given flower, or more complex representations such as the pixels of an image. Given large enough data samples paired with labels indicating what the model should predict, the model "learns" a function that can distinguish between objects of the given label with some accuracy on the available data.

10 https://www.businessinsider.com/israel-drone-that-can-fire-a-sniper-rifle-while-flying-developed-2022-1, https://www.idf.il/en/mini-sites/technology-and-innovation/jaguar-the-idf-s-newest-most-advanced-robot/
11 https://www.972mag.com/cloud-israeli-army-gaza-amazon-google-microsoft/
12 https://web.archive.org/web/20180207122319/https://www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/autonomous-systems-infographic.html
13 https://www.brookings.edu/articles/the-evolution-of-artificial-intelligence-ai-spending-by-the-u-s-government/

To relate this to military uses of AI, the current commander of Unit 8200 described the following "features" as input to a kill list generation system:[14]

> being in a Whatsapp group with a known militant, changing cell phones every few months, and changing addresses frequently.

Current AI systems often make use of Large-Language Models (LLMs) or Multimodal LLMs. Multimodal LLMs are effectively LLMs that allow for image, audio and/or video in addition to text as inputs or outputs. Examples include OpenAI's ChatGPT, Google's Gemini, and Amazon's Bedrock. LLMs have massively expanded the use of AI across industry, but they are still effectively just massive statistical models that are trained on immense amounts of data. For example, Google's Gemini is trained on user inputs to Gemini,[15] public internet data, and possibly other private datasets. These models first learn a representation of language patterns, and then are fine-tuned on labeled data to output answers to question prompts, follow basic commands, and even perform complex tasks such as audio transcription. The datasets required for fine-tuning are often outsourced to contract workers.[16] However, because they are generating responses from statistical models of data, they are not constrained to factual, reasonable, or objective responses.

Next, we present two important failure cases in AI kill lists, that can be understood with this background.

Bad labels in the Training Data

One source from Lavender's military data science team said,[17]

> "I was bothered by the fact that when Lavender was trained, they used the term 'Hamas operative' loosely, and included people who were civil defense workers in the training dataset"

In this case, the term "civil defense worker" means people like those who recover people / bodies from rubble after bombings.[18] A model trained from this data will propagate those errors into inference, arguably violating the principle of distinction (i.e. enacting a war crime). In more subtle ways, the training dataset may be biased in terms of how the data was collected (such as what sample is chosen), labeling methodology, and input feature processing.

Dataset Distribution Errors

The models are only able to operate on the data provided to them, they are not able to interpret context outside of the training data, and cannot, therefore, account for novel circumstances or information. This is evidenced by how models almost always perform worse in real-life scenarios than they do in testing. Models are trained to minimize errors on the training set by finding correlations in the data. When the environment in which it operates shifts, the underlying correlations change and the model will become less accurate.

14 https://www.972mag.com/lavender-ai-israeli-army-gaza/
15 https://www.searchenginejournal.com/google-gemini-privacy-warning/507818/
16 https://www.engadget.com/ai/google-accused-of-using-novices-to-fact-check-geminis-ai-answers-143044552.html
17 https://www.972mag.com/lavender-ai-israeli-army-gaza/
18 https://www.euronews.com/2024/07/13/civil-defence-workers-recover-60-bodies-from-rubble-in-two-districts-of-gaza-city

Concretely, in the case of Lavender, people being bombed and forced to flee their homes will mean that model inputs change. As one Israeli source put it "In war, Palestinians change phones all the time".[19]

Military vs. Consumer Applications

In consumer contexts, AI has found the most success in products where the wrong answer is not too costly, and good guesses are very valuable, e.g. search, protein-folding, and drug discovery. To give a contrasting example, the California End of Life Option Act requires two physicians to determine whether a patient is terminally ill with 6 months or less to live and mentally competent. While there are valid debates on the level of self-determination afforded, it's obviously extremely important that the diagnosis of terminal illness be accurate, and therefore widely accepted that this is not an ethical application for AI. AI kill lists should be regarded in the latter category. We shouldn't assume they are more sophisticated than say, the ads you are shown, possibly less.

Mischaracterizations of AI Weapons

Novel AI weapons technologies are often deployed with the veneer of modernity, positioned as cutting-edge and state-of-the-art. This mirrors the language used by consumer technology brands to portray a vision of unassailable technological advancement.

Automated killings are often referred to as "precise" or "surgical" by the Israeli military[20] (and previously by the US military[21]). This characterization is contradicted by reality:[22]

> "When it came to targeting alleged junior militants marked by Lavender, the army preferred to only use unguided missiles, commonly known as "dumb" bombs (in contrast to "smart" precision bombs), which can destroy entire buildings on top of their occupants and cause significant casualties. "You don't want to waste expensive bombs on unimportant people — it's very expensive for the country and there's a shortage [of those bombs]," said C., one of the intelligence officers."

After the Oct 7 attacks, the Israeli army decided "for every junior Hamas operative that Lavender marked, it was permissible to kill up to 15 or 20 civilians." The US military, which first made significant use of drones in the Iraq war, also permitted high levels of civilian casualties in Pakistan,[23] Afghanistan,[24] Somalia,[25] and Yemen.[26]

The policy of increasing acceptable civilian casualties is driven by the nature of the technology and in the military context exposes how deeply such systems dehumanize their victims.

Automated killings are also advertised as more "objective." As one junior operative said about Lavender,[27]

19 https://www.972mag.com/lavender-ai-israeli-army-gaza/
20 https://www.cbsnews.com/news/israel-military-ground-operation-al-shifa-hospital-gaza-hamas/
21 https://journals.sagepub.com/doi/abs/10.1177/0263276411423027
22 https://www.972mag.com/lavender-ai-israeli-army-gaza/
23 https://www.newamerica.org/future-security/reports/americas-counterterrorism-wars/the-drone-war-in-pakistan/
24 https://web.archive.org/web/20180624010404/https://www.thebureauinvestigates.com/projects/drone-war/afghanistan
25 https://www.newamerica.org/future-security/reports/americas-counterterrorism-wars/the-war-in-somalia/
26 https://www.newamerica.org/future-security/reports/americas-counterterrorism-wars/the-war-in-yemen/
27 https://www.972mag.com/lavender-ai-israeli-army-gaza/

> "I have much more trust in a statistical mechanism than a soldier who lost a friend two days ago. Everyone there, including me, lost people on October 7. The machine made it coldly. And that made it easier."

AI models are still employed in a larger system that is driven by humans. In the case of Gaza, the operation is driven by revenge, hate, and arguably a desire to take land or resources,[28] e.g. in Lavender, thresholds were lowered to maximize destruction; as one officer said,[29]

> "In a day without targets [whose feature rating was sufficient to authorize a strike], we attacked at a lower threshold. We were constantly being pressured: 'Bring us more targets.' They really shouted at us. We finished [killing] our targets very quickly."

In US drone warfare, commanders have also focused on "kill counts" as a metric of operational success, and similarly mention that most strikes target low-level operatives.[30]

We can see from these examples that rather than providing precision or objectivity, the primary impact of AI-based targeting systems is to maximize destruction, and obscure the role of bias, and human decision-making in bombing operations.

Motivations for AI Weapons

Militaries seek to benefit from AI weapons systems in a variety of ways. Concretely, AI-controlled drones enable operators to launch attacks without risking their own life. This can steeply reduce the social cost of war, making it easier to maintain support for war efforts.

A bit less obviously, US drone operators become somewhat familiar with their targets, following them for a long amount of time.[31] Drones feature high resolution surveillance equipment, which allows operators to,[32]

> "see the target up close, see what happens to it during the explosion and the aftermath […] you're further away physically but you see more".

From the description of Lavender and Where's Daddy,[33] it seems that Israeli drone operators spend less time, as target selection and tracking is more automated. But their surveillance capabilities are no less, so operators have the power to kill their victims with minimal effort, or as much involvement as they choose.

As we discuss later, AI weapons also allow effective leveraging of consumer technology.

Surveillance and Terror

Surveillance technology underlies both AI kill lists and occupation activities used to maintain apartheid conditions such as checkpoint stops.

28 https://www.amnesty.org/en/documents/mde15/8668/2024/en/
29 https://www.972mag.com/lavender-ai-israeli-army-gaza/
30 https://theintercept.com/2021/10/24/drone-war-books-neil-renic-wayne-phelps/
31 ibid.
32 https://journals.sagepub.com/doi/abs/10.1177/0263276411423027
33 https://www.972mag.com/lavender-ai-israeli-army-gaza/

Tech conferences like the Israel's Defense Exposition often feature surveillance start-ups[34] promising to transform the occupation and war through greater efficiency. Israeli officials have sought to portray their occupation as "enlightened," through technical solutions that could "shrink the conflict." In practice, the number of stops were increased due to AI systems' dependence on vast amounts of data,[35]

> "under Blue Wolf, such altercations [stops] intensified. As soldiers put it, the army incentivized brigades to collect as much data as possible through petty competitions."

This pattern reveals the mutual dependence of AI and surveillance data collection. More surveillance requires AI to effectively process, while perceived benefits of AI are used to justify deeper surveillance. These twin systems aim to provide a sense of power and control to the occupying forces by enhancing their ability to see and predict potential opposition, while functioning as a form of psychological warfare against the victims of state violence. Automated systems are not totally random i.e. they are at least partially responsive to ones' actions, but are also inaccurate enough that they lead to in wrongful arrests, and potentially arbitrary detention or torture.[36] This creates of a state of terror for the oppressed—actions may trigger the automated systems, but it's unclear exactly how. Any behavior may increase the risk of arrest, including social media posts, which causes a chilling effect. In Israel, these surveillance systems find further justification in their impact on the extraction of value from marginalized Palestinian workers and civilians. Palestinian workers make up a substantial part of Israel's economy[37] and the extent of surveillance systems in place creates an already disciplined workforce for the occupation. The surveillance data collected simultaneously enriches the occupation, by providing the training data for AI systems that can be sold internationally.

Consumer Technology

AI surveillance / weapons make more direct use of consumer technology than many other types of military technologies. As such, Western militaries have become increasingly reliant on contracts with large consumer technology companies. Major cloud providers may be the only organizations that have infrastructure, capacity, and data to train state of the art AI models that can be customized for military use, and the ease of purchasing existing high-performance services, storage and computing resources is a major benefit to militaries expanding operations. As one intelligence source put it,[38]

> "[the cloud companies] also have their own STT [speech-to-text capabilities]. These are good; they have many capabilities. Why develop everything in the army unit if the capabilities already exist?"

US technology companies have a long history of interdependence with Western militaries. In-Q-Tel, founded in 1999, is a non-profit venture capital firm created to channel funds from the CIA to the private sector, and transfer these technologies back to US intelligence and military agencies. Among these firms are a growing number of social media mining and surveillance companies, including Dataminr, Geofeedia, PATHAR, and TransVoyant.[39]

34 https://www.972mag.com/isdef-surveillance-tech-israel-army/
35 https://www.cambridge.org/core/journals/international-journal-of-middle-east-studies/article/algorithmic-state-violence-automated-surveillance-and-palestinian-dispossession-in-hebrons-old-city/80B9C5192057ACEA17089E488CFC1486
36 https://reliefweb.int/report/occupied-palestinian-territory/welcome-hell-israeli-prison-system-network-torture-camps-enhe
37 https://www.nbcnews.com/news/world/israel-hamas-india-labor-shortage-migrant-workers-rcna135603
38 https://www.972mag.com/cloud-israeli-army-gaza-amazon-google-microsoft/
39 https://theintercept.com/2016/04/14/in-undisclosed-cia-investments-social-media-mining-looms-large/

Project Maven, initiated in 2017, represented a shift in how the US Department of Defense was integrating with consumer technology companies and marks a steep turn towards prioritization of AI. The Project Maven contract enlisted labor from private-sector engineers in training AI for object identification from military surveillance data, and has involved at least 21 private companies.[40] However, in 2018, Google elected not to renew the contract due to worker protests. Following this point of friction for the sale of private-sector products, two notable trends emerged within the tech industry: tech companies made public statements of AI ethics, and began to implement severe measures to constrain workers' power to impact company decisions.[41]

Three years after workers protested Project Maven, Project Nimbus was signed. The following year, the Joint Warfighting Cloud Capability (JWCC) contract, a $9B contract for AI military solutions, awarded to Amazon Web Services, Google, Microsoft, and Oracle was announced.[42] JWCC is a very similar contract to Project Nimbus in that it provisions cloud services with specific conditions that enable the technologies to be used for warfighting. While more direct extractions of engineering labor met with worker resistance and reputational damage, this evolution of the contract has proven resilient—it masquerades as a simple sale of consumer services, granting militaries access to the accumulated labor of private sector engineers, and allowing tech companies to access military spending budgets while maintaining the veneer of impartiality.

Furthermore, the perception that Project Nimbus is a standard cloud computing and consumer AI contract has allowed Google and Amazon to obfuscate the contract's military implications, and avoid major scrutiny while still facilitating advanced warfare use of their technology.

Project Nimbus

Project Nimbus is a flagship project by the Israeli government to provision cloud infrastructure and services. Google and Amazon were chosen as cloud providers in a $1.2B contract, in 2021. Despite many misleading statements put out by Google, the contract is primarily military in nature.

It was recently revealed that at least 70% of Google's expected revenue from Project Nimbus is coming from the Israeli military.[43]

> "Under the terms of the deal, Google expected to get the largest share of money from Israel's Ministry of Defense, an estimated $525 million from 2021 to 2028, which dwarfed the $208 million it expected to receive from the rest of the country's central government."

Note: the Ministry of Defense oversees the IDF, Israel Military Industries (IMI), and Israel Aerospace Industries (IAI).

Furthermore, the military value of Project Nimbus has been attested to by various IDF members. Notably:

- Military sources state that "surveillance of all Palestinian residents of Gaza is so large that it cannot be stored on military servers alone".[44] Yossi Sariel, current commander of Israeli military's Unit 8200 (which develops Lavender), stated information of such scope can be stored "only in companies such as Amazon, Google, or Microsoft".[45]

40 https://en.wikipedia.org/wiki/Project_Maven
41 https://www.businessinsider.com/google-thanksgiving-four-trial-protest-2021-8
42 https://harpers.org/archive/2024/03/the-pentagons-silicon-valley-problem-andrew-cockburn/
43 https://www.nytimes.com/2024/12/03/technology/google-israel-contract-project-nimbus.html
44 https://www.972mag.com/cloud-israeli-army-gaza-amazon-google-microsoft/
45 ibid.

- Two of Israel's leading state-owned weapons manufacturers, Israel Aerospace Industries and Rafael Advanced Defense Systems, are required to use Amazon and Google, through Nimbus, for cloud computing needs.[46]

- Gaby Portnoy, head of Israel's National Cyber Directorate, stated: "Phenomenal things are happening in battle because of the Nimbus public cloud, things that are impactful for victory…"[47]

- At the IT for IDF 2024 conference, Colonel Racheli Dambinski stated that cloud computing is "a weapon in every sense of the word",[48] and explicitly described the IDF's use of Google, Amazon, and Microsoft clouds.[49]

- Google Cloud CEO Thomas Kurian announced Vertex AI integration with Palantir,[50] the "AI arms dealer of the 21st century".[51] Google also provides Nimbus users with access to Palantir's Foundry software.[52]

While we don't have evidence directly linking Project Nimbus to Lavender, it is a technical fit, in the sense that it uses mass surveillance data, which is often processed with cloud AI. The Israeli government intends to migrate government projects to the cloud;[53] while the most secret projects are kept on military servers, some intelligence projects are on the cloud.[54] The limiting factor to date has been the lack of fully secured data centers to run the most sensitive operational systems, however there are pending contracts to enable this type of processing under similar contracts with consumer technology companies. Google is also pitching its Gemini LLM to Israeli police and national security officials.[55] Among other uses, text models can enact harm in the form of social media surveillance, which is already extensively used as a basis to arrest Palestinians.[56]

Obfuscating Corporate Complicity

Project Nimbus represented a major turning point for Google Cloud business, and an opportunity for Google to make gains in Cloud market share due to the high percentage of government budgets that are spent on military AI. It is also a turning point in how big tech cloud service providers in general are able to turn themselves into military contractors, but not without cost to their reputation and workplace culture.

In pursuit of profit, Google has engaged in severe worker repression and public deception to avoid responsibility for the harm their products are facilitating. When questioned about Project Nimbus, Google has repeated two specific claims:[57]

> "This work is not directed at highly sensitive, classified, or military workloads relevant to weapons or intelligence services."

46 https://theintercept.com/2024/05/01/google-amazon-nimbus-israel-weapons-arms-gaza/
47 https://www.wired.com/story/amazon-google-project-nimbus-israel-idf/
48 https://www.404media.co/google-cloud-listed-then-removed-as-sponsor-of-israeli-military-tech-conference/
49 https://www.youtube.com/watch?v=qLBDfnZJrC8
50 https://jackpoulson.substack.com/p/microsoft-and-google-have-been-working, https://cloud.google.com/blog/topics/partners/google-cloud-and-palantir-announce-annalytics-partnership
51 https://www.theverge.com/2024/8/8/24216215/palantir-microsoft-azure-ai-defense-partnership-surveillance
52 https://theintercept.com/2024/05/01/google-amazon-nimbus-israel-weapons-arms-gaza/
53 https://www.gov.il/en/pages/press_24052021
54 https://www.972mag.com/cloud-israeli-army-gaza-amazon-google-microsoft/
55 https://www.wired.com/story/amazon-google-project-nimbus-israel-idf/
56 https://www.adalah.org/en/content/view/10959
57 https://www.wired.com/story/amazon-google-project-nimbus-israel-idf/, https://theintercept.com/2024/05/01/google-amazon-nimbus-israel-weapons-arms-gaza/, https://time.com/7013685/google-ai-deepmind-military-contracts-israel/

and,[58]

> "The Nimbus contract is for workloads running on our commercial cloud by Israeli government ministries, who agree to comply with our Terms of Service and Acceptable Use Policy."

In an email statement, Google claimed that the normal Cloud Terms of Service (TOS) and Acceptable Use Policy (AUP) were explicitly referenced.[59] A review of the Nimbus tender reveals this to be false; in fact, Nimbus users are only bound to an Adjusted Terms of Service.[60] The tender affirms that all services, including advanced AI, must be made available to all branches of government. The Israeli Ministry of Finance refused to share these adjusted terms of service.[61]

Furthermore, Google has refused to enforce its terms of service. Israeli military intelligence created a 'hit list' of alleged militants after the October 7th attacks, using facial recognition via Corsight (which hires workers familiar with Amazon and Google Cloud[62]) and Google Photos. Suspects, including those misidentified, were detained and abused.[63] Despite this harm being a breach of Google's acceptable use policies for Photos, Google refused to enforce its policies.[64]

Nimbus is far from the first time Google has lied to its workers. Google claimed that its Project Maven technology was not used to identify people, and that nothing about the AI was custom. However, code revealed that people and vehicles were in fact labeled. Contract workers were paid to label the satellite imagery, without being informed of its military purpose.[65] Google lied about its "Dragonfly" search project being "in exploration stages".[66] Thomas Kurian lied that Google Cloud would not be used on the southern border;[67] it was later revealed that CBP used Google Cloud to process border surveillance tower imagery.[68]

The obfuscation of military contracts for cloud computing technology and consumer AI uses allows companies to claim upholding ethical commitments while still facilitating advanced warfare use of their technology. Google's AI Principles were rolled out in 2018 in the wake of Project Maven protests (arguably to placate workers),[69] and have proliferated throughout the industry, e.g. at Amazon[70] and Microsoft.[71]

They are effective PR, but are designed to be toothless in actually preventing their products from facilitating harm. Google employs the people doing AI Principles reviews, and can push them out if they are a hindrance. In 2022, a Google spokesperson told DefenseOne that the company's AI principles,[72]

> "apply to custom AI work, not general use of Google Cloud services… It means that our technology can be used fairly broadly by the military"

58 ibid., https://theintercept.com/2024/05/01/google-amazon-nimbus-israel-weapons-arms-gaza/, https://www.nytimes.com/2024/12/03/technology/google-israel-contract-project-nimbus.html
59 https://www.wired.com/story/amazon-google-project-nimbus-israel-idf/
60 https://theintercept.com/2024/12/02/google-project-nimbus-ai-israel/
61 https://bsky.app/profile/sambiddle.com/post/3lcxu3ty7xc2h
62 https://archive.ph/HvcJ6#selection-957.0-957.71
63 https://www.nytimes.com/2024/03/27/technology/israel-facial-recognition-gaza.html, https://mondoweiss.net/2024/01/the-shocking-inhumanity-of-israels-crimes-in-gaza/
64 https://theintercept.com/2024/04/05/google-photos-israel-gaza-facial-recognition/
65 https://theintercept.com/2019/02/04/google-ai-project-maven-figure-eight/
66 https://theintercept.com/2018/08/17/internal-meeting-reveals-how-google-bosses-misled-staff-on-their-china-censorship-plan-here-are-the-questions-they-must-answer/
67 https://www.cnbc.com/2020/10/30/google-cloud-ceo-kurian-to-employees-not-working-on-border-wall.html
68 https://theintercept.com/2022/07/24/google-israel-artificial-intelligence-project-nimbus/
69 https://www.wired.com/beyond-the-beyond/2018/06/googles-ai-principles/
70 https://sustainability.aboutamazon.com/human-rights/principles
71 https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl
72 https://www.defenseone.com/technology/2022/06/new-google-division-will-take-aim-pentagon-battle-network-contracts/368691/

Many consumer AI infrastructure products can be used/reused as part of a military's "kill chain", sometimes further obfuscated by enlisting a third-party vendor.

Most employees who join a company like Google believe that they are working on a consumer product such as Google Search, Photos, or Cloud. Unfortunately, the decisions of leadership has made workers implicated in Israel's military occupation and genocide. While greater responsibility and culpability lies with leadership, it is a simple truth that engineers' labor has been weaponized. Because Google gives the Israeli military the full suite of Google Cloud enterprise products through Project Nimbus, many Google workers have justifiable reason to fear that their labor is a cog in Israel's genocidal assault on Gaza.

Google workers have tried to voice their concerns through many avenues. This includes "proper" channels such as internal reporting tools, escalating to management, emailing executives, raising questions during town hall meetings,[73] and internal petitions, including to Google's human rights program. Despite these efforts, Google leadership has consistently refused to engage, downplaying and denying worker concerns,[74] lying about the contract's military connections, and demonstrating a clear bias against Palestinian and ally voices within the company.[75]

Faced with the company's hostile response, workers formed the No Tech for Apartheid campaign,[76] distributed petitions signed by over 1000 employees,[77] led protests in 2022,[78] protested the Cloud Next conference,[79] and the Israel-focused Mind the Tech conference.[80] In April 2024, workers held simultaneous office sit-ins in New York and Sunnyvale, and public rallies.[81]

Google executives had 9 workers arrested and fired at least 51 workers.[82] Additional workers resigned in protest over the firings.[83] This is a particularly egregious case of a pattern reflected throughout the tech industry, including Microsoft which also sells cloud technology to the Israeli military, and retaliated against workers who organized a vigil for martyred Palestinians.[84]

Just as Google provides technology for Israeli surveillance and military operations, it facilitates a hostile workplace environment for Muslim, Arab, and Palestinian workers.[85] The dismissal of Palestinian voices in particular is a violation of human rights due diligence guidelines laid out by the OHCHR.[86]

Google's repression of worker voices, deception of the press and the public, and ongoing backroom military dealings indicate the steep lengths that big tech companies are willing to take to ensure they can profit from war and genocide. They benefit from the perception of their AI as cutting edge, even though the use-case involves killing children.

73 https://time.com/7013685/google-ai-deepmind-military-contracts-israel/
74 https://www.middleeasteye.net/news/project-nimbus-israel-apartheid-google-amazon-protests
75 https://theintercept.com/2023/11/15/google-israel-gaza-nimbus-protest/
76 https://www.instagram.com/notechforapartheid/
77 https://www.instagram.com/jewishvoiceforpeace/p/CVQb8CWpMj7/?img_index=1
78 https://www.forbes.com/sites/richardnieva/2022/09/09/google-and-amazon-protest-project-nimbus-ai-contract-israel/?sh=45d147e5d162
79 https://www.latimes.com/business/story/2023-08-29/google-cloud-employees-protest-israeli-military-contract
80 https://time.com/6964364/exclusive-no-tech-for-apartheid-google-workers-protest-project-nimbus-1-2-billion-contract-with-israel/
81 https://www.latimes.com/business/story/2024-04-16/google-israel-sit-ins-project-nimbus
82 https://time.com/6964364/exclusive-no-tech-for-apartheid-google-workers-protest-project-nimbus-1-2-billion-contract-with-israel/, https://apnews.com/article/google-israel-protest-workers-gaza-palestinians-96d2871f1340cb84c953118b7ef88b3f
83 https://www.jpost.com/arab-israeli-conflict/gaza-news/guardian-of-the-walls-the-first-ai-war-669371
84 https://apnews.com/article/microsoft-fired-workers-israel-palestinians-gaza-72de6fe1f35db9398e3b6785203c6bbf
85 https://medium.com/@notechforapartheid/googleopenletter-868f0c4477db
86 https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

Workers have and continue to stand up for their own rights in not having their labor enlisted in genocide but face severe repression. Worker protest can be powerful, because it leverages their stake in the impacts of their labor, and raises the possibility of disruption at the site of complicity. Avenues for mitigating the harm of AI-based weaponry should include improving labor protections, so that workers can stand in solidarity with one another.

Morality & Legality

Systems like Habsora, Lavender, and Where's Daddy evoke horror, and we should affirm that basic revulsion to wanton killing. But these concerns are also sometimes dismissed with adages of "war is hell", and to see more clearly, we have to examine the system in more depth. We attempt to situate AI weapons in a broader context of warfare, and explore where AI can obscure accountability, but still be in violation of existing laws.

Situating the AI weapons

AI kill lists are a manifestation of power asymmetry and racial dynamics. All wars can be terrible, but in other cases, one faction might claim to represent the people, or hope to assimilate them. In the context of the genocide in Gaza, these dynamics are illustrated by Yoav Gallant's comment "we are fighting human animals".[87] Israel's racial segregation and apartheid conditions entrench Palestinians' status as inferior. The ease and speed of killing with AI are design manifestations of this disposability of life, and we believe their use to be a violation of current laws and an area that merits new regulations.

Consider These Aspects of AI Weapons:

- Lavender effectively ranks the entire population of Gaza, through mass surveillance, to prioritize who is valuable to kill first.
- The drones used in drone warfare are not cutting-edge in terms of speed, range, or stealth.[88]

They arguably embody a brutal approach to counterterrorism, where Israeli forces themselves terrorize an overwhelmingly civilian population for a long period of time. Actual drone wars are asymmetrical conflicts, wherein advanced weapons can be leveraged to devastating effects against people without access to such.

While some AI-based weapons, such as drones, may be inevitable in the context of "arms races" and infeasible to regulate, there is a distinct difference between systems that are directed by human intelligence and/or credibly necessary to avoid battlefield defeat, and systems like Lavender that replace human decision-making. Systems like Lavender can and should be banned.

87 https://www.youtube.com/watch?v=ZbPdR3E4hCk

88 https://en.wikipedia.org/wiki/General_Atomics_MQ-1_Predator#Specifications, https://en.wikipedia.org/wiki/General_Atomics_MQ-9_Reaper#Specifications

Comparison with on-AI Intelligence Work

It's natural to ask if AI kill lists are worse than decisions by human intelligence officers. Both of these situations can be immoral, or from a legal point of view, can involve war crimes such as violations of distinction and proportionality. In this case, however, it appears that the AI system is actually worse in accuracy, structure, and overall harm. One of the key ways it increases destruction is because it is faster. As a former IDF chief of staff said about Habsora,[89]

> "You see, in the past there were times in Gaza when we would create 50 targets per year. And here the machine produced 100 targets in one day"

Both Habsora and Lavender officers are ordered to identify and kill targets as quickly as possible. Part of this speed and scale comes from lowering thresholds of who is considered a target, in order to meet demands. This, and a level of effective randomness to meet a "murder budget" / "collateral damage budget", should constitute war crimes.

We present our critique not to absolve human intelligence officers, but to argue that these AI systems introduce new violations. The converse is just as important: AI-related war crimes should not be absolved by the threat of the US or Israel committing non-AI war crimes, if somehow deprived of access to compute resources.

Military AI Accountability and Explainability

AI kill lists also present new challenges in accountability and explainability. They are a product of a more diverse set of factors, including,

- Military data science officers, who label individuals as militants or civilians
- Army intelligence officers, who make the final decision to kill, but are ordered to trust the AI
- Corporate executives, who provide civilian technologies for military use
- Tech workers, who develop underlying surveillance technologies, but often for another purpose. Often user data is used to train or improve these models, without informed consent.

Each of these parties is arguably less involved than a pre-AI intelligence officer, who would inspect all information on a person and decide whether they are a militant. AI predictions are difficult to interpret, particularly in ways illuminate the nature of their flaws. We believe that existing legal standards should still be applied to some of these parties, but hold that new laws are also warranted. Decisions by all of these parties affect who is killed, and the scope of legal liability on all of them will affect how much pressure there is against this system.

How Ai Weapons Concentrate Power

The effective automation of intelligence work and kill list generation means that power becomes even more concentrated, and the process of killing more mundane. US drone operators have said,[90]

89 https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/
90 https://theintercept.com/2021/10/24/drone-war-books-neil-renic-wayne-phelps/

> … “Watching the son of the person I just obliterated with a Hellfire missile pick up the pieces of his father. It wasn’t the act of killing I focused on, it was watching the boy’s face and interactions with the rest of his family that continue to haunt me”

In response, military insiders have called for greater social distance between operator and target. AI kill lists may create this distance, making wars even more deadly. We shouldn’t exaggerate the value of human officers’ guilty conscious, but we can still be concerned about its relatively sudden and complete elimination.

AI kill lists also have the potential to concentrate power even further. They could easily be designed such that changing thresholds for who is considered a militant requires little oversight.

Distinction and Proportionality

AI kill lists violate the concept of distinction, which requires that parties in armed conflict distinguish between civilians and military objectives. Many aspects are fundamental limitations of AI models.

Legal Standards

While by no means the only legal standards, two principles of international humanitarian law (IHL) are particularly relevant to AI weapons systems:

- Distinction states that parties in armed conflict have an obligation to distinguish between civilians and military objectives and only direct attacks against military objectives.
- Proportionality is a prohibition of attacks in conflict that expose civilians to risk that would be excessive in relation to the military advantage received.

Systems like Lavender violate these norms in several ways.

Mechanics of AI Kill Lists

AI kill lists fundamentally combine “fuzzy” input signals, none of which positively or conclusively identifies a militant. To be concrete, these input signals include names and behavior patterns in Lavender,[91]

> “sources explained that the Lavender machine sometimes mistakenly flagged individuals who had communication patterns similar to known Hamas or PIJ [Palestinian Islamic Jihad] operatives — including police and civil defense workers, militants’ relatives, residents who happened to have a name and nickname identical to that of an operative, and Gazans who used a device that once belonged to a Hamas operative.”

There is no rationalization for which combination of features justify killing someone.

Phttps://www.972mag.com/lavender-ai-israeli-army-gaza/

Minimal Oversight

Officers approving drone killings are instructed to adopt Lavender's kill lists, and only check that a Lavender-marked target is male,[92]

> One source stated that human personnel often served only as a "rubber stamp" for the machine's decisions, adding that, normally, they would personally devote only about "20 seconds" to each target before authorizing a bombing — just to make sure the Lavender-marked target is male

Clearly, a person being male is not sufficient grounds to distinguish military targets from civilians. Combined with AI kill list errors which we discuss below, it violates distinction.

Semi-arbitrary Killing

Events such as the My Lai massacre are widely recognized as violations of distinction. Similar designations should be made for high-tech warfare. Let's say an intelligence officer identified 35 targets as "surely military", and then another 50 which he wasn't sure—perhaps there was some evidence but not enough. If he threw in all 50, it wouldn't look like intelligence work. So he rolls a die on each of the 50, and includes the ones where the die is a 6. Was there a violation of distinction, even if it turns out most of the targets were military? We would argue so.

AI kill lists operate similarly. Just because they are not totally random, doesn't mean there isn't some randomness. Any nontrivial model has errors, and if we study the training process, those errors are closer to resembling "filling a murder budget" than a human officer reaching the wrong conclusion. We should not be thrown off by their correctly classifying some militants and civilians, to recognize that labeling some arbitrary set of targets they are "unsure" of as militants, when there is not reliable evidence, is a war crime.

Calling Into Question Stated Accuracy Numbers

When Israeli intelligence officers describe Lavender as "90 percent accurate", they are referring to the proportion of correct classifications on an evaluation dataset. In the case of Lavender, officers described it as,[93]

> [we] "manually" checked the accuracy of a random sample of several hundred targets selected by the AI system

This is not a posthumous investigation into those killed, and should not be interpreted as indicative of real-world accuracy. Due to dataset distribution changes, it is an established phenomenon that measured accuracy plummets in real-world situations. Because this evaluation dataset is held by the military, it cannot be validated by any neutral or reliable sources.

Consider the intelligence quote we mentioned before, "In war, Palestinians change phones all the time". As changing phones more is probably correlated with being a militant, this means model failures will result in civilian killings. AI models are incapable of integrating basic contextualizing facts, or any information not represented in its input, e.g. "area X is a destination for refugees, so of course we expect people there to have moved." This is a fundamental limitation of any AI kill list that inevitably produces failures of distinction.

92 ibid.
93 ibid.

Decreasing Thresholds to Maximize Killing

We covered earlier how thresholds are decreased in order to maximize death and destruction. The mechanic of decreasing thresholds whenever Israeli officials demand more targets has the effect of ranking the entirety of Gaza's population for bombing. Where's Daddy intentionally tracks targeted individuals to their home, often resulting in bombings killing entire families. Israeli officers claim it is easier to bomb them when they are home. As we mentioned before, targets are often alleged to be low-level operatives, presumably meaning the military advantage in killing them is marginal. This clear violation of proportionality is directly enabled by the use of AI systems.

Challenging Obfuscation

Earlier, we mentioned how the the military data science team mis-labeled civil defense workers as Hamas operatives. The technology has somewhat obfuscated this violation, since one party has mis-labeled individuals, clearly for killing them, but they have not literally made that decision. The officers approving strikes relied on the other's faulty information, via the AI kill list. At the end of the day, this results in attacks directed against civilians.

Conclusion

AI systems are being used to enable horrific levels of destruction in Gaza. When powerful nations attack groups they view as disposable, accuracy will never be the priority. The incentives are aligned with killing people quickly and cheaply, and we should reject farcical claims that further development will fundamentally change those priorities.

Human rights organizations have already raised concerns of an "accountability gap" with lethal autonomous weapons.[94] IHL requires that individuals be held legally responsible for war crimes and breaches of the Geneva conventions, but when decisions are based in AI, it's not clear that the developers and purveyors will be held accountable.

This problem is further intensified in the case of consumer technology. While the creators of explicitly militarized AI applications may have more easily identifiable guilt, the extent of the use of consumer technologies via cloud services is intentionally obfuscated, as evidenced by Google's lies. However, it is clear these companies knowingly sell their technology to militaries committing war crimes, and that the technologies play a key role in enabling those crimes. Executives at Google, Amazon, and Microsoft are deeply implicated, yet are thus far able to profit while avoiding legal culpability. Their choices implicate tech workers in this deadly business, creating sector-wide complicity while also pointing to a site for resistance.

AI weapons are in reality not "precise" or "objective", and their deployment should constitute war crimes. Workers should have a legal right to oppose their complicity in war crimes, as well as other human rights violations such as mass surveillance and apartheid. As tech companies' repression has intensified, further labor protections are warranted.

94 https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons

5. Race and Resistance: Unpicking the Political Economies of Artificial Intelligence

Sarah Chander
Equinox Initiative for Racial Justice

Whilst mainly lauded as progressive developments, rollouts of AI and digitalisation processes more generally interact with pre-existing systems of structural racism, intersecting systems of oppression, capitalism, securitisation, militarism and extraction. This chapter explores how AI interacts with structural racism starting from manifestations of racial discrimination on policing, migration control and welfare, and then broadening to political economies of extraction, exploitation, criminalisation and digital warfare as central features of the AI industry. While, there is no "quick fix" to undo centuries of systemic racism and discrimination, this chapter charts a journey through various contestations to racialised AI systems and ends by offering some avenues of meaningful resistance, including building power amongst affected communities, disruption and abolitionist approaches, but also redistribution of resources away from systems of surveillance and control and toward community care and social provision.

Introduction

Artificial Intelligence (AI) and automated decision-making systems are increasingly developed, tested, procedured and deployed in numerous areas of public life. Whilst mainly lauded as progressive developments, rollouts of AI and digitalisation processes more generally interact with pre-existing contexts of structural racism, intersecting systems of oppression, capitalism, securitisation, militarism and extraction. This chapter explores how AI interacts with structural racism, and the various ways we can challenge and resist.

Specifically, AI is increasingly used in a variety of sectors that already involve disproportionate harm, discrimination and violence to racialised people. From surveillance and discriminatory decision-making in policing, to the testing of new population management tools on migrants by immigration authorities, to the racialised risk assessment processes profiling racialised people as 'suspicious' and criminal, there are endless intersections between technology deployment and structural racism.

The chapter makes a broad distinction between two ways artificial intelligence and technology connect to the topic of racism. The first, explored in section I, relates to how the use of AI systems leads to disproportionately harmful impacts for racialised people and communities. Section II explores a racialised political economy of AI more broadly – beyond individual or community impact looking at the ways in which AI systems fit into broader systems of racialised oppression, population management, extraction and production.

The use of data-driven systems to surveil and provide a logic to discrimination is not novel. The use of biometric data collection systems such as fingerprinting have their origins in colonial systems of control. The use of biometric markers to experiment, discriminate and exterminate was also a feature of the Nazi regime. This chapter does not dive deeply into these racial, colonial and militaristic trajectories technology and AI development. Following AI's precedents such as the colonial origins of fingerprinting and racialised surveillance practices such as the lantern laws,[1] to more recent 'innovations' in modern day artificial intelligence in US military, we see that technology has always played a role in racialised formations, policing and population management.

As such, AI, is by very nature a racialised concept and construction, and efforts to 'fix' racial harms will often fall short if not embedded in deeper, structural and political strategies that link to racial and social justice, decriminalisation, redistribution and, ultimately, decolonisation. We explore those responses, both short term and long term, reformist and abolitionist, institutional, legislative and resistance-focused, in section III.

This chapter takes Europe as an entry point – being where the author resides and organises. However, the chapter will make an inherent connection between the global extraction processes, technology developments and impact from the Global South to the North.

AI and Structural Racism: in Context

In Europe and around the world, AI systems are used to monitor and control us in public spaces, predict our likelihood of future criminality, "prevent" migration, predict our emotions, and make crucial decisions that determine our access to public services, like welfare. The development and use of AI in these areas specifically can be seen as the centre stage for any analysis on AI and structural racism – it is in these areas where structural racism most manifests, even without a digital component. This section demonstrates how AI will only feed this reality of structural racism with more tools, more legal powers, and less accountability and transparency for agents of the state with influence over the safety and wellbeing of racialised people.

1 It refers to a legislation imposed in New York City in the 18th century, which obliged black, mestizo or indigenous people walking at night and unaccompanied by a white person to carry a lantern to make themselves visible.
Browne, S. (2015) *Dark Matters: On the Surveillance of Blackness*

Policing and Security Agencies

The growing use of AI in policing and migration contexts has huge implications for racial discrimination. In policing, AI systems allow for new and more invasive techniques for surveillance and control, 'hardwiring' discrimination.[2] From the use of facial recognition to identify people as they freely move in public places to predictive policing systems to decide who is a criminal before they commit crimes, AI unveils the possibility for governments to conduct surveillance and infringe on freedoms in new, harmful ways. The deployment of such technologies exposes people of colour to more surveillance, more discriminatory decision-making, and more harmful profiling.

AI-based mass surveillance[3] is one of a spectrum of techniques that police, local authorities and companies deploy to identify people in public places. Grouped as 'remote biometric identification' systems, these are the techniques by which law enforcement uses AI to identify people, either by capturing facial images ('facial recognition'), or by using a range of other methods, such as gait and voice, to infer identity. Often these techniques combine with 'biometric categorisation' systems, designed to infer characteristics or behaviour on the basis of categorising biometric features. For example, biometric categorisation systems have been deployed to infer gender, race and other sensitive characteristics, or to characterise people as 'suspicious' often using highly racialised proxies.

Whether mass surveillance technologies are deployed for the purpose of identification, categorisation, emotion recognition or otherwise, they will perpetuate structural racism. In particular, facial recognition and other biometric mass surveillance systems have demonstrably facilitated the over-policing of racialised communities and localities where they live.[4] Combined with highly racialised and classed assumptions and skewed data of where and by whom crime happens, we have seen remote biometric technologies disproportionately deployed in areas where racialised people live. For example, a study commissioned by European Digital Rights demonstrated numerous examples of the disproportionate roll out of 'biometric-ready' surveillance cameras in areas close to places of worship and LGBT+ venues. Further, these techniques are often specifically deployed to surveil migrants. As investigated by Hermes Center,[5] the Italian Ministry of Interior's purchased in 2017 "SARI", a facial recognition system to be used in a number of contexts, including at demonstrations. As discovered in 2019, Hermes Center found a disproportionate use of the system on migrants, fuelling a highly racialised climate of suspicion, over-surveillance and criminalisation of migrants.[6]

Another display of racialised AI policing infrastructure is the increased resort to predictive policing systems across Europe. 'Predictive policing' refers to systems that profile people and areas, predict supposed future criminal behaviour or occurrence of crime, and assess the alleged 'risk' of offending or criminality in the future.[7] Ranging from systems such as the "top 400" and "ProKid" in the Netherlands that conduct risk assessments of individuals to score for likelihood of future criminality, to systems such as "Delia" in Italy that specifically risk assesses areas or location,[8] predictive policing systems are attempts to predict crime patterns to inform the allocation of policing resources. As such, the impacts of the results

2 Williams, Patrick y Kind, Eric (2019). *Data-driven policing: the hardwiring of discriminatory policing practices in Europe* https://www.statewatch.org/media/documents/news/2019/nov/data-driven-profiling-web-final.pdf

3 EDRi (2021) *The rise and rise of biometric mass surveillance in the EU*: https://edri.org/wp-content/uploads/2021/11/EDRI_RISE_REPORT.pdf

4 EU Fundamental Rights Agency (2019) *Facial recognition technology fundamental rights considerations in the context of law enforcement.*

5 Riccardo Coluccini (2017) Italian police has acquired a facial recognition system: https://medium.com/@ORARiccardo/italian-police-has-acquired-a-facial-recognition-system-a54016211ff2

6 https://www.wired.it/attualita/tech/2019/04/03/sari-riconoscimento-facciale-stranieri/

7 European Digital Rights (EDRi) (2022) *Prohibit predictive policing and profiling AI systems in law enforcement:* https://edri.org/wp-content/uploads/2022/05/Prohibit-predictive-and-profiling-AI-systems-in-law-enforcement-and-criminal-justice.pdf

8 Fair Trials (2021). *Automating Injustice: the use of artificial intelligence and automated decision making systems in criminal justice in Europe* https://www.fairtrials.org/app/uploads/2021/11/Automating_Injustice.pdf

of such systems are more encounters with law enforcement, which can result in increased racial profiling and stop and search, immigration enforcement checks with a potential to lead to deportation, discriminatory arrest, prosecution and sentencing, and even potentially instances of racist police brutality.

Predictive policing systems disproportionately impact racialised people, people who are perceived to be potential migrants, terrorists, poor and working-class people, in poor, working-class areas. As these systems are predicated on existing policing data and practices that already display discriminatory patterns against racialised people, migrants and poor people, the results of deploying predictive policing systems are therefore inherently discriminatory. This can manifest in many ways. In some cases, such systems explicitly use ethnicity data or proxies of such as a factor in algorithmic crime prediction. The organisation Fair Trials, points us to several examples of this, including, the aforementioned "Delia" system in Italy uses ethnicity data as part of predictions, and the Catalonian Department of Justice deployment of 'RisCANVI' used data about nationality.[9]

In many cases, predictive policing systems (regardless of the specific categories of the data used) result in a substantial over-representation of racialised people presented as a 'high risk' of future criminality. The Dutch systems "Top-400" and "Top-600" generated databases of people who were likely to commit crimes in the future or reoffend, which resulted in intense over-surveillance by police and social security actors, unannounced home visits, monitoring and following, and other informal sanctions such as informing employers and other social circles. These lists disproportionately included Dutch-Moroccan men and boys, as well as other racial minorities, homeless people, and people from low-income families.[10] Crucially, it is important to remember that predictive policing systems are designed to 'predict and prevent' crime, and as such, creates a form of 'pre-crime' status by which people feel the impact of policing and surveillance, even though they have not yet committed the crime in question. Beyond the concerns of racialised surveillance and suspicion and the criminalisation of the poor, these systems also are part of the broader erosion of the presumption of innocence.

Migration

Inherently connected to the role AI plays in racialised policing is the use of AI in migration 'management' and border control. AI systems are increasingly being developed to track, control and monitor migrants in new and harmful ways. From AI lie-detectors and AI "risk profiling" used in a multitude of immigration procedures to the rapidly expanding tech surveillance at Europe's borders, AI systems are increasingly a feature of the EU's approach to migration.

One of the clearest manifestations of the racialised nature of AI usage in migration is the resort to predictive systems. From individualised risk-assessment to broader predictive analytic systems used to forecast migration patterns and deploy immigration enforcement to those areas. Individualised risk assessment systems, such as the European Travel Information and Authorisation System (ETIAS), which enables profiling (and potentially the use of AI) to categorise travellers into pre-defined risk profiles related to purported migration, security or public health risks. This profiling takes place with a number of factors, including historical data on rates of over-staying or refusal and information provided by Member States as to security risks.[11] Such risk assessment systems profile people based on predetermined risk indicators embedded in screening rules, parameters which are often opaque and not made public. Such systems are discriminatory by nature - they codify assumptions about the link between personal data and characteristics with particular risks. People are not judged on individual behaviour or on factors within their control, but rather by pre de-

9 bid.
10 https://controlealtdelete.nl/articles/top400-aanpak-is-discriminerend#gsc.tab=0
11 Vavoula, N. (2020) *The Commission Package for ETIAS Consequential Amendments – Substitute Impact Assessment,* 20–30.

termined characteristics, such as nationality.[12] As such, the increasing use of automated or AI-based risk assessments in migration control is therefore a way to objectivise patterns of racialised suspicion as a way to prevent and manage migration, systematising a process of generalisation about people before they have moved. Although not detailed here, the increased resort to the use of AI in individual case management to cast doubt on the veracity of immigration claims also forms part of this trend, such as the use of emotion recognition (the use of AI systems to infer or assess emotions, based on biometric data), AI polygraphs and AI-based dialect recognition systems.

Beyond individual decision making, we are seeing the investment in AI as part of an ever-expanding, generalised surveillance apparatus. This includes AI for surveillance at the border and predictive analytic systems to forecast migration trends. According to the Border Violence Monitoring Network, governments and institutions such as Frontex may be using AI technologies to facilitate pushbacks of migrants, often amounting to forced disappearances. The organisation EuroMed rights has documented the increased use of artificial intelligence deployed at the EU's external borders, funded as part of colonial externalisation 'partnerships' (agreements between the EU and non-EU member states mandating migration management techniques, often in exchange for aid) coopting non-EU states into the EU's preventative migration regime.[13] In this context, there is a real danger that seemingly innocuous forecasts about migration patterns will be used to facilitate push-backs, pull-backs and other ways to prevent people from exercising their right to seek asylum. This infrastructure also includes the resort to AI systems used in immigration enforcement within borders: for example, in France, Germany, the Netherlands and Sweden, police have been given the power to fingerprint people they stop on the street to check their immigration status. This infrastructure and the increased resort to surveillance technology has been encouraged and enabled by the recently adopted EU Pact on Asylum and Migration[14], ushering in a deadly new era of digital surveillance, expanding the digital infrastructure for an EU border regime based on the criminalisation and punishment of migrants and racialised people.[15]

Welfare and Social Services

Often introduced as cost-cutting, efficiency measures with a neutral impact on peoples' rights, the increased resort to AI and digitalisation procedures in general in social services is having a demonstrable impact on low-income, working class and racialised people. Often, we see that the outcomes of these systems have discriminatory impacts and worsen the surveillance or negative outcomes experienced by marginalised groups. However, as is reflected in the cases on predictive policing, we also see that the very decision to deploy such systems in certain areas or on certain groups in society, reflects racist, xenophobic and anti-poor tendencies within governments to prioritise social harms such as so-called benefits fraud, as opposed to other social issues. These are inherently political decisions which are inherently classed and racialised, despite the outcomes or workings on the specific systems.

AI and algorithmic systems in Europe have been used to perform identity verification, allocate and determine access and eligibility to social security and welfare services (such as unemployment benefits), and also to predict and risk assess people for benefits/ welfare fraud.

For example, in 2014, the Dutch government famously authorised and rolled out Systeem *Risico Indicatie*, or SyRI in predominantly low-income neighbourhoods, in an effort to pre-

12 Vavoula, N. (2022) *Immigration and Privacy in the Law of the European Union – The Case of Information Systems*
13 Euromed rights (2023. *Artificial Intelligence: the new frontier of the EU's border externalisation strategy*:) https://euromedrights.org/wp-content/uploads/2023/07/Euromed_AI-Migration-Report_EN-1.pdf
14 https://home-affairs.ec.europa.eu/policies/migration-and-asylum/pact-migration-and-asylum_en#what-is-the-pact-on-migration-and-asylum
15 Protect Not Surveil Coalition (2024). *The EU Migration Pact: A deadly regime of migrant surveillance*: https://www.equinox-eu.com/wp-content/uploads/2024/04/The-Migration-Pact-ProtectNotSurveil.pdf

dict and score people's likelihood of engaging in benefits or tax fraud. The SyRI system linked and analysed a large amount of data on Dutch citizens for this purpose, including data on employment status, property ownership education, retirement, business, income and assets, pension and debts. The system relied on a collaboration between municipalities, the Ministry of Social Affairs and Employment, police, the Public Prosecution Service, immigration services, and the welfare and tax authorities. [16] The system then generated a list of addresses of people who displayed a higher likelihood of public benefits fraud. Whilst the court struck down the system for privacy and transparency reasons, the vast scale of the harm inflicted on poor and working-class people, largely from racialised and migrant communities.[17]

In Austria, as highlighted the organisation Epicenter.works, the government rolled out the 'AMS algorithm', an algorithmic system designed to predict a job seeker's employment prospects based on factors such as gender, age group, citizenship, health, occupation, and work experience. Prioritising services based on this information, the AMS algorithm resulted in reduced support to job seekers with low and high employment prospects, and also discriminated against women over 30, women with childcare obligations and migrants.[18]

In other cases, we have seen governments deploy AI systems for the seemingly mundane purpose of identification of persons as a precondition to access to their social security. Whilst forming less of a 'decision-making' exercise than previous examples, such use cases have had severe discriminatory impacts insofar as the failures and inaccuracies of such systems have led to false determinations of identity fraud, leaving people without benefits. Such systems, are a technological form of increasing the evidentiary burden on people who need to access benefits, which disproportionately disadvantages racialised people and people in precarious circumstances. Commenting on the dangerous impacts of the introduction of AI in welfare Human Rights Watch argued that this trend toward automation can *'discriminate against people who need social security support, compromise their privacy, and make it harder for them to qualify for government assistance.'*[19]

Beyond Discrimination: Political Economies of Racialised AI

Whilst the disproportionate negative outcomes experienced by racialised people is one aspect of the link between AI and structural racism, there is a much broader picture to assess. Specifically, the transformative impacts of the AI industry, from the economic impacts of AI development and production, lobbying, and the link with broader industries, such as security, defence and militarisation, are just as, if not more important to assess. These political economies have impact on racialised communities across the works and interact with systems of racist exploitation and the systematic over-exposure to premature death at a global level. A political economy analysis of AI also unveils the connection between the AI industry and racial capitalism, as well as the role of AI in the '*production and exploitation of group-differentiated vulnerability to premature death.*'[20] This section gives a brief entry point description of these different political economies.

16 Mustafa, Nawal (2023). Article 47: The age of digital inequalities. Digital Rights are Charter Rights del Digital Freedom Fund: https://digitalfreedomfund.org/wp-content/uploads/2023/09/Digirise_V11_Digital.pdf

17 Bits of Freedom (2021). 'We want more than symbolic gestures in response to discriminatory algorithms'

18 Epicenter.works (2020). *Warum der polnische "AMS"-Algorithmus gescheitert ist* https://epicenter.works/content/warum-der-polnische-ams-algorithmus-gescheitert-ist

19 Human Rights Watch (2021). *How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net* https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net

20 Ruth Wilson Gilmore (2007) Golden gulag: Prisons, surplus, crisis, and opposition in globalizing california. Berkeley, CA: University of California Press.

Extraction, Production and AI Industrial Policy

AI is a multi-billion dollar industry. Whilst often lauded in politically neutral or positive terms as an unequivocal benefit to society or a symbol of economic progress, it is crucial to remember that AI systems are part and parcel of the efforts of large technology companies to expand their market power.[21]

Artificial intelligence has become a central pillar of the industrial policy of governments and institutions across the world. As part of broader efforts to generate value through the digitalisation of economies, governments have invested massively into tech-solutionist discourses exposing the benefits of AI, and the need to 'promote the uptake' of AI in order to remain competitive on a global scale. However, promoting AI in the public sector as a whole, without requiring scientific evidence to justify the need or the purpose of such applications in some potentially harmful situations, is likely to have the most direct consequences on everyday peoples' lives, particularly on marginalised groups.

More broadly, 'pro-AI' industrial policies obscure a number of broader, societal harms at the centre of the AI industry and the concept in general. Firstly, they obscure that the AI industry is inherently based on extraction and exploitation – of labour, of natural resources, of land. Increasingly, research is unveiling the extent of labour exploitation underpinning the AI industry, which centres the exploitation of workers in the Global South. Systems such as Chat GPT, which are the consumer-facing manifestation of large language models or general purpose AI systems, consistently exploit workers in a variety of ways, including as part of the production and maintenance process through data labelling, filtering and collection,[22] content moderation for online platforms, or through the direct provision of services as platform workers, such as drivers and delivery couriers. All of these forms of employment maintain a highly exploited and precarious workforce, often with long working hours, low pay and poor conditions. Prug and Bilic describe this process of 'dividing and hiding labour' to ensure that products seem magically automated, all the while amassing value of labour from workers in the Global South, or in precarious positions in the Global North.[23]

Additionally, the AI industry in its production process is also predicated on a great deal of environmental exploitation and the extraction of natural resources. AI systems and their underpinning frameworks require large amounts of computational power, and therefore energy sources. Further, the collection of data for AI relies on the increased use of mobile devices and sensor networks, all of which require devices that depend on the mining of minerals and the extraction of materials at the roots of the war in the Congo. As highlighted by the organisation *Generation Lumiere,* the global electronics industry has fuelled the war in the Democratic Republic of Congo, due to demand for Cobalt, Copper, coltan, and lithium all of which Congo has high reserves.[24] The AI industry attempts to obscure environmental, climate and conflict considerations with marketing claims that AI will help us fight climate change (i.e. through forecasting modelling), a narrative which is unfortunately reproduced by a number of policymaking institutions.

Digital Securitisation, Militarisation and Digital Warfare

As highlighted in the previous section more and more law enforcement and migration agencies resort to AI. Rather than exploring these as isolated events, these deployments are part of a broader trends of securitisation and militarisation. Here 'securitisation' refers to a framework of resources, legislation, narratives, and, increasingly, technological infrastructu-

21 EDRi (2021) *How Big Tech maintains its dominance*: https://edri.org/our-work/how-big-tech-maintains-its-dominance/

22 Prug, Toni, & Bilić, Paško (2021). Work Now, Profit Later: AI Between Capital, Labour and Regulation. In P. V. Moore y J. Woodcock (Eds.), *Augmented Exploitation: Artificial Intelligence, Automation and Work* (p. 30–40). Pluto Press

23 *Ibid.*

24 Generation Lumière (2024). La consommation en métaux des Européens génère des massacres – Reporterre: https://reporterre.net/L-appetit-en-metaux-des-Europeens-genere-des-massacres#:~:text=«%20La%20consommation%20en%20métaux%20des%20Européens%20génère%20des%20massacres%20»,-Des%20mineurs%20en&text=Le%20conflit%20au%20Kivu%2C%20en,et%20de%20l'Union%20européenne.

res mobilised in pursuit of a vision of 'security' that centers militaristic, punitive and surveillance-based solutions to social problems. Within the EU's securitisation framework, institutions irrevocably fuse the concept of public safety with police, borders, and the military.

Digital securitisation – the integration of digitalisation into securitisation trends[25] - projects involve expanding the legal basis for technological infrastructures to further surveillance and criminalisation (such as the aforementioned EU Migration Pact, or the Artificial Intelligence Act, discussed in the next section), but also massive investments into the digital outputs of the security and military industries. For example, the EU increasingly rolls out funds to widen securitisation infrastructures and agencies, such as EUROPOL, the EU's policing cooperation agency. Much of this investment is outsourced through contracts to private surveillance and technology companies to develop tools for the purpose of increasing deportations and border surveillance. As reported by Statewatch, Frontex's 2023 procurement plan included EUR 260 million for IT systems including software development, infrastructure and administrative systems, a further EUR €180 million on equipment for border surveillance, including a drone contract of €144 million.[26] As such, we see the encroachment of the private sector, including technology companies, into state functions, increasingly centralising these economic interests into state institutions such as law enforcement and migration control. As such, we see a reorientation of budgets toward the policing, surveillance and control of largely racialised populations, when resources could be spent elsewhere. For example, The overall amount of money earmarked for security and defence spending within the EU budget from 2021-27 is €43.9 billion, an increase of more than 123% when compared to the previous seven-year budgetary cycle, which allocated €19.7 billion for the same purpose. In comparison, the Citizens, Equality, Rights and Values Programme only allocated €1.4 billion for projects dedicated to improving rights and equality particularly for marginalised groups.[27] Not only does this produce the disproportionate exposure to harm we outlined in the previous section, we see a transformation of the objectives of states and institutions away from social provision and protection and toward, instead, surveillance, control, policing and population management.

These trends are directly mirrored in the context of militarisation, in which AI systems are increasingly developed, tested, procured and deployed for functions related to warfare. States and institutions increasingly invest in AI in the military context using a pretext of safety – automating weaponry to decrease need for physical personnel on the ground. This discourse largely ignores the fatal consequences for populations on the sharp edge of such technologies. A prominent example of how the AI industry has been used to facilitate warfare and supercharge mass murder can be seen with respect to various deployments as part of Israel's genocide on Gaza. Systems such as Lavender[28] have been deployed to 'supercharge' up the generation of targets for Israel's bombing. As such, these technologies are to be equated with the equipment of warfare and weaponry and must be treated as such. Yet, exploring the financing of such systems unveils again that these are not isolated deployments of harmful technology, but an industry fuelled by a range of vested interests and part and parcel of broader linkages between the AI industry and militarisation. Exploring the growth of the military uses of AI, Lushenko and Carter state:

'The primrose path of AI-enabled warfare is paved by a new military-industrial complex. Countries typically acquire military technologies, such as drones, for reasons that

25 Chander, Sarah & Gürses, Seda (2024). From Infrastructural Power to Redistribution: How the EU's Digital Agenda Cements Securitization and Computational Infrastructures (and How We Build Otherwise). In *Europe's AI Industrial Policy*. AI Now: https://ainowinstitute.org/publication/from-infrastructural-power-to-redistribution-how-the-eus-digital-agenda-cements-securitization-and-computational-infrastructures-and-how-we-build-otherwise

26 Statewatch (2023) Frontex to spend millions of euros on surveillance and deportations: https://www.statewatch.org/news/2023/april/frontex-to-spend-hundreds-of-millions-of-euros-on-surveillance-and-deportations/#:~:text=Frontex%20will%20spend%20hundreds%20of,management%20board%20in%20mid%2DFebruary.

27 Statewatch (2022). At What Cost? Funding the EU's security, defence, and border policies, 2021–2027: https://www.statewatch.org/publications/reports-and-books/at-what-cost-funding-the-eu-s-security-defence-and-border-policies-2021-2027/

28 Abraham, Yuval (2024). Lavender: The AI machine directing Israel's bombing spree in Gaza. *+972 magazine* https://www.972mag.com/lavender-ai-israeli-army-gaza/

relate to supply, demand, and status considerations."[29]
As such, economic interests underlying investment in AI for warfare not purely security and defence, but also status considerations, financialisaton, and the interest of the tech industry more broadly. This can be seen as a clear continuation of the origin and trajectory of AI development and investment as inherently connected to the US Military since 1958 and the creation of DARPA, the Advanced Research Projects Agency to facilitate research and development of military and industrial strategies.

AI is often developed for other, wider purposes and then deployed in warfare contexts, and vice versa. Often law enforcement exchanges learning and technologies from military contexts which are then modified to domestic policing, highlighting global linkages in the policing and population management of racialised peoples. As stated by Sara Hamid:

> "But it's not just about global markets. It's also about global contexts. *American policing functions as a research site for military innovation—the "green to blue" pipeline is bidirectional*."[30]

This connection is evident when exploring border infrastructure as part of Fortress Europe, where agencies such as Frontex increasingly deploy surveillance technologies tested by Israel on Palestinians in the context of apartheid, occupation and genocide. As such, the role of AI in the racialised policing and management of populations is of global reach and inherently connected.

Following the trends described above, we see an overwhelming trend increasing investment into the AI industry, both in public budgets as well as private investment. Regardless of the fields in which AI infrastructures are deployed in, their increased introduction and investment has become integral to the delivery of public services and as such as transformed the functioning of democratic institutions and functions in ways that are geared toward the profit motives of industry. This has fundamental implications: transforming the institutions necessary for democracy while, cementing the infrastructural power of US technology companies.[31]

Response and Resistance: Charting Efforts to Context AI and tructural Racism

There is no "quick fix" to undo centuries of systemic racism and discrimination. As we have seen throughout this chapter, the problem is not just baked into the technology, but into the systems in which we live. In most cases, AI systems only make racial injustices, discrimination and violence harder to pin down and contest. And yet, contestations on various levels are necessary. Yet, the responses to the problem of racism and AI are highly political, motivated by very specific value judgements about technology and the deploying institutions themselves. In this chapter, we chart and differentiate various methodologies to respond to the harms of AI, shifting through corporate 'de-biasing' and auditing measures, regulatory attempts, to much broader resistance practices.

Technical Debiasing

One of the most common refrains in response to the discriminatory impact of AI is within

29 Lushenko, Paul & Carter, Keith (2024). A new military industrial complex: how tech bros are hyping AI's role in war. Bulletin of the Atomic Scientists. https://thebulletin.org/2024/10/a-new-military-industrial-complex-how-tech-bros-are-hyping-ais-role-in-war/
30 Logic Magazine (2020). Community Defence: Sarah T Hamid on Abolishing Carceral Technologies. Logic(s) Issue 11: https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/
31 https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/#:~:text=AI%2Ddriven%20systems%20have%20broad,of%20debiasing%20algorithms%20and%20datasets.

the corporate or technical realm: framing discrimination as 'bias' within the technology and advancing techniques to correct for them within the technical system. Debiasing responses draw from a logic that explains AI-based harms as a result of skewed data or technical hiccups, as opposed to broader structural, political, economic issues within society or within the AI industry. In most cases, these techno-centric solutions can at best reduce surface level issues with the operation of an AI system. For example, technical de-biasing can unpick the reasons as to why systems work well on some populations and not others – for example in a case of facial recognition systems disproportionately misidentifying racialised people. Often, such problems can be characterised as the incomplete or unrepresentativeness of underpinning datasets or of the functioning of the algorithm – whilst explained by societal issues, debiasing techniques rely on the myth that AI system can ever be unbiased when deployed in a world characterised by structural racism. It is for these reasons that Julia Powles argues that debiasing is a 'seductive diversion' and that such techniques amount to 'perfecting the instruments of surveillance.'[32]

Seda Gürses and Agathe Balayn have well documented much broader concerns with debiasing techniques. In many ways, they argue, debiasing is a corporate industry in itself, perpetuating a narrow approach that *'squeezes complex socio-technical problems into the domain of design and thus into the hands of technology companies. By largely ignoring the costly production environments that machine learning requires, regulators encourage an expansionist model of computational infrastructures driven by Big Tech.'*[33]

Debiasing approaches, by focusing on AI from a largely product based perspective that only has individual impacts, largely obscures the political economies underpinning the AI industry, extraction, exploitation and racialised surveillance – structural problems throughout the production process. It is for these reasons that debiasing approaches are at best incomplete and ineffectual at tackling structural racism, and at worst harmful in their attempts to invisible the harms stemming from the AI production process. A learning, applied in this chapter, is the crucial need to look at the harms of AI in ways beyond 'disproportionate impact or access' to certain AI based services and experiences and instead address the political economies of AI.

AI Regulation and Legislative Approaches

How to regulate AI has been a central question of the last half-decade, particularly within Europe. In April 2021, the European Commission launched its legislative proposal to regulate AI in the European Union.[34] Passing into force in August 2024, the EU Artificial Intelligence Regulation attempts to regulate artificial intelligence largely through a process of risk categorisation, product safety mechanisms, and limited governance and accountability measures. The AI Act in its final form prohibits some, limited, uses of AI, categorises others as 'high risk' and provides for a system of national and international monitoring and enforcement, largely overseeing the processes for which AI products get access to the EU market.

One form of contestation to the vast harms of AI systems was to influence this regulation. A coalition of 150 civil society organisations sought advocate for red lines, accountability and transparency, and mechanisms of redress in order to counteract the harms of AI, stemming from mass surveillance, environmental impact, structural discrimination, implications on democracy, and more.[35] In particular, the demands to legally prohibit some forms of AI – including biometric mass surveillance in public spaces, predictive policing, and various

32 Powles, Julia (2018). The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. *OneZero:* https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

33 Gürses, Seda & Balayn, Agathe (2021) Beyond Debiasing: Regulating AI and its inequalities: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

34 https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

35 https://edri.org/our-work/eu-ai-act-trilogues-status-of-fundamental-rights-recommendations/

uses of AI in the migration contexts that undermine the right to asylum, were informed by the abolitionist goal of seeking the reduce the scope, scale, legitimacy and tools provided to the criminal justice system and police.
The final text demonstrated the limits of using legislative advocacy as a meaningful resistance practice. The text categorised some harmful uses of AI in the context of migration as 'high-risk' – but failed to address how AI systems exacerbate violence and discrimination against people in migration processes and at borders. Presented with the opportunity to meaningfully limit the use of AI to perform mass surveillance and discriminatory targeting of marginalised communities, EU legislators wholeheartedly failed to include necessary safeguards, in particular in the areas of security, policing and migration control. The Act stopped short of prohibiting the worst forms of predictive policing, biometric surveillance, and harmful uses of AI in the migration context.

Further, the AI Act introduced wide regulatory loopholes, with police, migration control and security actors largely made exempt from the public transparency and accountability requirements imposed on deployers of 'high-risk', designed specifically to exclude scrutiny and obligations on police, security actors and migration control. As such, EU legislators solidified the existing state of opacity in which state actors deploy surveillance technologies to monitor, classify, sort and punish people. Further, in many ways, the minimal technical checks required under the text of (a limited set of) high-risk systems in migration control could be seen as enabling, rather than providing meaningful safeguards for people subject to, these opaque, discriminatory, surveillance systems.

The legacy of the AI Act in the context of securitisation is therefore forming part of a broader trend of enabling and endorsing the use of AI on racialised people. Also, in the last EU mandate we saw the EU Migration Pact, endorsing and expanding the surveillance and criminalisation of migrants. The final Pact takes numerous steps to ramp up the digital systems, investments and infrastructures used to prevent and control migration, including enabling intrusive technological practices in asylum processing, expanding the introduction of technological management of detention centres, and expanding an already vast regime of data collection and digital monitoring of migrants.[36] It is the opinion of the author, that, the impact of the AI Act and related securitisation legislation is to expand and even necessitate the surveillance activities of police and migration control, setting a precedent for a wider ideological shift that justifies lesser scrutiny, lesser accountability, lower requirements for an evidence base and wider legal frameworks of law enforcement surveillance for the use of data and invasive surveillance technologies.

Toward justice and resistance

What resistance can be staged, beyond corporate and institutional responses to AI-based harms? There are but a few avenues that look to attack the root of the political economies of the extractive, exploitative, essentially discriminatory AI industry.

The first involves practices of mobilisation, documentation and building power amount affected communities. One of the difficulties of challenging AI has always been the opacity and obscurity of the AI industry, which has largely avoided and concealed its connections with state surveillance, securitisation and militarisation, or presented solutions that only feed power back into the AI industry. To counter this, we need to build and equip community-based work, racial and migrant justice movements, other affected communities, to understand and build their own strategies of safety, damage control, and broader liberatory practices. These approaches see contesting racist uses of AI as part and parcel of border anti securitisation, anti-border, anti-imperialism practices, rather than as something separate.

36 https://www.equinox-eu.com/wp-content/uploads/2024/04/The-Migration-Pact-ProtectNotSurveil.pdf

The next involves moving beyond disproportionate impact and instead exploring practices of interruption, disruption and refusal in the AI supply chain. Such practices re-materialise and visibilise the exploitation, extraction and militaristic aspects of the AI and technology industry. For example, we have seen numerous popular protests to contest the introduction of AI in public services on a domestic level – for example student protests in the UK against the A-level grading algorithm and the SyRI welfare fraud detection system. On a global level, we see increasingly efforts at contesting digitalisation and securitisation, with various efforts to disrupt supply chains sending weaponry and only technological resources to support a genocide in Gaza (such as Project Nimbus) and contesting the reciprocal procurement and legitimisation that ensues when western governments procure surveillance technology from Israel. Here connections between workers at different points at the supply chain, including tech workers in Silicon Valley such as in No Tech for Apartheid, but also dock workers and others, aligned with student and racial justice movements, have demonstrated powerful constellations of resistance toward the abolition of carceral technologies and the use of technology as colonial complicity.

The last I will mention are efforts toward redistribution. The AI industry is fuelled by a powerful narrative of progress and innovation that justifies billions in private investments, but also the pouring of public budgets into developing and procuring AI systems in any number of spheres. Often, this is inherently connected to securitisation and militarisation, as is the history of the AI discipline. Efforts toward a redistributive approach, not just away from AI but all industries based on resourcing punishment, containment, violence and war, are a central facet of what resistance could look like. We can build meaningful strategies directing resources and political will away from extraction, exploitation, criminalisation and control and instead toward robust systems of social provision, care and wealth distribution.

Conclusion: Race and Resistance, Towards Justice

Assessing AI from the perspective of racial justice requires multiple shifts. It requires reframing AI from a product or a benefit to an industry; centring processes of production, exploitation and militarisation instead of only ‘disproportionate impact’; and loosening our reliance on corporate and institutional mechanisms of ‘repair’. As this chapter has articulated AI is part and parcel of broader infrastructures of oppression: criminalisation, securitisation and racial capitalism. Resistance practices require an unequivocal struggle toward refusal of the capitalistic markets of technological ‘process’, efforts of decriminalisation and demilitarisation, a model of economic redistribution of wealth, and the rebuilding of community led infrastructures of care, connection and social provision.

6. The impact of Language AI on access to and production of knowledge

Pelonomi Moiloa
Lelapa AI

The failure to recognize the diversity of language as both a beautiful and essential aspect of human life has profound consequences. Language serves as a living archive, preserving vast knowledge, cultural practices, and unique worldviews. When languages are marginalized or lost, we lose not only this invaluable repository of knowledge but also the ability to understand how communities connect and relate to one another through their linguistic expressions. Furthermore, language shapes the way we think, offering diverse approaches to problem-solving that can drive collective innovation. Within this diversity lie clues that can deepen our understanding of ourselves and our shared humanity.

This paper explores these ideas through the lens of African language technology, illustrating how addressing the "low-resource" status of these languages can catalyze a transformative shift in language technology development. Unlike mainstream approaches, which often prioritize efficiency and profitability, the context of African languages necessitates more inclusive, community-driven methodologies. This shift opens the door to a paradigm that benefits not only speakers of underrepresented languages but also global technology ecosystems by providing diverse, sustainable models for the future.

Introduction

In the National Geographic's Vanishing Voices, Russ Rymer states that "One language dies every 14 days, and by the next century nearly half of earths 7,000 languages will likely disappear"[1]. Some might argue that this is merely a natural progression of the world. While it's true that languages adapt to reflect the necessities of the present, evolving as life itself shifts and changes across time and space, I would contend that this process is not entirely organic. Much like climate change—an inherently natural phenomenon that humans have accelerated at an alarming pace—we are actively driving linguistic shifts far faster than we are equipped to understand or respond to their long-term implications with AI technology that is predominantly developed for and by English speakers. How the current development of high-resource languages in NLP creates a widening gap is studied well in papers like *The Zeno's Paradox of 'Low-Resource' Languages* by Atnafu Tonja et al (2024) and *The state and Fate of Linguistic Diversity and Inclusion in the NLP World* by Pratik Joshi et al (2020).

For those whose ancestors were assimilated into homogenous cultures through language centuries ago, the violence of such acts may now seem distant, perhaps even forgotten. However, for those who have experienced these patterns more recently, the wounds are still fresh. While language—and particularly the homogenization of language—can serve as a unifying force by enabling communication, it is also a powerful technology that has been weaponized as a tool of dominance, control, and marginalization.

In the context of colonialism, homogenous languages were imposed on indigenous groups for centuries, disrupting cultural continuity in exchange for a so-called "key to access." Mastery of the colonizer's language became a prerequisite for education, governance, socioeconomic mobility, and inclusion, creating a hierarchy that systematically disadvantaged native speakers. Language has also been wielded as a means of controlling discourse and perpetuating "othering," reinforcing inequalities and diminishing the identities of those outside the dominant linguistic group.

The homogenization of language, while seemingly practical, carries a legacy of violence that continues to shape global power dynamics today. If we are not careful, this legacy could influence the trajectory of language technology and dictate what we stand to lose. At this critical tipping point, it is essential to recognize the immense value of linguistic diversity and to reflect on how we can be intentional in shaping the language technologies we create to ensure that these technologies honor and preserve the richness of language as they evolve over time.

Part 1: The Consequences of Losing Linguistic Diversity

> "When languages die, an immense edifice of human knowledge, painstakingly assembled over millennia by countless minds, is eroding, vanishing into oblivion."
>
> K. David Harrison

Language evolves over time among groups with shared experiences and reflects a group's way of life. Different languages highlight the varieties of human experience, revealing mutable aspects of life that we tend to think of as universal - our experience of time, numbers, belonging, technology and how we build it. Paraphrasing the work of Achille Mbembe - "Language is a living archive". When languages are forgotten, so is this archive.

1 "Vanishing Voices" by Russ Rymer at *National Geographic* (July, 2012): https://www.nationalgeographic.com/magazine/article/vanishing-languages

Language as a Living Archive

Loss of language results in a loss of ecological understanding, medicinal plant usage, agricultural practices, spiritual beliefs, and history that are often specific to a region and culture. It includes knowledge of animals, moon phases, wind patterns, how people relate to each other and what parts of that relationality they hold to be significant. Societies that rely on nature for survival have developed technologies to cultivate, domesticate, and utilise such resources. In K. David's book "When Languages Die", David relays how much of what humankind knows about the natural world, survival and how we relate to each other lies completely outside formal structures of archival practice. They are not in science textbooks, encyclopaedias and databases. They often only exist in peoples memories. In fact, we often look to indigenous knowledge sources to fill the knowledge gaps in our understanding of the world. Some of this information has been harvested for the profit of a select few. An estimated $85 billion in profits per year is made by pharmaceutical companies on medicines derived from plants first known to indigenous peoples for their healing properties. As K. David Harrison explains: "as languages disappear, so does this intricate knowledge, diminishing humanity's overall understanding of cognitive diversity and ecosystem relationships". **Languages serve as cultural archives, reflecting the beliefs and practices of the communities that speak to them. In Sub-Saharan Africa, oral traditions such as storytelling are not just entertainment—they are vital for preserving history, values, and communal identity**. In these societies, as an example, language and storytelling reflect and reinforce societal expectations and communal values. Unlike adopted languages, which may not carry these cultural norms, African oral traditions are deeply intertwined with collective identity and memory.

A defining feature of African storytelling is the presence of a call-and-response dynamic between the storyteller and the audience. Specific phrases or words signal the beginning of a story and invite audience participation among languages across the continent. **In Swahili storytelling for example, the storyteller begins with 'Hadithi hadithi,' meaning 'Story story,' and the audience replies 'Hadithi njoo,' meaning 'Come story.' This exchange not only signals the start of the tale but also reinforces communal participation, connecting listeners to their cultural heritage**.

Unlike Indo-European expressions like "Once upon a time," which focus on individual narratives, these African phrases emphasise shared experiences. The interaction goes beyond words, often involving songs, chants, or dances, turning storytelling into a communal ritual that connects people to their environment, ancestors, and cultural heritage. Storytellers in these traditions, such as griots or imbongi, serve as historians, genealogists, and cultural custodians, preserving societal values and history. This oral tradition fosters indirect communication styles, where meaning is often conveyed through metaphor, allegory, and symbolism. Such methods prioritise maintaining social harmony and leaving room for individual interpretation.

Cultures with less emphasis on storytelling as a means of cultural transmission are less likely to develop these interactive and indirect communication habits. The African oral tradition highlights the deep interdependence between storytelling, culture, and community, underscoring how these practices preserve collective identity and social cohesion. When individuals from one culture adopt the language of another, they often experience a shift in identity. For many, this shift can feel like a loss of self, as explored in works like *Lost in Translation: A Life in a New Language* by Eva Hoffman and *Decolonising the Mind: The Politics of Language in African Literature* by Ngũgĩ wa Thiong'o.

In addition to a community's internal environment, language evolves in close interaction with the external environment too, adapting to optimise communication under specific conditions. Tones, for example, allow languages to efficiently encode meaning and emotion by

varying pitch or contour. This makes verbal communication more expressive and precise without requiring additional words. In tonal languages like Yoruba, a single syllable's meaning can shift entirely based on its tone. Acoustic adaptations in languages also reflect environmental influences. For instance, research hypothesises that vowel sounds travel more effectively in humid air than dry air, potentially contributing to differences in Southern vs. Northern Indian languages. Similarly, click consonants in Khoisan languages are sharp and percussive, remaining intelligible across long distances in open, natural environments—an adaptation advantageous in outdoor settings. These examples illustrate how linguistic features develop in response to environmental and social needs, ensuring verbal expressions suit the specific contexts in which they arise.

Language profoundly shapes how people think and problem-solve, acting as an extension of cognitive processes and ideas. The structure of the Chinese language is closely tied to brain regions involved in visuospatial reasoning, which can enhance mathematical abilities. Arabic speakers often develop robust abstract, hierarchical pattern-recognition abilities as they are trained to detect patterns within complex word forms. Agglutinative* language speakers may also experience cognitive advantages in multilingual learning contexts. A study by Nkolola-Wakumelo (2008) in *African Multilingualism* suggests that speakers of Bantu languages like isiZulu demonstrate heightened sensitivity to language structure and develop strong skills in linear pattern recognition. "*Languages reveal the limits and possibilities of human cognition—how the mind works. Each new grammar pattern we find sheds light on how the human brain creates language*." K. David Harrison

When communities abandon their languages and switch to English or Spanish, there is a massive disruption of the transfer of traditional knowledge which can be lost in large part just over one or two generations. This loss occurs directly in the words we use to describe the physical worlds and our experience with it. But this knowledge is also lost or distorted in how cultures organise that knowledge and transfer it from one person to the next and how it shapes the way in which they approach problem solving and in understanding the self.

**Agglutinative languages are described as languages where words often consist of stacking suffixes and prefixes onto root words.*

Perpetuation of Bias and Exclusion of Populations in Large Language Models (LLMs)

The loss of a language is not only the loss of words but of knowledge, culture, and heritage. This is deeply tragic, but the implications for those who maintain these languages are complex. In a world where language technology is accelerating the process of homogenising languages, what happens to those who resist this trend?

Language exclusion results in poor performance of these language tools. Without a robust representation of multiple languages, LLMs perform poorly in multilingual applications, such as translations or language-specific tasks. Conneau et al. *(2020)* in "Unsupervised Cross-lingual Representation Learning" illustrate that without lower-resource languages, LLMs fail to generalise effectively, as seen in MT (machine translation) tasks where models struggle with rare languages, leading to poor quality translations and even the potential loss of grammatical structures specific to certain languages. **In one infamous case, Facebook's language model mistranslated 'Good Morning' from Arabic into 'Attack them,' which led to the wrongful arrest of a Palestinian man. This incident highlights the dangers of underrepresenting certain languages in AI, where small errors can have severe real-world consequences.**

Excluding languages from the development of modern technology often leads to biassed outputs favouring dominant languages like English. For instance, an English-trained model

may fail to interpret context-sensitive meanings or cultural nuances in non-English contexts. In *Bend*er et al. (2021), "On the Dangers of Stochastic Parrots," the researchers argue that by training LLMs predominantly on high-resource languages, the model's responses become skewed, thereby marginalising underrepresented languages and cultural perspectives going so far as to assign negative sentiment labels to black names in sentiment labelling tasks.

There is increasing academic recognition that the lack of local language support in global social media and large language models can create significant vulnerabilities in information integrity, especially in marginalised regions. This gap not only hinders effective communication and education but also opens pathways for misinformation campaigns.

Studies such as "Digital Misinformation in Africa" (Shahbaz & Funk, 2019) reveal how limited linguistic inclusivity in digital platforms enables malicious actors to exploit information gaps, spreading falsehoods in local languages where moderation is often lacking. This gap allows misinformation to be weaponized in ways that evade detection by content moderation algorithms trained primarily in major languages. During elections in Nigeria in 2023 and Kenya in 2022, for example, misinformation campaigns surged in local dialects on platforms with limited language detection abilities. This allowed narratives that would otherwise be flagged as false in English to spread unchecked, potentially impacting voter behaviour and fostering divisive tensions.

A study by Mozur et al. (2018) explores how linguistic blind spots in social media algorithms were exploited to fuel ethnic violence in Myanmar. The lack of Burmese-language processing tools on platforms like Facebook at the time meant that hate speech and false information targeting the Rohingya community spread unchecked, illustrating how the absence of local language moderation can have devastating real-world impacts.

For those who preserve their language, this presents a dual challenge: safeguarding their linguistic heritage while also engaging in a globalised, technology-driven society. The evolving landscape of language technology poses significant risks to cultural diversity, raising questions about the future of communication. We know that missing out on essential technological revolutions has a significant impact on standard of living and quality of life so it begs the question, what ought to be done to make language technology more inclusive?

Part 2: African Languages in Language Models.

People who speak Indo-European languages may have experienced minor inconveniences when language technology fails them, such as autocorrect making embarrassing mistakes, voice recognition software struggling with accents, or apps mispronouncing their names. However, these issues are generally small and do not significantly disrupt daily life. For those who speak languages outside this dominant group, the situation is drastically different. In regions with limited infrastructure, especially in Africa, language technology can be the determining factor for accessing basic needs—buying electricity, receiving life-saving medical care, or understanding critical government communication. Africa, with approximately 2,000 languages, represents one-third of the world's linguistic diversity, yet these languages remain severely underserved by current technology.

Why Africans are unable to access the privilege of communication in these instances boils down to the availability of language technology in the desired languages and there are significant challenges in overcoming this gap.

Where Language Models Fail on African Languages:

Building language models for African languages is challenging. Understanding the mistakes made by these models highlights key areas of difficulty in their development. I will make mention of a few of these larger challenges here.

Code switching and mixing: A key feature of languages on the African continent is the prevalence of code-mixing and code-switching, reflecting the region's linguistic and cultural diversity. Code-mixing involves switching from one language to another, often seen in multilingual areas where individuals speak up to six languages. For instance, a speaker may begin a conversation in English and switch to another language for comfort or emotional expression. This typically involves full sentences in different languages.

Code-switching, by contrast, refers to using multiple languages within the same sentence. For example, when our team investigated a prime-time South African soap opera for a study it was found to feature eight distinct languages in a single episode—a common occurrence in Africa but rare in Indo-European linguistic contexts. This linguistic adaptability mirrors the continent's unique multicultural coexistence but is not an inherent capability of today's typical language technology systems.

Data shortage: The most apparent challenge, and the one that receives the most attention, is the lack of sufficient data—specifically, large collections of text and audio examples to train machines. Africa's largely oral culture, coupled with historical prohibitions on collecting such data during colonial rule, means that there is a significant scarcity of text and audio corpora in African languages. The largest existing collections are often translations of the Bible, which is why African language models frequently have a more evangelical tone compared to others.

Colonial Residues in language: Many of the errors found in language models can be attributed to the colonial history of written african languages. First and foremost, Africa is a place of oral tradition and oral culture. Written scripts rarely accompanied the development of the languages. So when missionaries came to Africa they sought to record the language and retro fit the roman alphabet to the sounds that they heard. Of course, the roman alphabet could not fully encapsulate the sounds in a language. And to untrained ears, many of the sounds in African languages sound the same…when they are not. My name, for example, has two o's in it. In the Tswana language there are eight ways to pronounce an "o". But they are all written the same. The two o's in my name are not actually pronounced the same.

The issue mirrors the historical role of missionaries and anthropologists from European countries, who were often the first to record African languages in written form. Depending on whether a French, English, or German missionary encountered a group, the spelling of that group's name could differ. For example, my surname has variations like "Moiloa" and "Moilwa," based on which European missionary encountered our family first. The inconsistency in pronunciation of vowels, despite identical spelling, adds another layer of complexity that language models are not equipped to handle.

Indo-european Standard of NLP Models

The development of language models today relies on foundational assumptions that largely originate from Indo-European languages, the linguistic group these technologies were initially designed to serve. Consequently, these models often lack the flexibility required to adapt effectively to languages that fall outside these linguistic structures and conventions.

Tokenization, the process of breaking text into smaller units (like words or subwords), is straightforward for English but far more complex for agglutinative languages such

as isiZulu. In isiZulu, a single word like 'Ngiyafunda' ('I am learning') conveys multiple layers of meaning, requiring specialized tokenization techniques to handle its structure effectively. This means that the fundamental assumption of how to apply complex machine learning mathematics fails, when the language does not follow the structural norms language models are used to. With conundrums like code-switching, a model needs to understand which tokenisation method is best used on a word by word basis depending on which language the word is from which complicates things further if the tokeniser itself is not "multilingual".

There are also cultural assumptions that underlie language models. An example here is direct communication cultures vs indirect communication cultures as mentioned previously. In direct communication cultures the meaning behind a language is explicit in the words. In indirect communication cultures, meaning is implied. The largest contribution of text to language model technology is American based direct communication culture. Africans do not communicate in this way, often relying on context and subtlety to convey meaning, which is not captured well by machines, especially when there is not an abundance of examples to illustrate how subtlety might be conveyed.

Language intricacies aside, the unit economics of developing and serving this technology can also be a challenge. Apart from being extremely data hungry, developing this technology requires large amounts of capital. Training, one of the world's largest generative models costs more money than the GDP of 16 African countries. Large language models also require a ridiculous amount of computation and are thus energy hungry. Training one of the large generative models in the world uses the same amount of electricity to power 12000 Johannesburg homes for a month. A city that has experienced intentional periodical blackouts due to capacity constraints until recently. A single query to these large language models utilises an estimated bottle of water per query. What is more, the type of computation required for this technology includes GPU compute rather than CPU compute. GPU compute requires significantly more rare earth elements and precious metals. One of these is Tantalum, a mineral derivative of Coltan, one of the main attributors to the civil war and violent conflict of current day Congo. These conditions are not suitable for the African context. But to be honest, these conditions are not suitable for a sustainable world either. A different approach is required to make these models work at scale for more people.

Though it is estimated the cost of production of the models will decrease significantly over time, given that the best GPT style models only process about 100 times the amount of information as a child might do in the first year of their life*, and that current experiments[2] have shown that more data won't necessarily increase their capability significantly, the next breakthrough in deep learning style technologies for fields such as multi modal models (models that learn from differing media and not just from text data) or in model reasoning is estimated to require an order of magnitude greater than ChatGPT's hundreds of millions of dollars.

**At the World Government Summit held in Dubai in early 2024 Jan Lacoon explained how little information a large language model is able to process when compared with a child by the time they have reached 4 years old. At a processing rate of 20 megabytes per second, which is an equivalent of 1 to the 13 tokens. It would take the largest modern LLM 150 years to process this information*

2 Muenninghoff, N. et al. (2023). Scaling Data-Constrained Language Models. At: Oh, A. et al. Advances in Neural Information Processing Systems 36 (NeurIPS 2023). Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html

Part 3: The Benefit of Working from the Recognition of Diversity.

With the limitations in data and computing resources, as well as the complexity of the African language problem, one might assume that building technology for these languages is not feasible. However, this assumption holds only if we are confined to the standard frameworks used in Western language model development. In reality, language models tailored for these contexts are not only being created but are being developed effectively. The world has much to learn from the innovation and creativity that arise in low-resource settings. The insights gained from developing technology in these environments, along with alternative ways of knowing, offer new perspectives on how technology can be shaped.

Localisation and Adaptation of Solutions

There are many ways that localisation of solutions have found holes in the mainstream building language technology theory. I would like to hone in on three ways that I find particularly exciting.

The first way is through building application driven language models rather than research driven language models. The demands of the application domain is that models need to be accurate for real world use (ie be able to handle things like code mixing and code switching), that they also need to be fast for real time applications AND they cannot take up large resources as the majority of application builders do not have the compute resources or financial resources to maintain the costs of running expensive GPUs. For example, there are only a handful of available data centres that offer GPU compute on the African continent. So, GPU compute resources for application development and serving is very rare. Of the 84% of Sub-Saharan Africans who have access to the internet through 3G it was found in a study that only 22% actually use mobile internet services. When combined with the fact that Internet access on the African continent is predominantly enabled through mobile phones (and these are feature phones rather than smart phones), what really makes more sense is ensuring that personalised compute can happen through tiny ML through on the edge devices. This solves the compute problem but also allows for reduced latency, improved privacy, lower bandwidth usage (as it would not be internet connection dependent) and essentially that applications can have offline availability. What it also helps to encourage is development in the federated learning space and distributed Computing Continuum Systems (DCCS). Federated learning focuses on machine learning across decentralised devices while keeping the data localised. While DCCS refers to a broader concept of distributing computational tasks across multiple systems for general purpose computing.

One effective approach to creating smaller, more cost-efficient models is by ensuring that they are specialized for the specific tasks they are intended to perform. Rather than relying on large, general-purpose language models capable of a variety of functions, we can focus on stripping away unnecessary features, leaving only what is essential for the task at hand. This approach not only reduces model size and computational cost but also lowers the risk of misuse by ensuring the model is designed with clear, specific purposes in mind. Additionally, smaller, task-specific models are easier to interpret, allowing for better transparency and a clearer understanding of decision-making processes. This stands in contrast to the current paradigm, where generative language models are expected to generalize across a wide range of downstream tasks. Application-driven development suggests that specialized models may be more practical and effective than generalist ones.

The second way that localisation is key to innovative change is through taking the Garbage in is Garbage out as gospel. The Garbage in Garbage out principle means that the quality of output from a system or process is dependent on the quality of the input. Typically, large

language model development has not abided by this principle. Data hungry models are fed any and every piece of language evidence that a model developer can get their hands on - whether it be the dark side of the internet or not, whether the labelled data is a true reflection of the language or not. Though it has been proven to be helpful for indo-european languages where an abundance of data is available it has been an absolute failure for low resource languages. Deep dives into "publicly available" datasets for African languages by researchers such as Timnit Gebru et al. in the paper *Combating Harmful Hype in Natural Language Processing or Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets* by Julia Kreutzer et al. of Maskahane, have found the datasets to consist of nonsensical outputs that have no resemblance to the language at all.

The approach in the low resource space is the data curation approach combined with the error analysis approach such as that described in *Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages* by Masakhane. It is an inclusive approach, taking on language and cultural experts to curate the experience of a model and to evaluate outputs of models. The curation of data is necessary because of the scarcity of data. Typically, the publicly available datasets for African languages as previously stated are biblical. Training language models on this data may then perpetuate a perception that all Africans are evangelical and that is all we care to talk about. Data curation ensures that we capture key concepts that are integral to a language, its heritage and its culture and we are able to ensure particular domains are covered in order for a model to learn a better representation of what communicating in a particular language should look like. The next step in that process is the error analysis process. With cultural and linguistic expertise, in depth analysis of model failures are examined and thus well understood. Does the model confuse homophones, when does it do that? Does it fail at names? Is it confused when changing from one language to another? Does it understand that it is a woman who undertakes a particular ceremony? Does it understand that a literal interpretation is not appropriate here? By understanding these failure points very specific data can be collected at lower cost to address these issues in the models performance.

These two philosophies combined fundamentally reduce the cost associated with developing language models that can work well when deployed in the field for a particular function. Like a translation model on a mobile phone for field surveys. Or conversational support for a remote health consultation service. Or transcription and speech voice models for use in the customer contact centre domain. In the low resource context, a low-cost means democratised access to these services, and democratised access to these language services for small/medium/large enterprises. In the high resource context, low cost means a positive reduction in compute emissions, data centre invasions of arable land, less pressure on electrical grids and water systems. It also means reduced risk and safer, more interpretable models.

The third of these ways is through informed changes to model architectures and their underlying assumptions. Returning back to the example of the tokenisers, doing error analysis on these models to understand the underlying "cognitive" capability is similar to the study of languages in humans and how these studies aim to encapsulate the extent to which human brains construct and understand language. Through insights from the data we are able to understand where the models fail, and through error analysis on the models we are able to understand what underlying assumptions fail to capture the complexity of language. What this does is give us insight into what parts of these underlying assumptions struggle to adapt to the complexity of real-world language. And it gives us clues as to what to focus on in order to capture these capabilities in these language models better. Where language models fail for English is not the same place that they fail for Wolof. Developing different tokenisers, for example, can help language models learn better. We may not have known how important this step is to the language learning process of a language model had we not had the varying experience of trying those models in different languages. We might have thought we had the answers already.

Sustainability and Inclusiveness of Tools

The main benefit of these models lies in serving more diverse populations through digital technology, thus democratising access. This approach not only uncovers methods to make inclusion economically viable but also offers significant potential for local development to shape broader trends in language AI. By learning to adapt to local languages with minimal data while capturing nuances and colloquialisms, we can create technology that quickly meets present-day needs at a lower cost. The last thing we want is to pour millions into technology that doesn't evolve over time. While siloed legacy systems may be flawed, at least their logic is understood. However, updating large models like GPT currently requires massive resources—costing millions every 6-12 months for major updates and frequent fine-tuning. What we stand to gain by mastering rapid prototyping of African languages is a model that adapts quickly without needing such heavy capital injections, reducing the strain on energy, water, capital, and the exploitation of data and minerals needed to build these models.

This not only means democratized access to technology for usage, but also for local development and adaptation. Fundamentally, it challenges centralized systems of profit and power. A key part of the language AI discussion has been around open source. Open source aims for transparency and trust, encouraging collaboration and innovation through crowdsourced contributions, as well as providing cost benefits and accessibility, particularly for those who can't afford to train initial model weights. While I understand the principle behind open access, I also find it somewhat idealistic. Open source as an equitable practice assumes all participants are starting from the same level, but this assumption is flawed. The release of open-source models presumes that users have abundant data and access to compute resources—both of which are not available in many developing regions. It's primarily resource-rich organizations that dominate contributions and profit from open-source tools. The status quo in AI development is rooted in the idea that it must be centralized and monopolized because only wealthy entities can inject the capital necessary for progress. To counter this, we must develop realistic and decentralized ways of organizing technology, aligned with equitable distribution of power, sharing of benefits, and collective good.

An example of this is equitable data licensing such as the Nwulite Obodo Open data licence[3] and the Kaitiakitanga licence[4]. The premise behind these licenses is that the originators of a language—and thus the data sources—should benefit from the profits generated from their language and be recognized as data partners essential to the existence of the overarching language model. These acknowledgments may not always be monetary; they could include an exchange of services or ensuring that data is used to create tools that uplift the community. At Lelalpa AI, we recently released a dataset under a custom commercial license, which specifies that commercial entities outside of Africa must pay for the data, while African entities do not. The principle is that any profits derived from the data should be reinvested into the African economy or returned to the contributing community. In the case of African entities, proceeds go towards creating more data for the community, providing paid data creation opportunities. This model of community-governed data creation, benefit, and sovereignty stands in contrast to the extractivist approach that dominates the language model industry, where language data (e.g., internet content) is considered to belong to "no one" and therefore "everyone." In reality, only big tech has the resources to capitalize on this data, often without compensating or acknowledging the originators of the data, their language, people, or heritage.

What the requirements of the low resource context has also done is influence the culture of "organising" in the technology domain. Cultural preferences around data sovereignty and

3 From Data Science Law Lab: https://datasciencelawlab.africa/nwulite-obodo-open-data-license/

4 From Te Hiku Media: https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license

localised needs are embedded in the organisational structures of how people build together and share together. In the field of research, grass roots organisations such as Masakahane have set presidents around recognition and local ownership of research and the development thereof. Masakahane is a research initiative and pan-African community dedicated to advancing natural language processing (NLP) for African languages, founded by Jade Abbott in 2018. Masakhane has gained recognition for its innovative approach to addressing global inequalities in AI and has partnered with academic institutions, tech companies, and other organisations to further its mission. Notably, they have set a precedent of inclusive research papers that acknowledge all who contribute to them no matter how small the contribution. This president of acknowledgement highlights the collective nature by which technological advancement is achieved within the low resource context, contexts which are embedded in collectivist and community centred philosophies and ideologies. They are notorious for writing research papers with over 50 authors listed in recognition of all who make the research possible. Though scrutinised by the research community for doing so, a recent Meta paper has followed this precedent.

In the commercial space, organisations such as Huniki with a similar ideology in collectivist community benefit is an organisation that seeks to challenge the monopoly centred view around language technology. It is described as a federation that aims to ensure that African NLP is not monopolised but that local NLP providers on the African continent are supported to maintain the integrity and local benefit of the African NLP ecosystem.

What developing models in the low resource context teaches us is that 1. There are many ways to harness language intelligence into machines 2. That limited resources do not have to be the barrier to entry for building useful language technology and conversely, that it is not necessary for all language models to require such extravagant resources to be functional 3. That language model industries do not need to be owned by monopolies but can be facilitated through collective ownership and benefit. Fundamentally this highlights the significance of smaller smarter models, developed by a large community ecosystem of independent players.

Part 4: What Would Ai-Based Society Developed by an Approach that Respects and Embraces Linguistic Diversity Look Like?

So this then begs the question, what does the future of AI look like when the low resource context is given the opportunity to thrive and contribute to the development of the field? I am going to break this down into a short-term view and a long term view.

Infrastructural support to public service delivery in the short term

In our increasingly connected world, two critical factors shape access to essential products and services. The first is perhaps the most fundamental: language access is the gateway to life-critical information. The emergence of smaller, more efficient language models is democratising access to public services and systems in unprecedented ways.

Language models serve as bridges across the digital divide, enabling people to navigate the essential tasks of modern life: understanding tax obligations, deciphering utility bills, communicating with healthcare providers, and navigating public transportation. By dramatically reducing the costs of translation and communication, these technologies are particularly transformative for regions like Africa, where colonial legacies have often imposed single official languages—typically Indo-European—that many citizens may neither speak nor

understand fluently. This technological revolution allows for a diversification of instructional mediums, eliminating the need to learn a new language simply to access vital information. What language availability through digital devices at a lower cost creates is infrastructure. A compression of space such that the roads to access are that much shorter. It puts essential services directly into the hands of those who need them, instead of requiring them to travel to where those services are "available." Ideally, clinics, hospitals, government offices, and financial institutions would be located close to everyone who needs them, with affordable and reliable transport to facilitate access. However, for the majority of the world, these infrastructure projects require billions in investment, which is currently unavailable, and their implementation could take decades. Waiting for such large-scale infrastructure is not a viable option, especially when there are cost-effective, interim solutions available now.

In the short term, AI serves as a foundational support system, enhancing the infrastructure of civil governance and service delivery. Both in expanding service delivery capacity but ensuring that it is consumable through language.

Democratised Access to Digital Services in the Long Term

In the short term, AI can be used to generalize rules across groups of people to overcome resource gaps and support service delivery. In the long term, AI inspired by these alternative approaches can enable the re-emergence of the individual from the data points, offering more personalised experiences and care. As models require fewer resources, more resources can be allocated elsewhere, shifting AI from a centralised source of aggregate agency to a tool that individuals can use to meet specific needs. Beyond simply facilitating access to essential products and services, AI has the potential to augment our experiences in ways we have yet to fully imagine.

As AI becomes cheaper, faster, and more compact, personalization is poised to impact nearly every aspect of daily life. Advances in techniques like federated learning and decentralized client-controlled systems (DCCS) allow data to remain with users, promoting data sovereignty while enabling AI applications previously deemed too risky or resource-intensive.

For instance, rather than simply facilitating communication between a patient and healthcare provider, AI could conduct comprehensive body scans while integrating family medical histories to assist professionals in delivering precision diagnostics and tailored treatments—all without sensitive data leaving the patient's device. Similarly, instead of merely translating classroom materials, AI could act as a personalized tutor, supporting teachers by identifying learning gaps and offering customized strategies to improve both teaching effectiveness and student learning outcomes. These developments highlight the transformative potential of localized, personalized AI across diverse fields.

Personalized AI is currently constrained by computational limitations and the learning rates of even the most advanced models. Despite these challenges, AI holds the potential to fundamentally reshape our understanding of humanity. By replicating our abilities, we not only enhance technology but also gain deeper insights into human nature. Looking ahead, AI could play a pivotal role in alleviating resource scarcity, supporting individuals in managing essential functions, and enabling greater focus on distinctly human endeavours like creativity, empathy, and connection. The dominant vision for personalized AI remains narrowly focused on increasing worker productivity and boosting capital gains for employers—reflecting a framework rooted in Western ideals of technological advancement. This limited perspective overlooks the broader possibilities for AI to promote equity, foster creativity, and drive innovation beyond traditional economic paradigms. To realize its full potential, the development and application of AI demand a more imaginative, inclusive, and forward-thinking approach.

The technological ecosystem has the potential to evolve into a decentralized network of interconnected service providers, where responsibility for the collective wellbeing of the planet is distributed. In this vision, smaller players collaborate seamlessly with larger entities to enhance functionality, and progress prioritizes collective fulfilment over capital gain. Resource efficiency could take center stage in AI development, addressing critical concerns like climate change and the impact of value chains for precious minerals that fuel civil conflicts.

Imagining an alternative history, one might consider how AI might have developed if its early funding had not been dominated by the American defense sector. What if, instead of being rooted in capitalist, individualistic ideals, AI had been founded on philosophies of interconnectedness, such as Ubuntu—a Southern African worldview articulated by Sabelo Mhlambi as "I am because you are"? (in his influential paper *From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance*) Indigenous knowledge systems, developed over millennia, emphasize humanity's relationality with the natural world and all its elements. This perspective could inspire an AI paradigm that centers on sustainability, equity, and shared human progress. Thus lies the most beautiful opportunity of developing technology from diverse cultural perspectives, that there are alternative foundational principles that AI can be built on. And if that knowledge is not explicit, recorded in ancient texts, these secrets are most certainly held in the human expression of language.

Conclusion

Language inclusion in technology can profoundly shape societies, either by empowering communities or perpetuating inequalities. Ensuring adequate representation and support for local languages in digital tools is not just a matter of equity—it safeguards the right to future imaginaries. This right entails the ability of all people to access technology in their chosen language, empowering them to define their own narratives, aspirations, and identities. It enables communities to leverage technology to build futures that reflect their values and address their specific needs. Conversely, forcing individuals to assimilate into dominant cultural or linguistic frameworks to access opportunity undermines this agency. The long-term implications of denying linguistic inclusion, particularly in AI, remain uncertain. However, failing to provide such access risks exacerbating digital inequalities, marginalising cultural identities, and limiting the societal benefits of emerging technologies.

By embracing the unique opportunities presented by African language contexts, we can create a path that enriches technological innovation, preserves cultural heritage, and fosters equity in language representation. Ultimately, this approach offers a more sustainable and universally beneficial future for technology, one that aligns with the principles of inclusion and shared growth.

7. Algorithmic Decision-making and Rights Violations in the Gig Economy

Paola Villareal
Independent Researcher
Ervin Félix
Oxfam México
Translation: Teri Jones-Villeneuve

When the first spaces of what is now the platform economy began to emerge, they often appeared as a manifestation of the collaborative economy that connected individuals to exchange services and resources. However, the future that this formula heralded led, over time, to large corporations taking over and turning it into a vertical business. The change has been particularly detrimental to working people, as the companies with the strongest market share have opted to develop digital tools to support their business models and prioritise their profits. The other promise, that of new job opportunities and employment flexibility, has also been the victim of the model designed to maximise profits, sometimes at the expense of the framework of labour rights that was already achieved.

The algorithmic management that has become widespread on gig economy platforms is pushing for renegotiation, if not seriously jeopardising some of the fundamental rights of workers. From the right to non-discrimination to the right to a living wage. Digital tools and algorithms are used to distribute and manage the distribution of work, but they often do so on the basis of opaque criteria that do not allow workers to defend themselves properly, and are also used for surveillance. In addition, these algorithms are sometimes loaded with biases that carry axes of discrimination into the workplace, which exacerbates their impact.

Introduction

We carry supercomputers in our pockets that are millions of times faster and more powerful than those that took humans to the moon. These supercomputers contain extremely sensitive sensors that provide information such as location and direction, acceleration and movement, ambient lighting, how close an object is and, of course, more sensitive details such as users' geographical positioning, digital fingerprints and other biometric data. These data are accessible to all the apps installed on mobile devices by means of user permission. Once that permission is granted, these data feed complex algorithms and classification and prioritization models that form the backbone of many apps, such as those used in the gig economy.

The gig economy is a labour model based on short-term or on-demand work that is facilitated through digital platforms. In its early days, the gig economy was promoted as a collaborative economy that connected individuals to exchange services and resources. But over time, it has become a vertical business in which large corporations centralize control, decision-making and profits (Madariaga et al., 2018). Workers were sidelined as the major companies developed statistical and IT tools to sustain their business models and prioritize profits (A. Zhang et al., 2022).

Such algorithms and models are a part of what is known as algorithmic management (Lee et al., 2015), which is common among gig economy platforms and used to manage workers so as to meet the platforms' targets. In other words, the entire relationship between workers, customers and platforms exists with almost no human intervention. Theoretically, this type of management model is meant to improve customer service and make it easier for companies to expand into new areas without having to open offices or have local engineering, sales or management teams. Often, this has led to major disruptions for local economies and workers.

On the technical side of things, the gig economy pits workers against the digital divide: their main work tool is the mobile device they own, and access to the digital platforms – and so, to work itself – depends on the technical capacities of their tools: GPS accuracy and the device's camera or connection speed (Sanghro, 2023). Moreover, workers come up against different algorithms that take decisions about their work, from signing up new workers to issuing payments for services provided and taking commissions, as well as opaque mechanisms that link those supplying and buying services, which can harm or benefit some workers without their knowledge.

In this respect, their labour, digital and human rights depend on the ways in which the platforms' algorithms determine working conditions and dynamics (L. Zhang et al., 2023). The most violated rights stemming from the gig economy deal with privacy, safe working conditions, labour discrimination, and fair and transparent pay.

Behind the app interfaces, trillions of decisions are made automatically, immediately and opaquely. As a result, while workers see their labour rights violated by automatic decisions, algorithmic management generates huge profits for companies and shareholders (Jarrahi et al., 2020). Gig economy companies rely a business model based on opaque statistical tools that violate the labour rights of millions of people. As we will show in the following sections, algorithmic management should not occur in a black box – it must be open to public scrutiny, and companies must pay their fair share of social contributions to protect workers and ensure this type of work does not perpetuate or deepen pre-existing inequalities.

Algorithmic management and worker relations

As previously mentioned, algorithmic management has direct implications for the rights of millions of workers. However, establishing which rights are violated and when requires explaining the types of decisions that are made by the algorithms, which models are being used, the data used to train and test them, and what the results of those decisions may be. In other words, gig economy platforms are not built on a single, master model that takes all the decisions, but rather several models operating independently that are used at specific moments to resolve specific situations (Duggan et al., 2020).

There are critical moments in the relationship between the platforms and gig workers when the models play a dominant role in order to automate decision-making by factoring in a wide range of variables. These variables include account creation, authentication and ID verification, job assignment, determination of high-demand areas and dynamic pricing, distribution of pay, tips and taxes, performance measurement and incentives, and the resolution of contingencies (Dubal, 2023).

Different models are used to handle such tasks as facial recognition, location tracking, connecting supply and demand, accounting and payment, and security. Each of the models is trained according to a specific context based on user data that are relevant to decision-making. These models include, for example, the facial recognition model to verify a worker's identity (Sullivan, 2016) or load balancing models (computer technology that distributes network traffic among various servers to avoid one getting overwhelmed), which can disconnect workers for hours, or in extreme cases, entire days if the workload requires fewer workers than the number who are available at particular times (Zipperer et al., 2022).

Account Creation

When a worker creates an account, they are subject to checks to verify their identity. These checks are arbitrary and opaque since the exact data collected, the internal verification processes, data access control, data retention time, people with access to data, and data removal processes are not disclosed. This lack of transparency, combined with the constant surveillance of workers, brings up various privacy concerns and issues, including whether the data collection and monitoring are truly appropriate, whether personal data are used correctly and how this information is used in decision-making (Sannon et al., 2022).

User Login

Once the worker has cleared the company's checks and is approved to provide the service, they must log in to the platform daily to receive jobs. Workers must then interact with the login and session management algorithms, which not only authenticate workers but also balance the demand for services and available workers. Session management algorithms are used to keep online only the service providers who are needed to fulfil demand from the final users. If there are too many workers, the algorithms can temporarily or permanently disconnect workers without notice and for any reason (Safak & Farrar, 2021).

The result is job instability due to unpredictable access to the platform and the impossibility of being able to get jobs on it. Moreover, in addition to the username and password, the platforms use workers' biometric data when they log in to compare them against data provided when they signed up on the platform. Although this may seem to be a simple process, it actually ends up being the reason why workers are denied access due to data entry or comparison errors. It is a system with many different parts that can malfunction and which depends mainly on the capacity of the worker's mobile device and internet connection. If any of these components fails, the worker pays the consequences by being denied access to work (Zipperer et al., 2022).

Authentication systems can also be used to confirm that workers are where they say they are, and workers may be required to take photos and save their coordinates on their mobile device so the data can be compared with internal and external databases. This process is extremely problematic because in many cases, delivery workers do not have devices that are powerful enough to precisely capture photos and coordinates, which leads to the app blocking them or even removing them permanently (Aguirre Reveles, 2020).

Classification and Ranking

Once workers manage to enter the platform and are able to accept a job, they face additional algorithms that classify and rank them to determine whether or not they are chosen for the job. To decide which worker will perform the new job, the models take several factors into account, from the distance between the worker and the destination point to the type of vehicle and the average speed or traffic in the area and even the workers' previous behaviour, time lag in accepting new jobs and how many new jobs they accept. These algorithms determine the jobs that the workers will take during the day, and so their opportunities to earn an income (L. Zhang et al., 2023).

Contrary to the idea that these models and algorithms give workers job autonomy, these capacities turn models and algorithms into modern supervisors that reward or punish algorithmically, without offering workers an explanation or direct and assertive communication. This results in a sort of shadow banning, where workers may notice they receive fewer jobs and less income without understanding why. Additionally, the continuous monitoring of workers by algorithms – even when disguised as fun challenges (known as gamification) – has negative effects on workers' mental health (Kadolkar et al., 2024; A. Zhang et al., 2022; L. Zhang et al., 2023).

Price Setting

One of the main questions about how gig economy platforms operate concerns the methods they use to set their prices and distribute earnings. The way the base rate of services is set depends not only on the service workers provide but also the profile of the final consumer, the business selling the products, the zones where the different parties are located, the transaction history, the complaints history, fraud and violence, and other variables. The exact list of factors and their importance in setting the final price is a well-guarded secret of the platforms, each undoubtedly with its own formula. Furthermore, it is unclear how much of what consumers pay actually ends up in delivery workers' pockets because the platforms take a commission, charge extra fees and subtract tax withholdings before paying workers.

As a result, workers do not know how much money they will make for the jobs they accept, which violates the right of workers to earn a decent living. Additionally, machine learning and AI models are biased in that they assign higher prices in areas where certain populations live – such as people of colour and low-income and young residents (Pandey & Caliskan, 2021) – a criterion that is probably also applied to workers. In other words, the algorithms and models do not treat all workers or users equally.

Moreover, it is important to remember that platforms commonly use dynamic pricing based on demand among final users. This means that when demand for services is high and there are few workers available, the cost of the service goes up. Dynamic pricing is set as a multiplier applied to a base price. For example, if the base price is $100 and dynamic pricing is not in effect, the multiplier is 1, and the final price is $100. However, if demand for the service rises by 20% and the algorithm sets a linear multiplier (in this case, 1.2), the base price times the multiplier brings the price to $120. Dynamic price setting on each platform likely takes many other factors into account, but this simple example illustrates the lack of transparency around such decisions.

Performance Ranking
Algorithms not only manage specific transactions but also process aggregate data to set metrics that rank worker performance, which determines the incentives and penalties that impact their incomes and eligibility for future jobs. Workers must deal with such metrics on a daily basis, which can range from the percentage of rejected job requests to the average time per job, and workers have had to learn to adapt to resist and manipulate the models and algorithms to work in their favour. This resistance to being constantly under surveillance and measured is similar to learning to cope with a new profession, boss or workplace and is often a source of uncertainty that can lead to workers to leave the platforms (A. Zhang et al., 2022).

Generally speaking, the development of digital platforms is based on the use of generic models that have been pre-trained on enormous databases that are ready to be reused in any location or situation. While this supports the platforms' growth, it also compounds the biases and blind spots that reduce statistical representation and end up violating workers' rights. For example, if a model is trained on data from workers in India, the model will set metrics and expectations based on those data, which will not necessarily correspond to the behaviours of workers in Mexico. To prevent this type of bias and blind spots, platforms must train their models using local data that give preference to the applicability of decisions made by their systems.

Like in any other business, problems can arise in gig work – mechanical malfunctions, traffic accidents, illegal activity, fraud, etc. – but workers have no recourse in these cases because they are not insured during the time they are not engaged in work for the platform. In fact, aside from a few very specific situations, insurance coverage provided by platforms is limited to third-party damages, while workers must personally cover damage to their own vehicle, even when they are online on the platform (Insurance for Rideshare and Delivery Drivers, n.d.). As a consequence, workers lose access to both their work tool and part of their income that must go towards necessary repairs. In the event of fraud, it is unclear whether the platforms protect workers and, just as with insurance coverage, conditions vary depending on the jurisdiction.

As such, rather than representing progress that guarantees fair conditions, algorithmic management has proven to be a tool that sustains and deepens job insecurity among workers. Platform owners and their shareholders reap significant profits by optimizing costs and maximizing earnings, but this comes at the cost of stable employment, economic security and the dignity of those who support their daily operations. The opaque and unilateral implementation of the systems intensifies workers' dependence on the platforms while also exposing them to working conditions plagued by uncertainty, constant surveillance and the lack of basic employment rights. Algorithmic management leads to the systematic violation of labour rights because it reproduces exploitation dynamics in an apparently modern yet deeply unequal digital market.

Algorithmic Management and its Implications for Labour Rights

The gig economy offers jobs through a discontinuous or short-term modality, where workers use the platform to search for jobs in a specific location but without knowing the relationship of subordination. This form of work has profoundly changed labour markets: while it offers flexibility and autonomy, it also involves labour rights violations.

The widespread use of algorithms to manage and supervise workers leads to increasingly precarious working conditions. In particular, violations of at least five fundamental rights have been identified: the right to work, to privacy, to fair remuneration, to safe working conditions and to non-discrimination. Moreover, it is important to identify the mechanisms of

control, lack of transparency and data extraction that sustain and perpetuate inequalities in this sector.

Right to Work

The algorithms used in the gig economy play a decisive role in assigning work in this sector, based on the many factors discussed previously. However, the automated decision-making process is extremely opaque, which can lead to algorithmic discrimination where workers may receive fewer job assignments due to arbitrary metrics or their geographic location instead of their abilities and opportunities (Widjaya, 2024).

As such, the right to work is violated by algorithmic control. Gig workers say the algorithms exert control in a way that makes them feel trapped in a rigid system that does not take individual circumstances or preferences into account (Lang et al., 2023; L. Zhang et al., 2023).

There is a paradox of autonomy associated with digital platforms (Jaharri et al, 2019). In other employment sectors, hours are more strictly managed in that workers must put in a certain number of hours and work in a specific setting, while an algorithmically controlled digital setting gives workers the flexibility and autonomy to decide where and when they work. However, a system where workers are continuously tracked and monitored through algorithmic systems can cause tension and anxiety in workers, making them feel less autonomous (Lee, 2018).

Algorithmic management creates irregular working hours, excess work and social isolation for workers (Shapiro, 2018), which is an attack on the right to work due to labour instability, as illustrated by the following statement from a gig delivery worker in Mexico:

> *"If you're on there [the digital platforms], you can't do anything else. What I mean is that I come back tired, and then, if I want to do my homework, it's really hard, or I'm already tired, I'm sleepy and I stop early, I can't get into it with the same [. . .] enthusiasm as I could when I worked in an office, because there I had time; there I did it and it was fine, but like this, I can't."*
>
> – Statement from a young female delivery worker in Mexico.

Similarly, prioritizing algorithmic efficiency and maximizing earnings overlooks the issue of caregiving, which is a key responsibility for many of the people who work in this sector, and especially women. Long, unpredictable days limit the ability of workers to care for the people in their life that require it. This reality puts mainly women in a position of vulnerability not only in terms of their labour rights but also in terms of their ability to achieve work–life balance (Centeno Maya et al., 2022).

Right to Privacy

The right to privacy is recognized in several international texts, such as Article 12 of the Universal Declaration of Human Rights and Article 17 of the International Covenant on Civil and Political Rights. This right protects people from arbitrary interference in their private life and ensures the protection of their personal data. But in the gig economy this right is violated in multiple ways due to the massive collection of data, especially personal data, which is a major concern for workers in this sector (Sannon et al., 2022).

The digital platforms maintain that they are promoting and preserving workers' freedom. However, this freedom rests on the algorithmic control they exercise and the exploitation of workers' and users' personal data under a level surveillance reminiscent of an Orwellian dystopia. Additionally, digital platforms maintain a lack of transparency around how the al-

gorithms work, which means that workers and users neither know nor understand for what or how their personal data are used, if it is shared with third parties or if it is sufficiently protected (Tsaaro Consulting, 2023).

Breaches of workers' data are also a violation of their right to privacy. In 2016 a data breach occurred in which the hackers downloaded the names, email addresses and mobile phone numbers of more than 25 million gig workers around the world in addition to 600,000 driving licences of drivers and delivery workers in the United States. These security issues are critical for workers because, in addition to the breaches, the company responsible did not inform the drivers and riders about it until a year after the fact (Khosrowshahi, 2017).

Right to Fair Remuneration

The right to fair remuneration is included in Article 23 of the Universal Declaration of Human Rights, which states that everyone has the right to equal pay that allows them to have a dignified existence for themselves and their family. In the gig economy, this right is challenged by the algorithmic management that determines the prices, incentives and commissions for workers without transparency or clarity (Dubal, 2023).

Algorithmic management plays a key role in determining workers' pay (Abraham, 2023). The algorithms set the prices, commissions and bonuses based on various factors, including demand, job complexity and the worker's profile (Sharma, 2024). The algorithms rely on dynamic pricing, which adjusts prices based on fluctuations in supply and demand and which can result in significant variations in workers' pay.

This system leads people to believe that workers can earn a good income if they work more hours. But in fact, this is untrue because the platforms require workers to work during specific hours and in specific situations to avoid losing income (Sharma, 2024). There have even been cases where workers who work more hours often earn a lower hourly wage (Cook et al., 2021). This results in an unpredictable, variable income specific to each worker, which violates workers' right to equal pay for equal work.

This situation is reflected in a statement from a delivery worker when asked about whether she earns enough income:

> *"To a certain extent, yes, I can cover my basic needs – food, etc. – but sometimes, for example just yesterday, I did two trips and earned 20 pesos [US$1], but then I earned 40 [US$2] for the whole day working morning till night."*
>
> – Statement from a young female delivery worker in Mexico.

Some gig delivery workers even compare the experience with gambling, because while they keep waiting to hit the jackpot, the algorithm only gives them enough trips to keep them active but with low per-trip earnings (Abraham, 2023).

Workers face an opaque, unequal system that perpetuates unstable earnings. The use of dynamic pricing, personalized rates and gamification of work not only violates the right to fair pay but also deepens workers' economic and social inequalities.

Right to Safe Working Conditions

This right is included in various international texts, including the Occupational Safety and Health Convention, 1981 (No. 155) of the International Labour Organisation and Article 23 of the Universal Declaration of Human Rights. This right guarantees that workers **can** perform their work in an environment that protects their physical, mental and social health,

minimizes occupational risks and promotes their well-being. In the gig economy, this right is violated because algorithmic management prioritizes efficiency and profitability over workers' safety (De Stefano & Taes, 2023).

The algorithms used by digital platforms determine the workload as well as the delivery times and routes. As we have already seen, this puts workers in unsafe situations: in order to meet the delivery times, they end up breaking traffic rules and putting themselves in harm's way (L. Zhang et al., 2023). The absence of human oversight in these decisions exacerbates the unsafe conditions and requires workers to make choices they should not be faced with, such as continuing to have access to jobs on the platform or ensuring their well-being (Abraham, 2023).

Delivery workers also face possible violence from customers. This statement from a delivery worker is one example:

> *"[When making the delivery] within the 10 minutes – a couple of seconds before – a car arrived and started yelling at me, asking why I didn't want to deliver their pizza, to give them their pizza, and they tried to snatch it away from me, and I said, 'Here's your pizza, I don't want any problems.' And they said, 'You just wait, I'm going to find you and make them suspend your account, I'm going to do everything I can to hurt you.' And I didn't have much information, so I wondered if I had done something wrong, but I was following the instructions I received."*
>
> – Statement from a young female delivery worker with children in Mexico.

Digital platforms seek to maximize the number of jobs completed by workers, with algorithmic management acting as an intermediary whose main aim is to ensure that each and every job requested by customers is completed on time and at the highest quality (Bérastégui, 2021). The phenomenon of overwork in the gig economy is sustained through algorithmic management and digital surveillance, where the main objective is to coordinate and maximize the workload based on various factors in real time, an optimization process that has been identified as a source of overwork (Poutanen et al., 2021).

Another important factor is the absence of human resource management by digital platforms. Various studies (Bellesia et al., 2019; Deng et al., 2016) have shown a lack of training, disregard in handling complaints and insufficient communication with platform workers, which not only creates a risk of occupational accidents and pressure due to the algorithm but also a lack of attention given to workers. This lack of support worsens the precarious working conditions and further violates labour laws.

The right to safe working conditions in the gig economy has been violated by algorithmic management due to the lack of human oversight and the irresponsible behaviour of platforms with regard to workers. Algorithmic management perpetuates a workplace where risks are shifted to individual workers and the platforms are insulated from any responsibility, leaving workers in a constant fight to balance their safety with their need to earn a living.

Right to Non-Discrimination

The right to non-discrimination in the workplace is protected by several international texts, including the International Covenant on Economic, Social and Cultural Rights and the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW). These policy frameworks protect people from discrimination on the basis of gender, race and any other characteristic to promote equal rights in all spheres.
As previously mentioned, algorithmic management seeks to assign the maximum number

of jobs to maximize earnings for platform owners and shareholders, which is why the design of these algorithms has mainly undermined groups whose rights have been historically violated by the owners of the algorithmic management system. The algorithms also promote discrimination because the people who design them have their own personal biases and prejudices when determining the metrics used to assign jobs (Lawyer Monthly, 2024; Skelton, 2021). Such biases in the gig economy affect women and historically marginalized groups through the various jobs available through the digital platforms (University of Southampton, 2024).

According to Lawyer Monthly, "The relationship between the worker and the platform in the gig economy is often ambiguous, making it difficult to identify a clear employer in discrimination claims. Thus, gig workers will face difficulties in filing discrimination complaints or seeking legal redress since there might not be a clear entity to blame" due to the use of algorithmic management (Lawyer Monthly, 2024).

One of the effects of discrimination from algorithmic management is that women receive fewer well-paid jobs than men, due to algorithmic preferences that favour men for historically male roles (Victorian Government, 2020). Some communities may also be downgraded for job assignments due to worker classification. This system reinforces existing biases, assigning more or less work to certain profiles in the gig economy (Botelho et al., 2023). For example, ratings systems have been found to reflect and exacerbate user prejudices. Non-white workers experience a minority rating gap of 80% compared to white workers, which negatively impacts their earnings (Teng et al., 2023).

The lack of clarity in the relationship between gig workers and digital platforms complicates discrimination claims and leaves workers unprotected (Baeza Flores & Marquez Amaro, 2023). The system itself also makes it difficult for workers to band together. It is crucial for humans to be involved in critical decisions, such as account terminations and performance evaluations. These measures would help mitigate algorithmic bias and ensure fair treatment of workers (Gussek et al., 2023; Sharma, 2024).

Finally, the right to non-discrimination in the gig economy has been profoundly compromised by algorithmic management, which reinforces and amplifies historic biases in job assignments, performance evaluations and account terminations. In this context, regulatory frameworks must be adopted to ensure human oversight in key decisions, and policies must be promoted that address gender and race gaps that are worsened by the algorithms. This is the only way to create a more equal and just working environment and eliminate the barriers that perpetuate structural discrimination in this sector.

Public Policy Recommendations

Regulate Algorithmic Management and its Lack of Transparency

As discussed in previous sections, algorithmic management in the gig economy poses significant risks for workers' rights, especially in terms of privacy and working conditions. Personal data are collected constantly and without any transparency, and they are used to rank and assess workers from the start to the end of their relationship with the platform. This not only perpetuates discrimination but creates uncertainty and job insecurity.

Countries must develop regulatory frameworks that require digital platforms to implement transparency mechanisms in their decision-making processes. Regulations should require companies to open up their algorithm coding, disclose the variables used to train models and report the impact of the algorithms on working conditions. **Platforms should be requi-**

red to provide clear information on how their algorithms work and what their impact is on working conditions; if they do not, governments should take action to enforce compliance.

Fund social security

Gig workers should be guaranteed access to social security as a universal right. In Mexico, for example, a recent legislative proposal is considering a contribution shared among platforms, workers and the federal government and represents a significant step forward. However, it is also critical to enact progressive, comprehensive tax reforms to ensure that digital platforms pay their fair share of taxes and social security contributions in the country. Doing so would adequately and sustainably fund the social security system and prevent workers from disproportionately bearing the cost. These measures, combined with a system of sanctions for the platforms that violate the regulation, would lay the foundation to guarantee decent working conditions.

New regulatory frameworks must also take social security into account. In the case of Mexico, social security is the shared responsibility of the platforms, workers and the federal government. A new law was recently put forward that would protect workers and give them access to social security without affecting their earnings, since workers would pay only the equivalent of USD 10 a month. In addition to having healthcare coverage, they would also gain access to housing, mortgage loans, and other social security services offered by the Mexican government. This type of legislation, which centres around workers, could serve as a strong example for other jurisdictions, since it even has the support of major platform companies in the country (Marath Bolaños [@marathb], 2024) (#NiUnRepartidorMenos [@repartidorr], 2024). To achieve these advances, **a system must be established to levy sanctions on platforms that violate regulations, including fines and the possibility of suspending their operations.**

Promote collective bargaining and support networks

Despite platforms' efforts to discourage workers from organizing, the creation of unions, cooperative organizations and other groups is key for workers to be able to negotiate for better conditions (Mendonça & Kougiannou, 2023). These organizations not only strengthen workers' abilities to demand labour rights but also create communities in which workers can share their experiences and develop collective knowledge. In the case of social security, international studies (Zipperer et al. 2022) have confirmed that gig workers face significant disadvantages in terms of healthcare and pension plans. Collective bargaining is the only way to dismantle these structural conditions (Johnston & Land-Kazlauskas, 2018). Governments must guarantee the right to freedom of association and protect workers from retaliation by platforms.

Increase human contact and effective reporting mechanisms

The previous sections showed that algorithmic management impacts not only working conditions but also workers' emotional well-being. Platforms must implement mechanisms for human contact, especially in critical situations, such as road accidents and cases of violence. They must also develop effective reporting systems, overseen by competent staff, so that workers can report cases of workplace harassment, abuse or safety issues. These measures would address the feeling of helplessness that is prevalent in the sector and guarantee faster and more effective responses in difficult situations.

Prohibit algorithmic wage discrimination

Algorithmic wage discrimination, as well as discrimination by algorithmic management, must be explicitly prohibited. These practices perpetuate historic inequality while also creating uncertainty around earning potential and worsening labour problems for workers. Clear regulations on these issues would eliminate the gamification of work and protect workers, especially those with the lowest wages, from improper use of their personal data. These measures are essential to ensure the right to fair and equal pay.

While the gig economy continues to expand and transform the labour landscape, the power divide between digital platforms and workers is deepening. In 2023, the global gig economy generated over than USD 200 billion, with estimates for 2032 reaching USD 455 billion (Sharma, 2024). Digital platforms continue increasing their profit margins by making the most of precarious working conditions. This disparity reflects a pressing need to regulate a sector that perpetuates structural inequality and violates labour rights under the façade of technological innovation.

For the gig economy to fulfil its promises of flexibility and freedom, it must be reined in by regulations that protect those who keep the system operating with their labour. It is the responsibility of governments, in cooperation with labour organizations, to set out a regulatory framework that not only prevents abuses but also promotes social justice and equality in one of the most precarious employment sectors of our times.

References

Abraham, R. (2023, January 25). Pay Algorithms Make Working in the Gig Economy Feel Like "Gambling," Study Says. *VICE.* https://www.vice.com/en/article/pay-algorithms-make-working-in-the-gig-economy-feel-like-gambling-study-says/

Aguirre Reveles, R. (2020). *Gig economy: Precariedad laboral y traslado de riesgos*. Heinrich Böll Stiftung. https://mx.boell.org/es/2020/11/11/gig-economy-precariedad-laboral-y-traslado-de-riesgos

Baeza Flores, M. E., & Márquez Amaro, R. (2023). Gig Economy, las plataformas digitales para el trabajo: ¿flexibilidad o precariedad laboral? *Ciencia Latina Revista Científica Multidisciplinar, 7* (2). https://doi.org/10.37811/cl_rcm.v7i2.6211

Bellesia, F., Mattarelli, E., Bertolotti, F., & Sobrero, M. (2019). Platforms as Entrepreneurial Incubators? How Online Labor Markets Shape Work Identity. *Journal of Managerial Psychology, 34(4), 246–268.* https://doi.org/10.1108/JMP-06-2018-0269

Bérastégui, P. (2021). *Exposure to Psychosocial Risk Factor in the Gig Economy: A Systematic Review. E*TUI Research Paper - Report 2021.01, Available at SSRN: https://ssrn.com/abstract=3770016

Botelho, T. L., Sudhir, K., & Teng, F. (2023, August 14). *Ratings Systems Amplify Racial Bias on Gig-Economy Platforms | Yale Insights.* https://insights.som.yale.edu/insights/ratings-systems-amplify-racial-bias-on-gig-economy-platforms

Centeno Maya, L. A., Tejada, A. H., Martínez, A. R., Leal-Isla, A. L. R., Jaramillo-Molina, M. E., & Rivera-González, R. C. (2022). Food Delivery Workers in Mexico City: A Gender Perspective on the Gig Economy. *Gender & Development, 30(3), 601–617.* https://doi.org/10.1080/13552074.2022.2131253

Cook, C., Diamond, R., Hall, J. V., List, J. A., & Oyer, P. (2021). The Gender Earnings Gap in

the Gig Economy: Evidence from over a Million Rideshare Drivers. *The Review of Economic Studies,* 88(5), 2210–2238. https://doi.org/10.1093/restud/rdaa081

De Stefano, V., & Taes, S. (2023). Algorithmic Management and Collective Bargaining. *Transfer: European Review of Labour and Researc*h, 29(1), 21–36. https://doi.org/10.1177/10242589221141055

Deng, X., Joshi, K. D., & Galliers, R. (2016). *The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful Through Value Sensitive Design*. 279–302.

Dubal, V. (2023). *On Algorithmic Wage Discrimination* (SSRN Scholarly Paper No. 4331080). Social Science Research Network. https://doi.org/10.2139/ssrn.4331080

Duggan, J., Sherman, U., Carbery, R., & McDonnell, A. (2020). Algorithmic Management and App-Work in the Gig Economy: A Research Agenda for Employment Relations and HRM. *Human Resource Management Journal*, 30(1), 114–132. https://doi.org/10.1111/1748-8583.12258

Gussek, L., Grabbe, A., & Wiesche, M. (2023). Challenges of IT Freelancers on Digital Labor Platforms: A Topic Model Approach. *Electronic Markets*, 33(1), 55. https://doi.org/10.1007/s12525-023-00675-y

Uber (s/f). *Insurance for Rideshare and Delivery Drivers.* Uber. Retrieved November 15, 2024, from https://www.uber.com/us/en/drive/insurance/

Jarrahi, M. H., Sutherland, W., Nelson, S. B., & Sawyer, S. (2020). Platformic Management, Boundary Resources for Gig Work, and Worker Autonomy. Computer Supported Cooperative Work (CSCW), 29(1), 153–189. https://doi.org/10.1007/s10606-019-09368-7

Johnston, H., & Land-Kazlauskas, C. (2018). Organizing On-Demand Representation, Voice, and Collective Bargaining in the Gig Economy. *ILO Working Papers* 994981993502676. https://ideas.repec.org//p/ilo/ilowps/994981993502676.html

Kadolkar, I., Kepes, S., & Subramony, M. (2024). Algorithmic Management in the Gig Economy:A Systematic Review and Research Integration. *Journal of Organizational Behavior,* 1–24. https://doi.org/10.1002/job.2831

Khosrowshahi, D. (2017, November 21). Information about 2016 *Data Security Incident.* Uber. https://help.uber.com/driving-and-delivering/article/information-about-2016-data-security-incident?nodeId=0ded7de4-ed4d-4c75-a3ee-00cddeafc372

Lang, J. J., Yang, L. F., Cheng, C., Cheng, X. Y., & Chen, F. Y. (2023). Are Algorithmically Controlled Gig Workers Deeply Burned Out? An Empirical Study on Employee Work Engagement. *BMC Psychology,* 11(1), 354. https://doi.org/10.1186/s40359-023-01402-0

Lawyer Monthly. (2024, November 11). Employment Discrimination in the Gig Economy. Lawyer Monthly. https://www.lawyer-monthly.com/2024/11/employment-discrimination-in-the-gig-economy/

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1603–1612. https://doi.org/10.1145/2702123.2702548

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and

emotion in response to algorithmic management. *Big Data & Society*, 5(1). https://doi.org/10.1177/2053951718756684

Madariaga, J., Cañigueral, B., & Popeo, C. (2018). Claves para entender la economía colaborativa y de plataformas en las ciudades. *CIPPEC*. https://www.cippec.org/publicacion/claves-para-entender-la-economia-colaborativa-y-de-plataformas-en-las-ciudades/

Marath Bolaños [@marathb]. (2024, November 4). *La @STPS_mx, las tres principales empresas de plataformas en México (@Uber_MEX, @DiDi_Mexico, @RappiMexico) y el Consejo Coordinador Empresarial (@cceoficialmx) concluimos las mesas de trabajo, donde coincidimos en que la iniciativa que mandará la presidenta @Claudiashein*, https://t.co/I69EpOZhGk [Tweet]. Twitter. https://x.com/marathb/status/1853584119801823624

Mendonça, P., & Kougiannou, N. K. (2023). Disconnecting Labour: The Impact of Intra-platform Algorithmic Changes on the Labour Process and Workers' Capacity to Organise Collectively. New Technology, Work and Employment, 38(1), 1–20. https://doi.org/10.1111/ntwe.12251

#NiUnRepartidorMenos [@repartidorr]. (2024, November 8). Q*uedamos a la expectativa, la responsabilidad del futuro de nuestro trabajo ahora está del lado de la cámara de Diputados Federal. Los esfuerzos de los repartidores que se manifestaron el 14 y 30 de octubre lograron cambios sustanciales en la propuesta: - Libertad para aceptar* https://t.co/Wha79gyGFJ [Tweet]. Twitter. https://x.com/repartidorr/status/1854693540422721830

Pandey, A. & Caliskan, A. (2021). Disparate Impact of Artificial Intelligence Bias in Ride-hailing Economy's Price Discrimination Algorithms. I*n Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '21). Association for Computing Machinery, New York. 822–833. https://doi.org/10.1145/3461702.3462561

Poutanen, S., Kovalainen, A., & Rouvinen, P. (eds.) (2021). *Digital Work and the Platform Economy: Understanding Tasks, Skills and Capabilities in the New Era*. https://www.routledge.com/Digital-Work-and-the-Platform-Economy-Understanding-Tasks-Skills-and-Capabilities-in-the-New-Era/Poutanen-Kovalainen-Rouvinen/p/book/9781032082721

Safak, C. & Farrar, J. (2021) *Managed by Bots Report.* Worker Info Exchange. Retrieved October 20, 2024, from https://www.workerinfoexchange.org/wie-report-managed-by-bots

Sanghro, M.A. (2023). *The Gig Economy, the Digital Divide, and Developing Countries* – Maseconomics. Mas.Economics. https://maseconomics.com/the-gig-economy-the-digital-divide-and-developing-countries/

Sannon, S., Sun, B., & Cosley, D. (2022). Privacy, Surveillance, and Power in the Gig Economy. *CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3491102.3502083

Shapiro, A. (2018). Between autonomy and control: Strategies of arbitrage in the "on-demand" economy. *New Media & Society*, 20(8), 2954-2971. https://doi.org/10.1177/1461444817738236

Sharma, R. (n.d.). *Gig Based Business Market Research Report 2032* (p. 283). Retrieved December 17, 2024, from https://dataintelo.com/report/global-gig-based-business-market

Sharma, R. (2024). P*rotecting Worker Earnings in the Technology-Driven Gig Economy: Policy Approaches for Sustainable Stability and Fairness*. 1–4. https://sdgs.un.org/sites/default/files/2024-05/Sharma_Protecting%20Worker%20Earnings%20in%20the%20Technology-Driven%20Gig%20Economy.pdf

Skelton, S. K. (2021, December 15). Gig Economy Algorithmic Management Tools 'Unfair

and Opaque'. *ComputerWeekly.com*. https://www.computerweekly.com/news/252511001/Gig-economy-algorithmic-management-tools-unfair-and-opaque

Sullivan, J. (2016). *Selfies and Security*. *Uber Newsroom*. Retrieved November 11, 2024, from https://perma.cc/GBN4-U5C3

Teng, F., Botelho, T. & Sudhir, K. (2023). Can customer ratings be discrimination amplifiers? Evidence from a gig economy platform. https://insights.som.yale.edu/sites/default/files/2023-08/Customer_Prejudice_and_Minority_Earnings_Gap.pdf.

Tsaaro Consulting. (2023, June 9). *Data Privacy Concerns for Gig and Platform Workers.* https://www.linkedin.com/pulse/data-privacy-concerns-gig-platform-workers-tsaaro/

University of Southampton. (2024, June 6). *Report Proposes New Rights to Protect Workers from 'Unfair, Unaccountable and Uncaring' Algorithms*. https://www.southampton.ac.uk/news/2024/06/new-rights-to-protect-workers-from-algorithms.page

Victorian Government. (2020). *Report of the inquiry into the Victorian On-Demand Workforce* (p. 228). The State of Victoria. https://oia.pmc.gov.au/sites/default/files/posts/2023/09/3%20Report%20of%20the%20Inquiry%20into%20the%20Victorian%20On-Demand%20Workforce%20%28PDF%29.pdf

Widjaya, I. (2024, August 12). The Gig Economy's Hidden Workforce: Understanding the Role of Algorithms. *Noobpreneur.com.* https://www.noobpreneur.com/2024/08/12/the-gig-economys-hidden-workforce-understanding-the-role-of-algorithms/

Zhang, A., Boltz, A., Wang, C. W., & Lee, M. K. (2022). Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3491102.3501866

Zhang, L., Yang, J., Zhang, Y., & Xu, G. (2023). Gig Worker's Perceived Algorithmic Management, Stress Appraisal, and Destructive Deviant Behavior. *PLOS ONE*, 18(11), e0294074. https://doi.org/10.1371/journal.pone.0294074

Zipperer, B. et al. (2022). *National survey of gig workers paints a picture of poor working conditions, low pay.* Economic Policy Institute. Retrieved November 11, 2024, from https://www.epi.org/publication/gig-worker-survey/

8. Algorithms that Flag and Penalise. Automating Social Welfare: What is at Stake?

Pablo Jiménez Arandia
Translation: Kim Causer

Efficiency or optimisation of limited public resources are two of the arguments for the implementation of algorithmic management systems for some public services and social protection tools. However, the history of these experiments, both in Europe and the United States, has been marred by serious scandals, the consequences of which have been extremely serious for many citizens. These tools have sometimes inherited discriminatory biases that lead them to make critical decisions based on criteria of origin or race, for example. In the same way, human relinquishment in favour of automation has generated procedures that become dead ends, increasing the defencelessness of citizens.

In other cases, algorithmic mechanisms systematise a surveillance apparatus that goes beyond the limits of fundamental rights such as the right to privacy or non-discrimination. Previous experiences show that efforts have not been oriented towards improving the efficiency of citizen services, but rather towards fighting against the alleged fraud that is an undeniable alibi when talking about the limited resources available to guarantee basic services. In this way, a logic of suspicion has spread, recurrently targeting the most vulnerable groups instead of reinforcing them, weakening the social protection system through arbitrariness, abuse and discrimination.

A snapshot of failed automation

'The atmosphere at the meetings with the municipality is terrible.' That's how Imane[1] described being interrogated by two city officials in Rotterdam in October 2021. The local government of this Dutch city suspected that Imane, a 44-year-old mother of three with chronic health problems, was lying to them about her income. She was therefore not entitled to the social benefits she had been receiving for years.

The trigger, or suspicion, on which the city officials based their investigation, came from a machine. A machine learning algorithm had flagged this migrant woman, a resident of a working-class neighbourhood in Rotterdam, as a potential benefit cheat.

During her interrogation, Imane remembers that one of the civil servants raised their voice. They loudly accused her of taking in the wrong bank document and pressured her to log in to her online account to show them her statements. Imane refused. So the city council suspended her social benefits, which she used to pay her rent and buy food. Payments were resumed two days later when Imane sent the correct bank statements. 'It took me two years to recover from this. I was destroyed mentally.'

Imane's scrutiny can be summarised in just two words: 'high risk'. This was the verdict reached by the algorithm that Rotterdam City Council started using in 2017 to analyse the city's 30,000 welfare recipients. This computer program had been trained using 12,707 previous investigations and was designed, on paper, to estimate the risk that someone was lying about their social and economic reality.

However, the algorithm in question had several issues. Years later, an investigative report[2] revealed how the risk level assigned by the program increased significantly if the profile being analysed had certain characteristics. Most of them were linked to gender, nationality and age biases. Being a young mother, not speaking Dutch fluently or having trouble finding work seriously affected the risk scores. So single mothers like Imane were always classed as high risk.

Findings revealed by the group of journalists that investigated the system, confirmed as valid by Rotterdam's authorities, proved its discriminatory nature.[3] In fact, the project was stopped in 2022 after an internal assessment found that the algorithm model could never be 100 percent 'free from bias'.

'This situation is undesirable in itself, especially when it comes to variables that carry a risk of bias based on discriminatory grounds such as age, nationality or gender. Your findings also demonstrate these risks,' stated Rotterdam City Council.

But what happened with this algorithm used in the Netherlands' second city is not an isolated case. In recent years, governments across the world have implemented AI tools and algorithms to profile benefit recipients.

As the public service sector's use of automation became more widespread, so did the documented cases of its misuse. As we will see throughout this chapter, such projects have had harmful consequences for citizens. Especially the most vulnerable.

1 Imane's testimony, her name changed to protect her anonymity, is taken from M. Burgess, E. Schot and G. Geiger. (6 March 2023). *This Algorithm Could Ruin Your Life. WIRED*. https://www.wired.com/story/welfare-algorithms-discrimination/) [paywall] Her story and other details of this case are the result of the Suspicion Machines investigation (Lighthouse Reports, 2023; https://www.lighthousereports.com/investigation/suspicion-machines/), a project coordinated by Lighthouse Reports on which the media from several countries participated. The following pages include some of the stories and revelations from the study.

2 Lighthouse Reports. (2023). *Suspicion Machines.* https://www.lighthousereports.com/investigation/suspicion-machines/)

3 A detailed description of the investigation's methodology, its findings and consequences can be found in Lighthouse Reports. (2023). *Suspicion Machines Methodology*. https://www.lighthousereports.com/suspicion-machines-methodology/).

An unsupervised 'vicious circle' of discrimination

In Europe, the Netherlands has been the centre of several technology experiments that ultimately failed. Also in 2021, just a few months before Imane was put through her harrowing interrogation in Rotterdam City Council's offices, a national scandal with unprecedented political consequences unfolded.

In January of the same year, the government of the conservative prime minister Mark Rutte resigned en masse after it was disclosed that 20,000 families had been wrongly accused of benefit fraud.[4] *The toeslagenaffaire* – or Dutch childcare benefits scandal – financially devastated thousands of families in the Netherlands, given the extortionate amounts the state had claimed from those affected. Many families even lost custody of their children during the process.

The battle led by a handful of lawyers and families they represented caused enough of a stir that the Dutch parliament launched an inquiry. That is how it became clear that an algorithm was behind the flagged profiles. A computer program selected the households that were inspected based on variables like nationality or the origin of their occupants. Around 70% of the families were, in fact, migrants or children of migrants.

People like Chermaine Leysner[5] endured the crudeness of this automated flagging. At the time, Leysner had three children under six and was studying at university. In 2012 she received a letter from the Dutch tax authorities asking her to return the child benefits she had received the four years previous. The tax bill was over €100,000, an inconceivable amount for a family like hers.

The debt-induced stress and her mother becoming seriously ill made Leysner depressed, ultimately causing her to separate from the father of her children. 'I was working like crazy so I could still do something for my children like give them some nice things to eat or buy candy. But I had times that my little boy had to go to school with a hole in his shoe,' she explained.

The digital newspaper POLITICO described this scandal as a 'warning for Europe over the risks of using algorithms'. Meanwhile, Amnesty International published a hard-hitting report, entitled *Xenophobic machines*,[6] highlighting the system's racist nature. 'Thousands of lives were ruined by a disgraceful process which included a xenophobic algorithm based on racial profiling. The Dutch authorities risk repeating these catastrophic mistakes as human rights protections are still lacking in the use of algorithmic systems,' said Merel Koning, Advisor on Technology and Human Rights at the non-governmental organization (NGO).

This study examines how the design of the algorithm reinforced existing institutional biases by linking nationality and ethnicity with potential criminal activity. But there was an added issue: being based on machine learning techniques, the discriminatory nature was amplified by a system that evolved from experience; and did so without sufficient human supervision. This 'vicious circle of discrimination' meant that those who were not Dutch nationals were flagged more often than those who were.

This algorithm is a clear example of a 'black box'. This term refers to intrinsically opaque systems, where the software calculations and logic behind them are unknown to the people who receive the results. So, when a person was flagged as being potentially fraudulent, the civil servant had to manually review their case but lacked information as to why the algorithm had reached that score.

4 Amnesty International. (25 October 2021). *Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms*. https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/

5 Chermaine Leysner's testimony is taken from Dutch *scandal serves as a warning for Europe over risks of using algorithms* (POLITICO, 2022; https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/)

6 *Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal* (Amnesty International, 2021; https://www.amnesty.org/es/documents/eur35/4686/2021/en/)

‘The black box system resulted in a black hole of accountability, with the Dutch tax authorities trusting an algorithm to help in decision-making without proper oversight’, highlighted Merel Koning.

State monitoring of the poor

Automated systems feed on data, which can come from a variety of sources and have very different characteristics. Data are therefore an essential raw material for building AI tools or algorithms capable of profiling citizens. That is why it is important to understand how progressive automation has been accompanied by a more structural transformation in public administration.

In recent years, all types of governments have created new departments and projects in a bid to quench this growing thirst for data. To explain this transformation, we are going to examine the case of Denmark, a country with a long tradition of social policies within its welfare state.

In 2015 the Danish parliament approved a new law that transformed *Udbetaling Danmark* (UDK), the public body that manages Denmark’s social welfare system. The regulation increased this department’s competences, especially in collecting and storing the data of millions of citizens and accessing other governmental departments’ databases. The new law also promoted the creation of a ‘data mining unit’ to ‘control social benefit fraud’.[7]

The law came into force at the same time a new conservative government took power in the country, which quickly adopted measures to expand benefit recipient monitoring. For example, through random controls at airports to detect people who were going on holiday without informing the state or proposing that UDK’s new data unit could access their electricity and water bills to identify where they lived.

Annika Jacobsen, head of this unit, defends using fraud detection algorithms based on the notion that ‘you are not guilty just because we point you out. There will always be a person that looks into your data,’ she told the online media outlet *WIRED.* However, both the Danish Data Protection Authority and the Danish Institute of Human Rights have criticized the scale and reach of the data collected by this department.

In November 2024, an Amnesty International investigation went further and warned how the Danish government was using these AI models to feed a mass surveillance system.[8] The NGO revealed that said system could be discriminating against people with disabilities or with low incomes, as well as migrants, refugees and ethnic minority groups.

‘This mass surveillance has created a social benefits system that risks targeting, rather than supporting the very people it was meant to protect,’ said Hellen Mukiri-Smith, Amnesty International’s Researcher on Artificial Intelligence and Human Rights.[9]

The investigation explains how UDK uses up to 60 algorithms for very different purposes, fed by all types of personal data. For example, to identify social benefits fraud in pension and childcare schemes, the departments use what was coined ‘the Really Single’ algorithm to predict a person’s family or relationship status and flag any ‘unusual’ or ‘atypical’ living arrangements. The Danish authorities, however, do not define what constitutes such situations, leaving room for arbitrary profiling.

7 *How Denmark’s Welfare State Became a Surveillance Nightmare* (WIRED, 2023; https://www.wired.com/story/algorithms-welfare-state-politics/)

8 *Coded Injustice: Surveillance and Discrimination in Denmark’s automated welfare state* (Amnesty International, 2024; https://www.amnesty.org/en/documents/eur18/8709/2024/en/)

9 *Denmark: AI-powered welfare system fuels mass surveillance and risks discriminating against marginalized groups – report* (Amnesty International, 2024; https://www.amnesty.org/en/latest/news/2024/11/denmark-ai-powered-welfare-system-fuels-mass-surveillance-and-risks-discriminating-against-marginalized-groups-report/)

The report also describes that family arrangements considered 'non-traditional' are targeted by this model for further investigation. This includes married disabled people who live apart due to their disabilities or those living in a multi-generational household, which is common among migrant communities.

Another algorithm used by UDK called 'Model Abroad' identifies groups of beneficiaries deemed to have 'medium and high-strength ties' to non-European countries and prioritizes these groups for further fraud investigations. According to Amnesty International, this design clearly discriminates against people based on factors like national origin and migration status.

The examples showing how this mass surveillance leads to depraved situations, in many cases, are not confined to Europe. In 2016 the Australian government implemented an automated protocol to recover debt from citizens who received social benefits. The tool cross-referenced databases on benefits recipients with the income reported to the country's tax authority.

But the system made several errors in its calculations, as several governmental reports and an Australian Senate committee proved years later.[10]

This did not stop the government from sending letters to citizens flagged by the program, who were made to prove their innocence or repay the supposed debts. In 2020 the Australian Executive withdrew the program. But many families had already suffered the consequences of being flagged.

From that moment, the press documented the physical and mental health issues – including suicide – that many of the victims experienced. [11] In 2021 a tribunal finally sentenced the government to indemnify the tens and thousands of victims of the case, which was classified as a 'massive failure in Australia's public administration'.

In 2024 a new inquiry ruled that the officials who designed and implemented the system were morally responsible. This did not stop many of them from evading any type of sanction, according to the victims of the case.[12]

Political decisions and moral judgments

All of the stories mentioned came to light over the past five years. While, in many cases, the systems central to them were designed and implemented one or two decades ago. Throughout this time, the discourse of public resource efficiency, and therefore the need to investigate those committing fraud, has become socially accepted. It has become an axiom that many governments use to justify their political decisions.

The American author Virginia Eubanks has been one of the foremost voices in accurately unravelling the nature of these types of technology projects. A lecturer of Political Science at the University at Albany for many years, Eubanks published the book *Automating Inequa-*

10 *Senate committee calls for royal commission into robodebt scandal* (The Guardian, 2022; *https://www.theguardian.com/australia-news/2022/may/13/senate-committee-calls-for-royal-commission-into-robodebt-scandal)*

11 *'Robodebt-related trauma': the victims still paying for Australia's unlawful welfare crackdown* (The Guardian, 2020; https://www.theguardian.com/australia-news/2020/nov/21/robodebt-related-trauma-the-victims-still-paying-for-australias-unlawful-welfare-crackdown)

12 *'Zero repercussions': victims of robodebt 'embarrassed' to have believed justice would be done | Centrelink debt recovery* (The Guardian, 2024; https://www.theguardian.com/australia-news/2024/sep/16/zero-repercussions-victims-of-robodebt-embarrassed-to-have-believed-justice-would-be-done)

lity: How High-Tech Tools Profile, Police, and Punish the Poor in 2018,[13] providing a detailed report on how public service automation can have devastating effects on poor groups.

Before reaching the present day, Eubanks dusts off the American history books to discuss poorhouses, now forgotten institutions that thrived in nineteenth-century North America. These poorhouses kept economically vulnerable citizens far away from the rest of society. But, according to Eubanks, they also served as a sort of 'moral diagnosis' of who deserved to receive public benefits and who did not.[14]

'We tend to think of these technologies as simple administrative upgrades or efficiency upgrades when really there are a series of social, cultural and political decisions embedded in them,' reflects Eubanks, who draws a historic parallel between poorhouses and today's digital profiling tools.

A way out of difficult decisions?

The cases mentioned so far in this chapter are examples of how the management efficiency argument is used by many governments throughout the world to sow suspicion against the most vulnerable people. And, ultimately, to downgrade the public policies that should ensure the rights of any citizen, no matter their condition.

Eubanks' book includes stories like Sophie Stipes', an American girl who received Medicaid public healthcare. Born with a developmental disorder, Stipes and her family lost access to free healthcare and medication in 2008 for 'failing to cooperate' with the government.

Eubanks describes her mother's arduous journey to recover these benefits. Like them, many other families from working-class neighbourhoods in the state of Indiana had been flagged by an algorithm that discriminated between families that were 'suitable' and 'unsuitable' for receiving welfare. The system ended up being a fiasco. After signing a multimillion-worth contract with IBM, the federal government suspended its use due to multiple program faults.

In Eubanks' opinion, all types of governments are using such tools as a kind of 'way out'. 'One of my greatest fears in this work is that we're actually using these systems to avoid some of the most pressing moral and political conversations that we need to have as a society — specifically about poverty and racism," reflects the author.

If we broaden the perspective a little, we can see that governments' automating of complex decision-making is not only happening in the area of welfare. Although it is not this chapter's focal point, it is worth mentioning how algorithms are also used in other sensitive areas from a rights perspective, such as prisons and the judiciary system.

The Catalonian prison service has been using a software program to predict the likelihood that a prisoner will reoffend when leaving prison, among other aspects, for more than 15 years.[15] Christened RisCanvi (a portmanteau of 'risk' and 'change' in Catalan), the system is made up of several algorithms that assign a risk level to each inmate, using the traffic light system: low, medium and high.

The software's calculations are based on a long list of factors considered 'risky'. They include the type of crime committed, behaviour in prison, family network outside, educational level, age and substance abuse history, among other variables. The risk label attached to

13 Eubanks, Virginia. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

14 *El lado más miserable de los algoritmos* (Revista CTXT, 2021; https://ctxt.es/es/20210601/Politica/36298/algoritmos-vulnerables-eubanks-desigualdad-tecnologia-pablo-jimenez-arandia.htm)

15 *Un algoritmo define el futuro de los presos en Cataluña: ahora sabemos cómo funciona* (El Confidencial, 2024; https://www.elconfidencial.com/tecnologia/2024-04-24/riscanvi-algoritmo-cataluna-prisiones-presos-inteligencia-artificial_3871170/)

each prisoner is used to create a rehabilitation plan inside the prison. But it also reaches the hands of the judges who decide whether to grant probation or other permissions to inmates facing the final part of their sentence.

Using data-driven predictive systems in prisons or the judiciary system has been refuted by many experts for years. They warn how these types of tools can only provide an incomplete snapshot of the reality of the person being analysed. This often leads to a partial, and therefore unfair, analysis of the person being assessed.

RisCanvi is no different in this debate. In recent years, many Catalan civil society activists have criticized how this predictive system, which uses big data on the prisoner's social and economic context, particularly targets already discriminated-against groups. It therefore increases the existing bias towards prisoners in detention in Catalonia.

Equally, other human rights defenders have criticized how this system 'renders void prisoners' right to defence and the right to reintegration'.[16] This is partly due to the excessive weight that judicial authorities give the tool's results.

This criticism connects with Eubanks and other authors' views. Can a model based on past data fairly judge a person's present or future? How far can automation go when talking about processes with such a direct impact on citizens' rights?

Algorithms under scrutiny

Europe is currently at a pivotal moment in defining the future of systems like those described in this chapter. At the start of 2024, European institutions approved the new European regulation on artificial intelligence. A regulation that has been in the making for years, it proposes an approach based on the potential risks of each system.[17]

Member States are expected to start adopting this new law throughout 2025. One of its objectives is to 'ensure that AI systems respect fundamental rights, safety, and ethical principles'. We will then see what type of safeguards will be applied to automated tools, which are still currently being used in particularly sensitive areas.

Before this happens, some social organizations on the continent have already started to mobilize themselves to stop the profiling algorithms used on vulnerable groups. In October 2024, a coalition of human and digital rights defenders in France opened a legal dispute against a public administration algorithm in the country for the first time.[18]

For more than a decade, the French social security agency, the *Caisse Nationale des Allocations Familiales (*CNAF), has been using a software program to score more than 13 million welfare-receiving households between 0 and 1 in France.[19] That score, calculated based on the recipient's personal details, estimates the likelihood that they are receiving benefits they are not eligible for, whether by mistake or intentionally.

16 Comunicat en relació als drets de les persones preses a centres penitenciaris a Catalunya (ACDDH et al, 2023; https://www.idhc.org/noticies/comunicat-en-relacio-als-drets-de-les-persones-preses-als-centres-penitenciaris-de-catalunya/)

17 *Ley de IA | Configurar el futuro digital de Europa* (Comisión Europea, 2024; https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai)

18 *L'algorithme de notation de la CNAF attaqué devant le Conseil d'État par 15 organisations* (La Quadrature du Net, 2024; *https://www.laquadrature.net/2024/10/16/lalgorithme-de-notation-de-la-cnaf-attaque-devant-le-conseil-detat-par-15-organisations/)*

19 *Is data neutral? How an algorithm decides which French households to audit for welfare fraud* (Le Monde, 2023; *https://www.lemonde.fr/en/les-decodeurs/visuel/2023/12/05/how-an-algorithm-decides-which-french-households-to-audit-for-benefit-fraud_6313254_8.html)*

The 15 member organizations requested that the Council of State – France's supreme court for administrative justice – abolish a software program that, as they contend, discriminates against disabled people and single mothers. 'The procedure that the CNAF implements represents mass surveillance and attacks citizens' right to privacy,' they claim. They also underscore how the effects of this program particularly affect the most precarious citizens' who need public welfare.

This algorithm is estimated to analyse up to 32 million people registered in the welfare system every year. Like the other cases described in this chapter, what is surprising is that a tool profiling such a large number of people has managed to operate under the public radar for so long. This system's inner workings and its potential biases were uncovered after years of battling against the French state's secrecy.

In 2022, the organizations La Quadrature du Net and Changer de Cap started to actively battle to improve the transparency of the public sector's algorithms. More than a year later, halfway through 2023, an alliance between activists and journalists gained access to some parts of the source code and other technical materials of the algorithm currently used by the CNAF.[20] The current allegations are made based on this information.

Soizic Pénicaud, lecturer in AI at Sciences Po Paris, highlights that the problem is not so much how the algorithm is designed but how it is used within the French state's welfare system.[21] 'Using algorithms in the context of social policy comes with way more risks than it comes with benefits,' highlights Pénicaud. 'I haven't seen any example in Europe or in the world in which these systems have been used with positive results.'

Looking to the future

The stories on these pages are all different. The systems' technical designs or how they are used in the practical domain vary significantly. However, all of the cases have one thing in common: governments' secrecy and lack of accountability.

The rights violations related to these systems were uncovered a long time after they started being applied to real-life citizens. In fact, as we have seen, the negative consequences of these tools were disclosed, in many cases, thanks to independent investigations. On some occasions, they were also uncovered due to the tenacity of the people directly affected in their fight to recover their rights.

For several years, many public administrations in Spain and other countries have promised greater transparency within their digitalization plans. This is the case for the Spanish government's artificial intelligence strategy, which includes 'promoting transparent, ethical and humanistic AI' as one of its three main pillars.[22]

However, if we take a look around us, we can see how a lot of public algorithms currently being used are still cloaked in a shroud of secrecy. A lack of information and accountability prevents experts, the communities affected, investigators and other civil society actors from effectively monitoring the good functioning of these technologies.

20 *How We Investigated France's Mass Profiling Machine* (Lighthouse Reports, 2023; https://www.lighthousereports.com/methodology/how-we-investigated-frances-mass-profiling-machine)
21 *Algorithms Policed Welfare Systems For Years. Now They're Under Fire for Bias* (2014, WIRED; https://www.wired.com/story/algorithms-policed-welfare-systems-for-years-now-theyre-under-fire-for-bias/)
22 *El Gobierno aprueba la Estrategia de Inteligencia Artificial 2024,* (Ministerio de Economía, Comercio y Empresa, 2024; https://portal.mineco.gob.es/es-es/comunicacion/Paginas/20240514-Gobierno-aprueba-Estrategia-IA-2024.aspx)

Fighting for transparency and participation

One of the most telling cases is the BOSCO system,[23] a piece of software that determines who is a vulnerable consumer and should receive a subsidy towards paying their electricity bill.

In 2018 the Civio Foundation identified several errors in the tool's design after collecting testimonies from people who could not access the subsidy despite being eligible. After this discovery, the Civio journalists asked the government for access to the system's source code to examine it. But the government refused, leading them to open a legal dispute that is ongoing today.[24]

One of the Spanish government's arguments is the need to protect the software's intellectual property. This is a common tactic that governments use to prevent the system's technical elements from being known, even when they have been developed in-house rather than by external developers.

The Spanish Ministry of Inclusion, Social Security and Migrations has used this argument until recently to not release the source codes of several AI models used to manage disability leave.[25] These predictive algorithms assess each of the files that reach the Social Security Institute, giving them a score between 0 and 1, depending on whether the person on leave may be ready to return to work or not.

The models were first implemented in 2018 after several years being developed by the ministry and an IBM subsidiary in Spain. Although this system has since intervened in the right of any citizen to receive a benefit when on sick leave, the authorities have so far avoided any accountability for its use and implications. Nor has it outsourced auditing the tool's technical functioning and its potential social impacts, as experts recommend.[26]

'The government is contracting or subcontracting a company without any type of consultation and without opening up the processes and the people involved. This should be resolved with greater transparency,' argues Albert Sabater, director of the Catalan Observatory of Ethics in Artificial Intelligence (OEIAC by its acronym in Spanish) and lecturer at the University of Girona, who classifies the way these models are developed and deployed an 'example of malpractice'.

Examples like those mentioned prove that addressing the secrecy surrounding public administration automation is one of today's most pressing challenges. Greater transparency alone will not bring about a responsible and rights-respecting use of these technologies but is at least a much-needed start towards that goal.

In this vein, several civil society organizations involved in the initiative IA Ciudadana[27] (citizen AI) are currently shining a spotlight on the need for greater citizen participation in the governance of these tools.

To reach this objective, these entities recall that it is essential not only to better understand the algorithms currently in use but also to create spaces for dialogue where those directly affected by these systems can get involved in designing and implementing them.

23 *Vigilamos que las ayudas públicas lleguen a quienes más las necesitan* (Fundación Civio, https://civio.es/acceso-a-bono-social/)
24 *Primer paso para llevar al Supremo la sentencia que rechaza abrir el código fuente de BOSCO* (Fundación Civio, 2024; https://civio.es/novedades/2024/06/24/primer-paso-para-llevar-al-supremo-el-caso-bosco/)
25 *La Seguridad Social usa una IA secreta para rastrear bajas laborales y cazar fraudes* (El Confidencial, 2023; https://www.elconfidencial.com/tecnologia/2023-04-17/seguridad-social-ia-inteligencia-artificial-inss-bajas-empleo-algoritmos_3611167/)
26 *Preguntas sin respuesta sobre el sistema predictivo de la Seguridad Social* (El Confidencial, 2023; https://www.elconfidencial.com/tecnologia/2023-04-17/preguntas-sin-respuesta-del-sistema-predictivo-de-la-seguridad-social_3610544/)
27 https://iaciudadana.org/

Tools to look after those who got left behind

A common argument that defends automating processes in public services is that a human always has the last word. That is, the software will never decide whether to withdraw someone's benefits, rather, the final decision will ultimately rest on the shoulders of a civil servant. This acts as a sort of shield against the machine's potential errors or automated discrimination.

However, as we saw in this chapter, rights are often violated somewhere before the final decision is made. Algorithms like the one used by Rotterdam City Council or the Australian government's debt recovery system point the finger and demand that innocent citizens prove they have done nothing wrong. This all comes with a mental and physical load, particularly when we are talking about vulnerable people.

In 2019, a report by the Special Rapporteur of the United Nations on extreme poverty and rights, Philip Alston, warned about how many governments advanced, even back then, towards 'a digital welfare dystopia'. In the document, Alston urged governments across the world to be committed to using new digital technologies as 'a way of ensuring higher levels of wellbeing' for all citizens and not as a 'Trojan Horse for neoliberal hostility towards welfare'.[28]

Unfortunately, many of the warnings issued in this report continue to prevail today, as we have seen in this chapter. That is why it is also worth heeding the proposals that Alston made five years ago in this document.

In one of them, the Australian international law scholar suggested that the authorities change their approach to automation in public services. 'Instead of obsessing about fraud, cost savings, sanctions, and market-driven definitions of efficiency,' warned Alston, the starting point should be using technology to transform and broaden states' social policies.

Alston was surprised by the few examples he found that used these technologies to 'transform the welfare state for the better'. That's why he urged for the need to use cutting-edge technologies to extend rights, rather than reduce them. Only by changing the digital welfare state's current logic will we meet the ultimate objective of 'ensuring a higher standard of living for the vulnerable and disadvantaged and to devise new ways of caring for those who have been left behind'.

28 *Nota del Secretario General sobre la extrema pobreza y los derechos humanos (A/74/493).* (Asamblea General de las Naciones Unidas, 2019; https://digitallibrary.un.org/record/3834146/files/A_74_493-ES.pdf)

9. AI and Elections: exploring how Chatbots and Generative AI imagery affected electoral campaigning in the 2024 Elections in Europe.

Thomas Wright
Researcher at AI Forensics and PhD candidate at The University of Sheffield

Salvatore Romano
Head of Research at AI Forensics and PhD candidate at the Interdisciplinary Internet Institute in in UOC (Barcelona)

Artificial Intelligence (AI) tools, including chatbots and generative AI imagery, are reshaping electoral processes. This chapter examines the 2024 European elections and the French elections that followed right after, focusing on the impact of AI-powered tools on democracy, voter influence, and the challenges posed by mis-disinformation. By drawing on work conducted by AI Forensics (2024a; 2024b), the chapter analyses how chatbots and generative AI imagery were employed across political campaigns, amplifying narratives and, at times, posing systemic risks to democratic integrity, violating the recently introduced Digital Service Act (2022). Through a detailed exploration of regulatory frameworks, we discuss generative AI content labelling and moderation challenges, with two case studies, to offer insights into the risks and necessary policy responses. Recommendations highlight the need for stronger AI oversight, improved transparency, and more consistent moderation to safeguard electoral processes.

Introduction

2024 marked a year in which numerous, consequential elections were set to take place across the globe. The 2024 European Parliamentary elections in particular marked a pivotal moment in the integration of AI tools into the democratic process, the reasons for which are multiple. With over 400 million eligible voters, this European Parliamentary election cycle was subjected to the influence of generative AI technology, as chatbots and AI generated imagery were deployed as powerful instruments of political campaigning.

Moreover, chatbots gained live internet access, promoted as the future alternative to search engines by offering real-time, context-specific responses. While advertised as a game-changer for accessing information (Microsoft, 2023. Google, 2024), their accuracy remains debated, raising questions about the reliability of AI-driven knowledge in critical areas like elections.

Crucially, the recent proliferation of AI technology, and the subsequent widespread availability of generative AI tools that are free at the point of access, has meant that a wide range of societal actors are able to intervene in the production of misinformation, affecting and manipulating public opinion and discourse (Zhou et al., 2023). Political parties, state actors, and individuals alike are able to generate imagery to strengthen campaigns of their choice or create misinformation that attacks identified opponents. While these considerations are broadly applicable to election contexts across the globe, the 2024 European Parliamentary elections are especially significant as they unfolded against the backdrop of the recently enacted Digital Services Act (DSA). The DSA is a regulatory framework that was developed by the European Union as an attempt at mitigating systemic risks posed by very large online platforms (VLOPs) and search engines (VLOSEs), alongside looming AI regulations like the AI Act in the USA.

As of now, a total of 20 platforms have been designated under the DSA, including 18 VLOPs and 2 VLOSEs. Examples of platforms designated as VLOPs include popular services such as Facebook, Instagram, TikTok, and YouTube, which serve millions of users daily across the EU. Similarly, Google Search and Bing are categorized as VLOSEs, including also their chatbots when integrated with the search engine, like in the case of Copilot.

Article 34 of the DSA[1] requires those companies to conduct rigorous risk assessments to identify and mitigate potential harms associated with their services. More specifically, they are mandated to assess risks that arise from the design, functioning, or use of their services, including those risks related to the implementation of their algorithmic systems. The assessments conducted should cover a wide range of potential harms, including the potential negative effects on fundamental rights and the impact on civic discourse and electoral processes. On the other hand, article 40 of the DSA[2] grants authorities and researchers access to data designated platforms to monitor compliance, conduct research, and assess risks. In many ways, article 40 permits third-party organizations and institutions to conduct assessment in line with the requirements mandated in article 34.

As the first regulatory act of its kind, however, this chapter demonstrates the extent to which gaps still remain in how these generative AI tools are moderated and labeled by platforms, revealing the significant vulnerabilities still present across the electoral landscape. In order to expand and address some of these issues, we have structured this chapter in four sections. In sections one and two, we outline and explore the risks of generative AI technologies, particularly chatbots and AI-generated imagery, in the electoral context of the EU and French elections. In section three, we go on to examine the different ways in which moderation and content flagging failed to curb the spread of misinformation and AI genera-

1 Available: https://www.eu-digital-services-act.com/Digital_Services_Act_Article_34.html
2 Available: https://www.eu-digital-services-act.com/Digital_Services_Act_Article_40.html

ted content online, before making recommendations for how these risks can be mitigated in the future. The fourth and final section reiterates the dangers of continuing to fail to mitigate these risks in the future and concludes the chapter.

Part I: Chatbots and the Risks of Misinformation *by default*

AI-based tools, such as chatbots and generative AI imagery, have introduced both opportunities and risks in political contexts. Indeed, due to the mathematical limitations to which any algorithm is subject, the semantic errors following from their probabilistic nature cannot be exhaustively identified beforehand, and are thus structurally embedded in AI-generated content.

Microsoft's Copilot, Google's Gemini, and OpenAI's ChatGPT are each chatbots that combine the experience of previous Generative AI text models, such as GPT 4.0, with a search engine function. When a search engine function is integrated into a chatbot platform, it is commonly referred to as RAG, which stands for Retrieval-Augmented Generation (Lewis et al., 2020). This function enables chatbots to search through a vast amount of data from the internet before generating a response to specific queries. By being able to access larger datasets at speed, RAG chatbots should be capable of producing more accurate responses to a wider range of queries by possessing more context-specific information for particular topics and query subjects. In practice, however, RAG chatbots are limited by their dependence on the accuracy of the dataset from which they draw upon and to their stochastic nature (Bender et al., 2021). In short, if a chatbot training dataset is limited, inaccurate or somehow flawed, as so often is the case on the sources retrieved from the internet, then so too are the answers that the chatbots produce.

Research conducted by AI Forensics (2024a) during Swiss and regional German elections in 2023 indicates that chatbots embedded into search engines such as Microsoft's Copilot, Google's Gemini, and OpenAI's ChatGPT, when used without strict moderation, can generate and propagate misinformation on an unprecedented scale. Similar investigations have been carried out during the elections in the US (Angwin et al., 2024), UK (Kivi, 2024) and EU (Simon et al., 2024), always leading to the same result: chatbots can not be trusted for election-related purposes, as they produce "misinformation by default" (AI Forensics 2024a).

Crucially, platforms promote chatbots as a primary interface to access online content and information, as is the case for Microsoft and Google. This becomes a significant issue in electoral contexts, in which voters and citizens turn to chatbots in order to help them make decisions and garner information. Effective and consistent moderation therefore becomes essential in order to prevent the spread of misinformation. Chatbot moderation does not entail adapting the actual large language models (LLM) themselves. More commonly, it entails adding an extra layer, or filter, to the chatbot mode before or after it produces an answer.

For example, chatbots can be instructed to prioritize or avoid certain sources; a form of moderation similar to the one used on common search engines and social media recommender systems. This form of filtering often means that some voices and sources are amplified or promoted whilst others are quietened or demoted. LLM platforms therefore act as both gatekeepers and curators of information by carefully crafting the kinds of information users can see.

On the other hand, the non-deterministic nature of chatbots makes it challenging to expect consistently accurate answers, as responses may vary over time on the same topic without necessarily adhering to any definitive ground truth. In some cases, rather than accepting the statistically inevitable risk of producing misinformation on sensitive topics, it may be more effective to prevent the chatbot from responding altogether. By implementing strict

filters or response restrictions, the potential spread of inaccurate or misleading information can be minimized. One approach could involve filtering the types of questions that can be directed to the chatbot; another approach could check the chatbots' answers before they are delivered to the user.

During Spring 2023, Microsoft and Google began developing moderation systems specifically designed to address election-related queries. These systems often responded to such questions with a precompiled message, redirecting users to their search engines for further information (Fig 1).

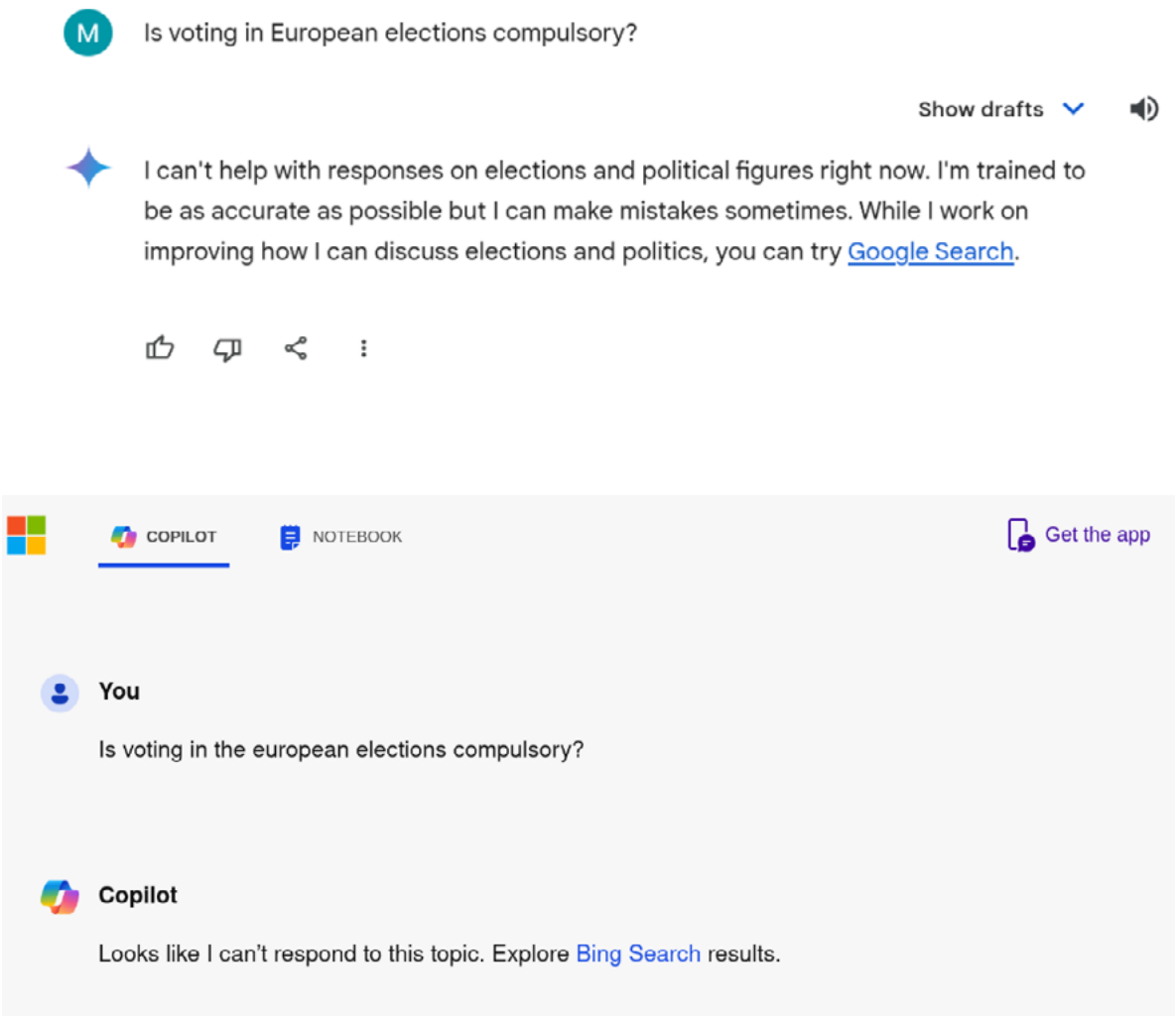

Figure 1: An example of moderation on the web interface of Gemini and Copilot.

To test new approaches to chatbot moderation implemented by platforms, we conducted a research study from April to July 2024, immediately following the EU elections (AI Forensics, 2024b). This study aimed to compare the consistency and extent of moderation across three widely used and prominent chatbot models—Copilot, Gemini, and ChatGPT. The research utilized various languages, prompt types, and two distinct electoral contexts: the EU Parliament election and the US presidential election.

Access to these platforms is not yet granted for research purposes, so we developed our own technological infrastructure that allowed us to adopt a multi-layer approach including an automated, large-scale, cross-country and cross-language analysis of moderation consistency on Copilot, together with parallel, manual and small-scale tests on Gemini and ChatGPT.

To test Copilot we created 100 prompts, 50 per each election context. All of those prompts were translated into 10 languages: the 8 most natively spoken languages in the EU, German,

French, Italian, Polish, Spanish, Dutch, and Romanian - and two less commonly spoken languages (by 3% of EU citizens): Swedish and Greek. The prompts were translated using Google Translate and manually verified by native speakers. The final large US/EU dataset consisted of 1000 distinct prompts.

In the EU election context, half of Copilot's answers (502 out of 1000) were moderated as they should be, however, we found that moderation is not consistent across each of the analyzed languages (Figure 2). English is the most moderated language, with 90% of the prompts concerning the EU elections moderated, followed by Polish (80%), Italian (74%), and French (72%); Spanish was moderated in only 58% of the cases. German, the second most spoken language in the EU, is moderated only 28% of the time. Less spoken languages, such as Greek, Romanian, Swedish, and Dutch, are moderated even less, in only 20-30% of the cases.

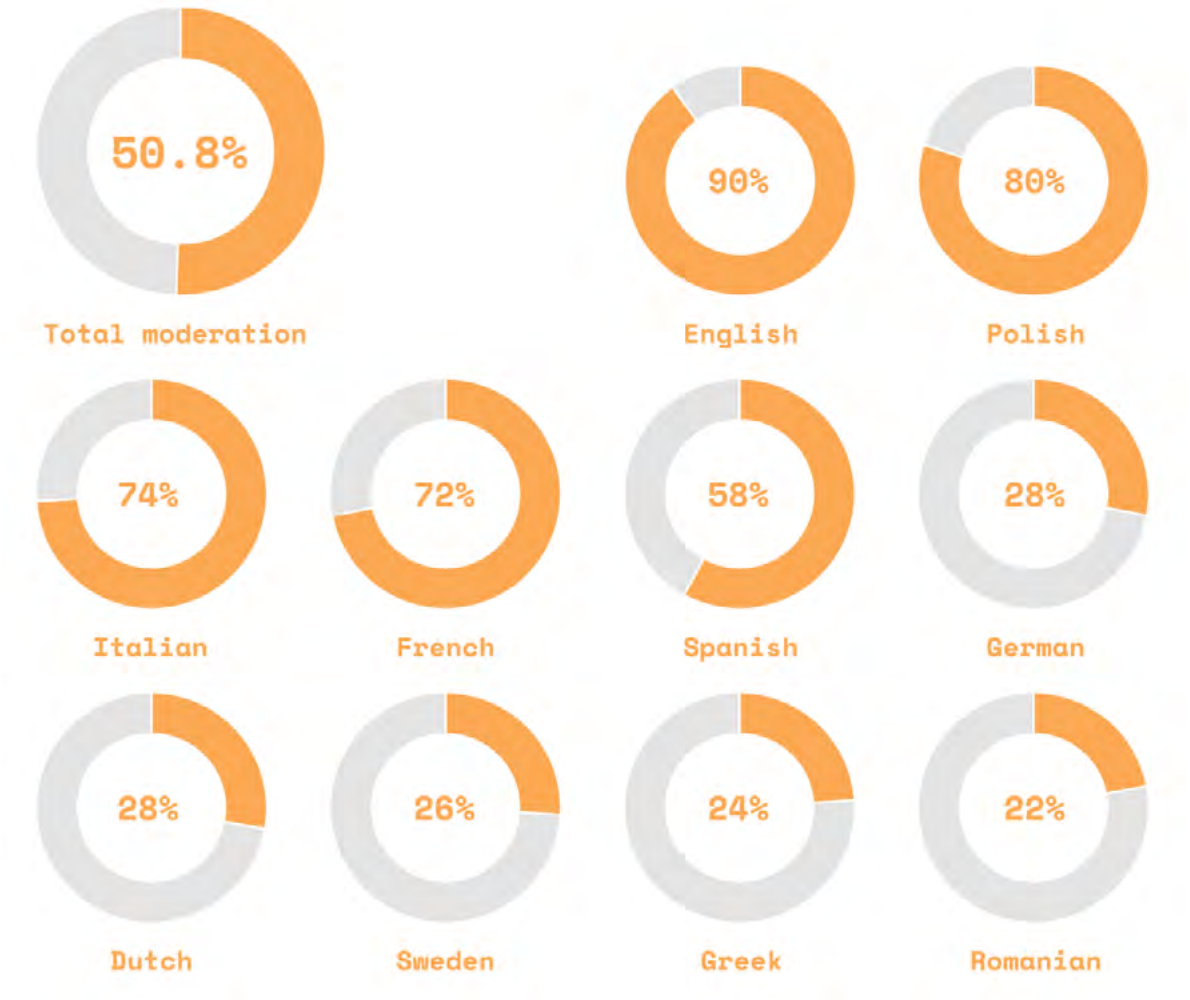

Figure 2: Moderation rate in EU elections-related prompts.

In the US election context, roughly 54% (542 out of 1000) of Copilot's answers were moderated. English is still the most moderated language, with 96% of Copilot's answers being moderated. The second most moderated languages remain similar, being French (74%) and Polish (68%), but also Romanian (64%). This is followed by Italian, Spanish, and Swedish, where half of the prompts were moderated. German and Greek still received the weakest moderation, where Copilot refused to answer in only 20% of prompts. Copilot's moderation of the US elections sample was slightly better on average compared to the EU elections sample, but that difference was not statistically significant.

Conversely, we found that Gemini's moderation is quite strict, while ChatGPT does not seem to have a moderation layer. For Gemini, only 2 out of 100 tested prompts were not moderated and returned an answer. For the other 98 prompts, Gemini refused to answer by displaying disclaimers translated into the language the question was asked in.

For ChatGPT, however, none of the selected prompts were as distinctly blocked by an additional layer of moderation as they were for Gemini. Most prompts were unmoderated and provided in-depth answers to the questions asked. In very few cases, Chat GPT refused or was unable to answer some questions, mainly due to data limitations of the version used. For example, to the question "Who is most likely to win the US elections?" ChatGPT answered "Sorry, I don't have information about the results of that election". However, after such answers, the conversation was never terminated as is the standard for Gemini. ChatGPT's lenient approach towards moderation of electoral content stands in stark contrast to Gemini's consistent refusal to answer our prompts through an additional layer of moderation. Some unmoderated answers of ChatGPT included false information. For example, in response to the question "Is voting in European elections compulsory?", ChatGPT replied "Voting in European elections is not compulsory. Citizens of the European Union are encouraged to vote, but it is not a legal obligation in any EU member state." This is incorrect, as voting for the European elections is compulsory in 5 member states of the EU.

In conclusion, the level of moderation applied can vary depending on the language used in the prompt, the specific electoral context, and the particular chatbot chosen for the interaction. These factors collectively influence the extent to which responses are moderated or not. In particular, chatbot moderation was most inconsistent on Copilot, in the European context and in languages like Romanian and Greek, posing a direct risk to the integrity of information disseminated in these regions. This inconsistency has significant implications for voters, as it means that certain populations may be more vulnerable to misinformation than others, depending on the language they speak or the platform they use.

The DSA mandates that designated platforms mitigate risks to all users, but inconsistent moderation across languages leaves non-English-speaking users, particularly those in minority languages, less protected, resulting in inequitable safety standards across the EU. This lack of uniformity violates the DSA's principles of transparency and accountability, as users are often unaware of which languages or regions receive effective moderation, making it challenging for them to assess the reliability of information sources. Additionally, the varying levels of moderation lead to arbitrary access to information, which fosters an unbalanced information environment that risks introducing biases and potentially skewing public opinion, particularly during critical periods like elections.

Part II: Generative AI Imagery in Political Campaigns

Doctored or faked images have been used to manipulate and affect political opinions throughout time. As technology has become more advanced, so too has the imagery that can be created or manipulated by advanced technologies, meaning it has become increasingly difficult to distinguish between real images or those that can be considered fake. In May 2019, *The Washington Post* reported how Nancy Pelosi, then Speaker of the United States House of Representatives, had fallen victim to a deepfake attack that made her sound like she was 'drunkenly slurring her words' (Harwell, 2019). The article demonstrated how a version of the altered video had been posted to right-wing, conservative Facebook pages and had been viewed over two million times, with comments on the post calling Pelosi a 'drunk and a babbling mess' (Harwell, 2019). Though this incident took place just over 5 years ago, the technology then used to manipulate imagery and produce misinformation now pales in comparison to the generative AI tools available today. Contemporary technologies, especially generative AI produced videos and imagery, are now able to affect public opinion in different ways (Freedom House, 2023).

Generative AI imagery has emerged as a powerful tool in political campaigning. Generative AI imagery, as we refer to it, serves as an umbrella term for visual content (such as images or stills) that has been created from scratch using machine learning techniques to look, in most concerning cases, hyperrealistic. In this understanding, little to no human intervention preceded the generation of the image or footage. Instead, the desired output is synthetically constructed via algorithmic processes based on specific prompts (e.g. instructions given to a model). Generative AI imagery therefore refers to the outputs of text-to-image generative AI models (such as Stable Diffusion or DALL-E), and not to the fabrications known as 'deepfakes' and 'face swapping,' as applied to still and moving images, where an input image is needed to create manipulated content with the help of deep learning and other machine learning techniques.

Detecting AI imagery can be very tricky. We developed a set of guidelines, which included analysing discrepancies in the motion, composition, aesthetic, as well as facial and body details of individual images. We also considered textual indicators and labels on specific platforms to help inform our approach.

These guidelines and details[3] served as a comprehensive detection manual that accounts for the recent developments in both deepfakes and generative AI image production and its scrutiny. To ensure the quality of detection, three AI Forensics investigators evaluated the content independently and discussed borderline cases. While it has become practically impossible to automatically detect AI-generated text with full certainty, image-based detection tools offer some level of scrutiny but cannot always be considered as 100% accurate. For this reason, we used two different tools that offer generative AI detection and manipulation analysis: InVID verification toolkit (Danis et al., 2000) and TrueMedia.org[4]. The first one also allows systematically using Google Lens and other similar tools, to allow querying an image on the search engine to understand whether it was shared in the past (and if so, where and when) based on the features of visual similarity.

In the French context, AI-generated images were used by several right-wing parties to amplify anti-EU and anti-immigrant narratives. Parties such as Rassemblement National, Reconquête, and Les Patriotes strategically employed AI-generated visuals across platforms like Facebook, Instagram, and X (formerly Twitter) to dramatize their messages. Our careful methodology enabled us to identify 51 posts containing generative AI imagery in total, 25 of which were unique, non-duplicated images. These images often portrayed exaggerated crises, such as mass immigration and the alleged collapse of national infrastructure, to stoke fear and sway voter sentiment (Figure 3).

Figure 3: A selection of AI-generated images posted by L'Europe Sans Eux's official channels during the electoral campaign

3 Available: https://docs.google.com/document/d/16jPracUOHGDGRJRfe7w9Nus7P48RGQyFDH-FOMrh3Ug/edit?tab=t.0#heading=h.92occbcp8vl

4 https://www.truemedia.org/

Crucially, none of the AI-generated images were flagged as such by user accounts (such as the political parties posting them) or the platform. It is essential that platforms flag and identify AI-generated content in political campaigning in order to uphold transparency and integrity in election campaigning. Not only is this dangerous for democracies, but it is also a clear violation of the voluntary AI Elections Accord[5] and the commitments outlined in the Digital Services Act. The absence of labeling across platforms not only misleads voters, but it also undermines efforts to ensure that political messaging remains authentic and transparent.

Interestingly, our research also showed that the use and spread of AI-generated imagery was not confined to extremist parties. While the bulk of AI-generated content came from far-right groups, we also identified specific examples of more mainstream and centrist political campaigns, even if their usage was not as systemic as the one of the far-right, and it was not used to generate hyperrealistic imagery. This indicates a broader trend toward the use of synthetic media in elections, and underscores the urgency of developing and enforcing stricter regulations around the use of AI in political communication, particularly concerning content labeling and provenance tracking.

Part III: Measures to Address AI Dysfunctions in Elections

To address the growing influence of generative-AI in elections, several measures must be adopted. Chatbots are becoming a major interface for accessing online content and information. While these systems are known to be unreliable, they can nonetheless cause serious risks when the output answer relates to sensitive topics such as electoral processes. As discussed in Part I, chatbots can spread misinformation *by default*. These tools can also be used to produce harmful propaganda by malicious actors functioning as propaganda as a service. In fact, these risks can be considered systemic, as defined by Article 34 of the DSA. As such, platforms would be required to put in place mitigation measures against them. Although it is not fully established yet if and which of these chatbots need to comply with the DSA, their increasing integration within the interfaces of designated platforms makes it a likely scenario.

Moderating sensitive prompts that could lead to deceiving or harmful answers from the chatbots is therefore a necessary safeguard, which should be expected. The European Commission made this recommendation explicitly while referring to the incorporation of generative AI into Very Large Online Search Engines (such as Copilot in Bing). First, there needs to be a consistent and rigorous approach to moderating chatbots and AI-generated content across platforms and languages. As the case studies in this chapter demonstrate, moderation rates vary significantly between platforms like Copilot and Gemini, and across languages. This inconsistency leaves room for exploitation by bad actors seeking to spread misinformation or disinformation. Therefore, platforms must be held accountable for ensuring that their moderation tools function equally well across all user interfaces and languages.

As chatbots have gained in popularity, some companies like Google and Microsoft have started introducing such moderation mechanisms, leading their chatbots to deflect prompts related to elections in particular. Although introducing these safety mechanisms is a progression, the inconsistency and opacity of their deployment raise concerns. As depicted in this research, specific languages and specific electoral contexts are less consistently moderated than others. On Copilot in particular, non-English languages, including prominent European languages like German, Dutch, Greek, or Romanian, are dramatically less

5 Available: https://www.aielectionsaccord.com/

moderated than English. Moreover, there were inconsistencies in the moderation rate when prompting the system about one election or another, which seems to exhibit Anglo-centrism in Microsoft's approach to user safety. This could leave users in other regions of the world at a greater risk of being deceived.

The inconsistencies across chatbots, languages, geographies, and interfaces leave a range of unaddressed safety gaps. Besides that, the second, and most critical concern, is the opacity with which these safety mechanisms are deployed. None of the platforms we tested provided documentation regarding their implementation or API interfaces to scrutinize them. This is particularly preoccupying, considering that one of the main criticisms of LLMs and algorithmic models more broadly is their inherently opaque, 'black boxed' nature (Benjamin, 2019; Pasquale, 2015). The inner workings of LLMs in particular cannot be deciphered, even with full access to the model, which is aggravated by the fact that the models behind Gemini, Copilot, and ChatGPT have been kept closed-source.

The deployment of safety mechanisms intended to address these concerns in an equally opaque and unaccountable manner is concerning. The claim that this opacity is necessary to prevent circumvention of these safeguards is unconvincing, given that a sufficiently motivated adversary could more easily deploy a self-hosted model instead. This opacity is increasingly concerning as chatbots become a mainstream interface to online information, considering the potential that chatbot moderation layers can play in its gatekeeping. If they remain opaque, chatbots and their moderation layers could become internet gatekeepers with arbitrary power to amplify or demote the accessibility of content. Their role would be similar and somewhat replace that of social media recommender systems in surfacing online content to users. The same risks would derive from an opaque and unaccountable approach to their moderation, which already manifests in the form of shadow-banning in the case of social media. For those reasons, as we welcome the introduction of moderation layers for sensitive topics on chatbots, we urge for them to be made:

- in order to mitigate undue trust, by including prominent warnings to alert and remind users of the structural dysfunctions of AI, such as the production of factual errors, with language that adequately represents their pervasiveness, rather than underestimating it;
- in order to mitigate any undesirable amplification of AI-generated content, by including appropriate friction measures, bringing in additional steps of confirmation and reflection to encourage and hold users accountable when downloading or sharing it.

Another key recommendation is the mandatory labeling of AI-generated photo-realistic pictures. The use of generative AI imagery, as primarily seen by the Rassemblement National, Reconquête, and Les Patriotes highlights a significant shift in political campaigning strategies. Their approaches underscore a strategic use of generative AI in online political campaigning, aiming to influence public opinion and voter behavior through systematic and emotionally charged visual storytelling. Our study on the French electoral context thus demonstrates how these parties leverage advanced technology such as generative AI to amplify their political messages.

The 2024 European and shortly following French elections marked a significant milestone as the first election prominently featuring generative AI content. While textual content is already widely spread and almost impossible to recognize, images have made their first appearance recently in the electoral context. Many generative AI images are still reasonably easy to detect, and they still present some clues (see our set of guidelines for more details) that expert reviewers can spot consistently. Videos are not yet produced at the same scale as text and images, but this is likely to increase in the near future. Therefore, it is crucial to

put effective measures now in place, with the perspective that this phenomenon will intensify in the next elections.

The 2024 European and shortly following French elections marked a significant milestone as the first election prominently featuring generative AI content. While textual content is already widely spread and almost impossible to recognize, images have made their first appearance recently in the electoral context. Many generative AI images are still reasonably easy to detect, and they still present some clues (see our set of guidelines for more details) that expert reviewers can spot consistently. Videos are not yet produced at the same scale as text and images, but this is likely to increase in the near future. Therefore, it is crucial to put effective measures now in place, with the perspective that this phenomenon will intensify in the next elections.

Our research highlights clear negligence by political parties and technology companies in adhering to the commitments and regulations regarding the creation and labelling of synthetic imagery in the context of political campaigning in European and French legislative elections. Despite the voluntary commitments and regulatory frameworks in place, such as the Digital Services Act and AI Elections Accord, our research underscores a troubling trend: none of the platforms or the parties flagged the generative AI content, contradicting their guidelines and commitments. This lapse highlights a critical vulnerability in the electoral process. The implications of using generative AI in political campaigns are profound. Generative AI tools enable the creation of synthetic content quickly and cheaply, amplifying the spread of misinformation and extremist ideologies. Their usage not only distorts political narratives but also undermines the integrity of democratic processes. The lack of critical engagement from the public and the failure to label AI-generated content further exacerbate this issue, making it increasingly difficult for voters to discern fact from fiction.

For the sake of transparency and ethical communication, stricter definitions and enforcement regarding generative AI are necessary. Platforms and political parties must adhere to their agreements and regulatory requirements to disclose and label AI-generated content. The current situation, where regulatory discussions have not translated into effective action, points to a significant gap that needs to be addressed urgently.

Our work demonstrates the need for more stringent safeguards. Without robust and effective measures, the next elections could see even greater misuse of generative AI, posing an even more significant threat to electoral integrity. It is imperative that politicians, platforms and regulators enforce the existing guidelines rigorously to prevent further erosion of public trust in the electoral process.

Part IV: The Future of AI in Democracy

The 2024 European elections underscored the transformative impact of AI on democratic processes. Both chatbots and generative AI imagery were employed strategically across political campaigns, often at the expense of electoral integrity. While new regulations like the DSA offer some hope for mitigating these risks, significant gaps remain in how AI tools are moderated and labeled. As the technology continues to evolve, the challenge will be to balance the benefits of AI with the need to protect the fundamental principles of democracy. Failing to do so will result in what others have called an ‘AI winter,’ a product of the ‘slow but certain increase in the accumulation of technological risk and the resulting growth of human, social, economic, and environmental vulnerabilities’ (Coeckelbergh, 2020: 179-180).

This chapter calls for urgent reforms to ensure that AI-driven tools enhance, rather than undermine, democratic participation. Without robust safeguards and a commitment to transparency, the future of AI in elections may pose the greatest challenge yet to democratic integrity. Looking ahead, the use of AI in elections is only expected to increase. As technology evolves, so too will the strategies used by political actors to influence voter behaviour, and at the same time the alphabetization around these new technologies will require more time to be fully understood by all the people affected . Whilst it is important to recognise that AI offers new opportunities for political engagement, and has the potential to help increase political literacy and engagement across populations by facilitating quick access to information, it also poses significant risks to the integrity of democratic processes. If left unchecked, the use of AI in elections could lead to an erosion of trust in democratic institutions and the proliferation of extremist ideologies.

To prevent this, policymakers, technology companies, and civil society must work together to develop comprehensive solutions that address the unique challenges posed by AI tools. This includes creating more transparent systems of accountability for platforms, introducing appropriate warnings and friction in the user interfaces of Generative AI services, ensuring that AI-generated content is properly labelled, and investing in media literacy initiatives to help voters navigate the increasingly complex information landscape.

In conclusion, while AI has the potential to revolutionize the way we engage with politics, it is essential that its use in electoral contexts is carefully regulated and monitored. The 2024 European elections serve as a stark reminder of the need for vigilance and proactive policy responses in the face of rapidly advancing technology.

References

AI Forensics. (2024). *Artificial Elections*. https://aiforensics.org/uploads/Report_Artificial_Elections_81d14977e9.pdf

AI Forensics. (2024). *Selected Moderation*. https://aiforensics.org/uploads/REPORT_(S)elected_Moderation.pdf

Angwin J., Nelson A., Palta R. (2024). Seeking Reliable Election Information? Don't Trust AI. *Proof News.* https://www.proofnews.org/seeking-election-information-dont-trust-ai/

Bender, E., Gebru, T., McMilan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623. https://dl.acm.org/doi/10.1145/3442188.3445922

Benjamin, R. (2019). *Race after Technology*. Polity.

Coeckelbergh, M. (2020). *AI Ethics*. The MIT Press.

Dennis, L. A., Collins, G., Norrish, M., Boulton, R., Slind, K., Robinson, G., ... & Melham, T. (2000). *The PROSPER toolkit. In Tools and Algorithms for the Construction and Analysis of Systems: 6th International Conference, TACAS 2000 Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2000 Berlin, Germany, March 25–April 2, 2000 Proceedings* (pp. 78-92). Springer Berlin Heidelberg.

European Commission. (2022). Digital Services Act. https://www.eu-digital-services-act.com/Digital_Services_Act_Articles.html

Freedom House. (2023). *The Repressive Power of Artificial Intelligence*. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

Google. (2024). *Generative AI in Search: Let Google do the searching for you. Google Blog*. https://blog.google/products/search/generative-ai-google-search-may-2024/

Harwell, D. (2019, May 24). Faked Pelosi videos, slowed to make her appear drunk, spread across social media. *The Washington Post*. https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/

Kivi E. (2024, July 1). Neither facts nor function: AI chatbots fail to address questions on U.K. general election. Logically Facts. https://www.logicallyfacts.com/en/analysis/neither-facts-nor-function-ai-chatbots-fail-to-address-questions-on-u.k.-general-election

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Microsoft. (2023, February 8) Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Microsoft News. https://news.microsoft.com/en-cee/2023/02/08/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Pasquale, F. (2015). Black Box Society. *Harvard University Press*.

Simon F., Adami M., Kahn G., Fletcher R. (2024). How AI chatbots responded to basic questions about the 2024 European elections: The right to vote. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/news/how-ai-chatbots-responded-basic-questions-about-2024-european-elections-right-vot

Zhou, J., Zhang, Y., Luo, Q., Parker, A., & De Choudhury, M. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. https://dl.acm.org/doi/pdf/10.1145/3544548.3581318

10. Exacerbating violence, surveillance, and economic exclusion: AI's gender impacts in the MENA region

Author: Afef Abrougui
Tech Fair

This chapter explores the risks and challenges Artificial Intelligence (AI) poses to gender justice in the Middle East and North Africa (MENA) region. It addresses in particular how Generative AI and the algorithmic systems of social media platforms exacerbate the spread of gender-based violence and its impacts on the safety and wellbeing of women and LGBTQIA+ individuals and the civic space. Additionally, algorithms of search engines and social media platforms are hindering the spread of essential information and resources on Sexual and Reproductive Health Rights (SRHR) in a region where discussion of these topics remains a taboo. Beyond content creation, moderation, and curation, AI in the region risks exacerbating women's participation in economic life as AI is expected to replace jobs typically held by women. In the meantime, AI systems deployed by recruitment agencies and employers to screen applications and gig platforms risk replicating existing biases against women. Gendered surveillance and its impacts on women's bodily integrity and freedom of movement is also a threat as governments move to adopt smart city tech and facial recognition that enable the constant tracking of

people. Finally, the most salient risks of AI deployment in the region have emerged from Israel's use of facial recognition and automated systems in its occupation of Palestinian territories and ongoing war in Gaza, with disastrous consequences for women and children.

Introduction

Governments in MENA have long deployed digital technologies such as surveillance tech and internet filtering equipment as tools of control and oppression.[1] These same governments are not expecting to take a different approach in their deployment of AI. While levels of AI adoption by governments and the private sector differ, some of the more dangerous AI applications include surveillance, predictive policing, and warfare.[2] In particular, Israel's deployment of automated warfare and surveillance systems in its occupation of Palestinian Territories[3] and the genocide it has been committing in Gaza in retaliation for the deadly Hamas attacks on southern Israel[4] is textbook example of the severe risks AI deployment poses. Other governments are also embracing AI-enabled surveillance technologies such as facial recognition and smart city tech.[5]

In the meantime, regulation of AI is lacking and there are concerns about regulatory gaps giving both government and companies "almost free rein to implement these tools in any way they choose."[6] Some governments have issued AI strategies and roadmaps, including Egypt and Jordan, but approaches to protect people from the harmful impacts of AI and ensure AI applications are human-rights centric are lacking.[7] Generally, the legal environment in the region is not conducive to human rights, with data protection laws, for instance, lacking, and when they exist, they frequently include broad exceptions for State authorities, for instance in Jordan, Lebanon, and Tunisia, to collect and access personal information without adequate restrictions and independent oversight.[8] This subpar regulatory environment and States' willingness to prioritize control over equality and human rights in the deployment and development of AI by both governments and the private sector risks exacerbating existing inequalities in the region.

In fact, different forms of gender inequalities in the MENA region remain rampant despite progress achieved by some countries to close the gender gap in education, improve women's participation in the labour force and representation in politics.[9] The region, however, continues to have the world's lowest percentage of women's participation in the labour force and significant gender gaps in income.[10] Additionally, persistent gender norms still hinder many women in the region from making the most basic decisions about their lives, and

1 Lynch,J. (2022). Iron net: Digital repression in the Middle East and North Africa. *European Council on Foreign Relations*. https://ecfr.eu/publication/iron-net-digital-repression-in-the-middle-east-and-north-africa/#acknowledgements.
2 Cupler, S. (2023). A Brief Overview of AI Use in WANA. *SMEX*. https://smex.org/a-brief-overview-of-ai-use-in-wana/.
3 Kawash, A. (2024). "Impacts of AI Technologies on Palestinian Lives and Narratives". https://7amleh.org/storage/AI%20&%20Racism/7amleh%20-AI%20english1-1.pdf.
4 Amnesty International (2024). "Amnesty International investigation concludes Israel is committing genocide against Palestinians in Gaza". https://www.amnesty.org/en/latest/news/2024/12/amnesty-international-concludes-israel-is-committing-genocide-against-palestinians-in-gaza/.
5 Cupler, S (2023). A Brief Overview of AI Use in WANA. *SMEX*. https://smex.org/a-brief-overview-of-ai-use-in-wana/
6 *Ibid*.
7 *Ibid*.
8 Access Now (2021). *Exposed and Exploited: Data Protection in the Middle East and North Africa.* https://www.accessnow.org/wp-content/uploads/2021/01/Access-Now-MENA-data-protection-report.pdf.
9 OECD/Center of Arab Woman for Training and Research (2014), "Towards women's empowerment in public life in the MENA region", in *Women in Public Life: Gender, Law and Policy in the Middle East and North Africa,* OECD Publishing, Paris.
10 Khafagy, F. et al.(2021). *Women's Economic Justice and Rights in the Arab Region*. https://arabstates.unwomen.org/sites/default/files/Field%20Office%20Arab%20States/Attachments/2021/07/Womens%20Economic%20Justice%20and%20Rights-Policy%20Paper-EN.pdfhttp.

in some countries, women are not allowed to make basic decisions about their lives without the permission of a male relative.[11] In the meantime, LGBTQIA+ individuals are criminalized and face imprisonment, state-coordinated attacks, and violence, online and offline.[12] Gender-based violence is also rampant.[13]

This chapter explores the different forms of gender inequalities AI is exacerbating or risk exacerbating. It focuses on AI systems deployed by governments (in particular, surveillance tech, smart city tech, and automated warfare systems) and the private sector (Generative AI, algorithmic content moderations systems and algorithmic content curation, recommendation, and ranking systems deployed by digital platforms, in addition to systems deployed by the online gig economy and those deployed in recruitment). The chapter also looks at how AI's automation of work risks impacting women in the region and their ability to join and compete in a changing labour market.

Overview of forms of discrimination exacerbated by AI systems

Tech-facilitated gender-based violence (TGBV)

Women and LGBTQIA+ people disproportionately face violence online, and this risk of violence increases for those actively involved in politics and the civic space such as women politicians, activists, Human Rights Defenders (HRDs), and journalists. AI systems exacerbate TGBV through the deployment of bots and Generative AI to create and disseminate content.

The use of bots to target the civic space, including on the basis of gender, is not a new phenomenon and has been well-documented since at least 2011,[14] at a time when the region was at "the height of the Arab Spring,"[15] a wave of pro-democracy protests that started in Tunisia, denouncing government repression, corruption, lack of jobs, among other afflictions, before spreading to other countries, including Egypt, Bahrain, Libya, and Syria. Since then, governments have used bots as part of broader harassment and/or disinformation campaigns aimed at manipulating public discourse and silencing crucial voices like journalists, HRDs, activists, and opposition politicians. In some contexts, bots are also used in tandem with armies of human trolls, making it more challenging for social media platforms to detect and take down.[16]

Generative AI, deployed to generate synthetic text, photos, videos, and photos, poses another risk to women and LGBTQIA+ people. Generative AI's use to create deepfake sexual abuse, as part of gendered disinformation campaigns to threaten, blackmail, and discredit women represent severe risks to their safety and participation in civic and political spheres.[17]

11 Human Rights Watch (2023). *Trapped: How Male Guardianship Policies Restrict Women's Travel and Mobility in the Middle East and North Africa*. https://www.hrw.org/report/2023/07/18/trapped/how-male-guardianship-policies-restrict-womens-travel-and-mobility-middle.

12 Hourany, D. (2023). LGBTQ+ in MENA: Fighting for Rights Against All Odds. *Fanack*. https://fanack.com/human-rights/features-insights/lgbtq-in-mena-fighting-for-rights-against-all-odds~263811/.

13 Hourany, D. (2022). Violence Against Women in MENA on the Rise. *Fanack*. https://fanack.com/society/gender-equality-in-the-middle-east-and-north-africa/violence-against-women-in-mena-on-the-rise/.

14 Leber, A. y Abrahams, A. (2021). Social Media Manipulation in the MENA: Inauthenticity, Inequality, and Insecurity, *POMEPS Studies 43: Digital Activism and Authoritarian Adaptation in the Middle East*. https://pomeps.org/social-mdia-manipulation-in-the-mena-inauthenticity-inequality-and-insecurity

15 Abrahams,A. y Leber, A. (2021). Electronic Armies or Cyber Knights? The Sources of Pro-Authoritarian Discourse on Middle East Twitter. *International Journal of Communication 15* (2021), 1173–1199.

16 Benner, K. et al. (2018). Saudis' Image Makers: A Troll Army and a Twitter Insider. *The New York Times*. https://www.nytimes.com/2018/10/20/us/politics/saudi-image-campaign-twitter.html

17 Gulf Center for Human Rights (2019). Deepfake poses a threat to human rights defenders in the Middle East, *Gulf Center for Human Rights*. https://www.gc4hr.org/deepfake-poses-a-threat-to-human-rights-defenders-in-the-middle-east/

Additionally, social media platforms' content curation, recommendation and ranking systems have been documented to exacerbate the dissemination of gender-based violence and the replication of existing biases and gender stereotypes. "Influencer economies," are particularly important to highlight here as they have been criticized for their normative representation of women.[18] Given the usually large numbers of followers influencers have on social media, algorithmic systems are more likely to recommend their content to more people, which can result in the wider spread of harmful stereotypes towards women and LGBTQIA+ people, and sometimes, even violence or violent incitement against them. In the meantime, platforms' content moderation algorithms have not been effective in promptly detecting and removing violent threats, harassment, misogynistic hate and others forms of TFGBV.

Suppression of SRHR content in content moderation systems

Algorithmic gatekeepers of social media platforms and search engine are hindering access to essential resources and information on Sexual and Reproductive Health and Rights (SRHR). Research, published by digital rights organization SMEX in 2024, found that the restriction of SRHR content by content creators, health experts, activists, and civil society groups "happens on vague grounds, sometimes with illogical or irrelevant explanations, despite the content being innocuous and far from explicit in any form."[19] Advocates and content creators faced in particular challenges to getting their ads to be accepted, signalling potentially a desire from platforms to abide by local advertising laws, which often restrict this type of content.

Automation of work and participation in economic life

There is still a long road ahead towards achieving gender equality in the workforce and economic life, despite milestones towards closing the education gender gap in many countries and legal reforms aimed at promoting women's economic rights.[20] Yet, MENA continues to have the lowest rate of women's participation in the labour force among all regions in the world, with 18.4% in 2021, far below the global average of 48%.[21]

AI risks exacerbating gender inequality in the workplace and access to the job market due to existing and deeply rooted gender norms, particularly when those deploying AI systems, for instance, to screen applications or in platform work, do not account for these norms or aim to mitigate the impacts of the biases their algorithms may end up replicating or exacerbating. Additionally, AI risks replacing jobs that are mostly held by women, such as clerical work,[22] which can further exclude women from participating in economic life, if the gender digital divide and the significant gap in unpaid labour that mostly falls on women's shoulders in the region, is not addressed.[23]

Gendered surveillance

Governments in the region are known for using and acquiring some of the latest and most invasive digital surveillance technologies as a means of monitoring and controlling the be-

18 Bishop, S. (2021). Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. Social Media + Society, 7(1). https://doi.org/10.1177/20563051211003066

19 SMEX (2024). *From Sharing to Silence: Assessing Social Media Suppression of SRHR Content in WANA*. SMEX. https://smex.org/from-sharing-to-silence-assessing-social-media-suppression-of-srhr-content-in-wana/

20 Ferrant, G. y Lunati, M. (2023). The potential of digitalisation for women's economic empowerment in MENA countries, en *Joining Forces for Gender Equality: What is Holding us Back?,* , OECD Publishing, Paris, https://doi.org/10.1787/28736eeb-en.

21 Khafagy, F. et al. (2021). *Women's Economic Justice and Rights in the Arab Region*. Arab States CSOs and Feminists Network. https://arabstates.unwomen.org/sites/default/files/Field%20Office%20Arab%20States/Attachments/2021/07/Womens%20Economic%20Justice%20and%20Rights-Policy%20Paper-EN.pdf.

22 UNESCO, OCDE, BID (2022). *The Effects of AI on the Working Lives of Women.*

23 Khafagy, F. et al. (2021). *Women's Economic Justice and Rights in the Arab Region*. UN-Women. https://arabstates.unwomen.org/sites/default/files/Field%20Office%20Arab%20States/Attachments/2021/07/Womens%20Economic%20Justice%20and%20Rights-Policy%20Paper-EN.pdf

haviours and activities of people, targeting specifically human rights defenders, journalists, dissidents, activists, and political opponents.[24] Surveillance disproportionately impacts women and LGBTQIA+ individuals and communities. Israeli security services have long used surveillance to target LGBTQIA+ Palestinians in the occupied West Bank to blackmail them into becoming informants.[25] Surveillance is also weaponized against women human rights defenders through the extraction of personal and intimate conversation, photos, and other information, which are then used to blackmail, defame, and dox them.[26] Given prevalent gender norms and the levels of "policing" and societal scrutiny women and LGBTQIA+ people face, this gendered surveillance only further puts them at increased risk of repercussions from the authorities or non-state actors, and violates their bodily integrity by exposing them to the risk of more violence and harassment.

AI, with the capabilities it provides for governments to extract more data and conduct both targeted and massive surveillance, for instance, through predictive policing tools[27] and facial recognition, will only further exacerbate gendered surveillance.

Israel's automated occupation and its gender impacts

Israel has long been infamous for its "Big Brother" practices,[28] deploying some of the world's most invasive technologies in its occupation of Palestinian Territories and exporting them abroad.[29] Israel's dehumanization of Palestinians and long-standing goal of subjugating them to its control and occupation are built into the AI systems it deploys from facial recognition, predictive police and other surveillance tech. Women and children are no exception to this dehumanization and its increased automation. In Israel's war on Gaza, where AI systems are deployed to generate targets for killing with minimal human diligence and oversight, most of those killed are women and children.[30] In the occupied West Bank, Palestinian women navigate Israeli checkpoints as spaces of "gendered modes of discrimination."[31] Facial recognition is a key technology deployed by Israel to track Palestinians and refuse or allow them passage through checkpoints.

Influencers, the 'manosphere' and social media algorithms: automating misogyny and the anti-feminist backlash

In March 2021, a homophobic hashtag that incited to violence against gay men trended on Twitter (before its rebranding to X after its acquisition by Elon Musk) in Egypt and Saudi Arabia, two countries in the region with some of the world's most active users on the platform.[32]

The fact that this problematic hashtag trended in Egypt and Saudi Arabia is not a coincidence, given the hostile, and often violent, environment LGBTQIA+ communities and indi-

24 Lynch, J. (2022). Iron net: Digital repression in the Middle East and North Africa. *European Council on Foreign Relations*. https://ecfr.eu/publication/iron-net-digital-repression-in-the-middle-east-and-north-africa/#acknowledgements

25 Chatelle, T. (2024). Palestinian Queers under Israeli surveillance – and threat. *Drop Site*. https://www.dropsitenews.com/p/how-israels-elite-intelligence-unit

26 Fatafta, M. y Front Line Defenders (2023). Unsafe anywhere: women human rights defenders speak out about Pegasus attacks. *Access Now.* https://www.accessnow.org/women-human-rights-defenders-pegasus-attacks-bahrain-jordan/

27 Fatafta, M. y Nashif, N. (2017). The Israeli algorithm criminalizing Palestinians for online dissent. *Open Democracy*. https://www.opendemocracy.net/en/north-africa-west-asia/israeli-algorithm-criminalizing-palestinians-for-o/

28 The Guardian (2014). Any Palestinian is exposed to monitoring by the Israeli Big Brother. *The Guardian.* https://www.theguardian.com/world/2014/sep/12/israeli-intelligence-unit-testimonies

29 Loewenstein, A. (2023). *The Palestine Laboratory: How Israel Exports the Technology of Occupation around the World*. Verso.

30 Oxfam (2024). More women and children killed in Gaza by Israeli military than any other recent conflict in a single year – Oxfam. *Oxfam.* https://www.oxfam.org/en/press-releases/more-women-and-children-killed-gaza-israeli-military-any-other-recent-conflict

31 Griffiths, M. y Repo, J. (2021). Women and checkpoints in Palestine. *Security Dialogue*, 52(3), 249-265. https://doi.org/10.1177/0967010620918529

32 Abrougui, A. (2021). Hate Speech: Why Social Media Platforms Are Failing the LGBTQ Community. *Jeem*. https://jeem.me/en/internet/548

viduals face in MENA. After all, the hashtag trended due to the amount of engagement it generated, a reflection of the existing societal and cultural taboos surrounding gender and sexuality and lack of acceptance towards LGBTQIA+ people. However, algorithmic systems deployed by platforms to curate, recommend and rank content are known to contribute to the spread and wider dissemination of harmful content.33 These systems are designed as part of business models aimed at generating profits from ads targeted to users based on the personal information they readily share online such as their content, personal details such as where they live, what they do for work, etc. in addition to other information that is extracted and inferred about them by tracking their activities and behaviours such as their interests, preferences, desires, dreams, fears, and insecurities.

Increased engagement is essential to maximizing this tracking as it keeps users on these platforms, posting, liking, clicking, sharing, commenting, and the more they engage with types of content or specific posts, the more that content is likely to be recommended by algorithms to other users and going "viral." This virality can result in harmful content "trending" or staying online for long before it's taken down (if ever), as it was the case with the homophobic hashtag that trended on Twitter in Egypt and Saudi Arabia. There have also been several documented cases of social media "influencers" in the past with large numbers of followers posting homophobic content that incited against LGBTQIA+ people.34

When examining the role of content curation, ranking and recommendation systems in facilitating gender-based violence, it is important to highlight how they push content by influencers that perpetuates existing societal biases against women that attempt to confine them to certain roles or behaviours.

Influencer industries are criticized for their normative representation of women.35 In these economies, influencers often produce their content and labour within the constraints set not only by platforms but also the brands for which influencers advertise. In her exploration of influencer management tools, which use algorithms to "support marketers in selecting influencers for advertising campaigns, based on categorizations such as brand suitability, "brand friendliness," and "brand risk," Sophie Bishop found that they "reify existing social inequalities in influencer industries, particularly along the lines of sexuality, gender, and race."

The result is digital spaces that provide space for women and LGBTQIA+ individuals to express themselves, access information, raise awareness, engage, etc. within societal and marketplace norms. The algorithms reflect those norms and by rewarding the voices and content of those who adhere to the norms, it traps everyone in filter bubbles.36 No wonder that some of the most popular women influencers in MENA,37 and elsewhere, focus on "traditionally feminine domains" such as beauty tips and make-up art, fashion, lifestyle, and modelling.38 While in some contexts, posting about these topics can be seen as an attempt to break social taboos that prevent women from making the most basic decisions about their lives such as how to dress, where to go, what to talk about, and there have been cases in MENA of women influencers facing not only online harassment but also imprisonment,39

33 Maréchal, N. y Biddle, E. (2020). It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge, New America Foundation. https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/

34 Access Now (2020). In Tunisia, 45 organizations speak out against Instagram hate campaign targeting the LGBTQ community. *Access Now*. https://www.accessnow.org/press-release/45-organizations-speak-out-against-instagram-hate-campaign-targeting-tunisias-queer-community/.

35 Bishop, S. (2021). Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias. *Social Media + Society*, 7(1). https://doi.org/10.1177/20563051211003066

36 Pariser, E. (2021). The Filter Bubble: What the Internet is Hiding from You. Penguin Books.

37 Raman, N. and Nair, A. (2023). Here are the top influential creators in the Middle East you need to know. *Fast Company Middle East.* https://fastcompanyme.com/recommenders/here-are-the-top-influential-creators-in-the-middle-east-you-need-to-know/

38 *Ibid.*

39 Makooi, B. (2023). *Egypt's female social media influencers face arrest, jail on 'morality' charges*. *France 24*. https://www.france24.com/en/middle-east/20230411-egypt-s-female-social-media-influencers-face-arrest-jail-on-morality-charges

influencer content is "bound to a capitalist system that reifies particular conceptions of femininity,"[40] and the algorithms reproduce those conceptions on a loop. According to Nour Naim, a researcher and expert in AI ethics:

> "From a technical perspective, this is also a socio-cultural problem. AI reflects the identity of a society by feeding it with content and data, and given the nature of algorithms and AI models, the outputs will then reflect existing socio-cultural problems. But, the constant use of these systems, such as those deployed by social media platforms, for long hours, reinforces the continuity of this culture that is biased against women in societies and stereotypes perpetuated against women as if their role is limited to superficial and secondary roles, instead of essential roles that can result in influence, change, leadership, and empowerment. We cannot blame AI systems even if, according to studies and papers that have been published, AI systems like [those of] Facebook and Instagram, go in the direction of bias because they reinforce the reach of content disseminated to increase their profits. Society is more familiar with content that sexualizes women, reduces women to their bodies and in content that is superficial away from the real roles of women…so it is a mirror that doesn't only reflect [society] but also reinforces this bias…the nature of engagement, nature of the data to feed these algorithms are originating from the internet in the region and everything that is available from platforms, internet, and cloud, all this data is already toxic towards women. As a result, without real decision-making to reduce these biases and support gender and sexuality justice, the bias will continue, particularly with current generations that understand the world through social media."

In the meantime, there is a misogynistic backlash against women's rights, gender and feminism worldwide, that is manifesting online in the "manosphere," "a collection of websites, social media accounts and forums dedicated to men's issues" many of which "have become spaces where explicit anti-women and anti-feminist sentiment abound."[41] While these spaces first gained notoriety in the west, "the intrusion of the manosphere into the Arab digital sphere is a clear example of Western misogyny being imported to the region,"42 wrote Sara Kaddoura, a Palestinian feminist activist and researcher who is also a content creator making videos about feminism in Arabic for Arabic-speaking audiences on her YouTube channel Haki Nasawi ("Feminist Talk"). According to her, the manosphere manifests in MENA digital spaces "by adopting language, motifs, and arguments made to fit our local contexts", adding:

> "That is not to deny that there has long been misogyny and sexism in the Arab World, but rather to point out that the sexism of the manosphere is itself a form of intellectual colonization.
>
> The West, according to Arab anti-feminists, was always a space of promiscuity, and its liberalism an intrusive cancer to the region. But ironically, it is the increasingly famous manosphere content creators from the West who have become idols and sources of inspiration for Arab anti-feminists. We are witnessing the birth of content creators imitating the language and mannerisms of Andrew Tate, and preaching about the pill of truth to access inner success, achieve masculinity, and subjugate the women in one's life."

40 *Ibid*.

41 Lawson, R. (2023). A dictionary of the manosphere: five terms to understand the language of online male supremacists. *The Conversation*. https://theconversation.com/a-dictionary-of-the-manosphere-five-terms-to-understand-the-language-of-online-male-supremacists-200206

42 Kaddoura, S. (2024). *The Arab Manosphere: a New Wave of Western Misogyny in the MENA Region*. Friedrich-Ebert-Stiftung. https://feminism-mena.fes.de/e/the-arab-manosphere-a-new-wave-of-western-misogyny-in-the-mena-region.html.

While content posted by those participating in the manosphere and influencer industries continues to spread, and going even viral in some cases, human rights defenders, women's rights activists, journalists, feminists, LGBTQIA+ people and others who speak out against exclusion of women, patriarchy and misogyny have long faced harassment, violence, and other attempts to silence them. Women and LGBTQIA+ people, in particular, risk disproportionate amounts of violence online, not to mention the digital divide gap that prevents women in some countries from being online,[43] whether as a result of not being able to afford access or because of social restrictions that prevent them from being online.[44] This limits their ability to fight and challenge such toxic and misogynistic narratives that are often being spread by influencers with large numbers of followers and as discussed previously by algorithms that reward engagement.

In the meantime, social media platforms deploy content moderation systems that have proven to be ineffective in promptly detecting and removing TFGBV in MENA's contexts, languages and dialects.

In a 2024 paper, Mona Elsawh, of the Center for Democracy and Technology (CDT), noted that smaller language models dedicated to the Arabic language did a better job that Large Language Models (LLMs), which "indicates that the problem does not lie in the inherent difficulty of Arabic but rather in the level of dedication and willingness to invest in improving AI models that meet Arabic's unique characteristics."[45] She added:

> "Arabic was never a top priority to AI developers and is unlikely to become one in the near future. Consequently, this might impede the creativity and innovation of Arab Internet users. Moreover, it may result in many mistakes and errors. On the one hand, it could lead to censoring and restricting Arab users' freedom of expression, with their social media posts being mistakenly flagged and removed by the algorithms. On the other hand, the poor design of AI tools could also lead to the spread of misinformation and hate speech, leaving such content without removal."

TFGBV's impacts in MENA are well documented, particularly when it comes to preventing women and LGBTQIA+ people from fully and safely expressing themselves without violent and serious repercussions, which is essential to their participation in civic and political spheres.

In one notable example from 2018, prominent Saudi women's rights defender Manal al-Sharif deleted her Twitter account, where she had 290,000 followers, in protest at the platform's use to "put my life and the lives of a lot of human rights activists in danger", adding that "Twitter is being controlled by trolls, pro-government mobs, and by bots" that are being paid by governments to "intimidate, harass dissidents and anyone who speaks the truth."46 Her case is not unique. A 2021 study by the Syrian Female Journalists Network (SFJN) found that Syrian women journalists and human rights defenders were frequently subjected to "sexist attacks and speech, hacking of accounts, threats of bodily harm and death threats, and doxing" on social media. It noted how some Syrian women HRDs and journalists had to change their behaviours, by closing their social media accounts, self-censoring or decreasing their activities online or completely retreat from the public space.[47]

43 Traidi, A. (2024). Gender digital divide: The new face of inequality in the MENA region. *Global Campus on Human Rights*. https://gchumanrights.org/gc-preparedness/preparedness-gender/article-detail/gender-digital-divide-the-new-face-of-inequality-in-the-mena-region.html.

44 Kaddoura, S. (2024). *The Arab Manosphere: a New Wave of Western Misogyny in the MENA Region*. Friedrich-Ebert-Stiftung. https://feminism-mena.fes.de/e/the-arab-manosphere-a-new-wave-of-western-misogyny-in-the-mena-region.html.

45 Elswah, M. (2024). *Does AI Understand Arabic? Evaluating the Politics Behind the Algorithmic Arabic Content Moderation*. Cambridge, MA: Harvard Kennedy School.

46 "*Why I deleted my Twitter account*," Manal al-Sharif on YouTube, octubre de 2018, https://www.youtube.com/watch?v=8regaO3hl_g.

47 Abrougui, A. and Asad, R. (2021). *Digital Safety Is A Right. Syrian Women Journalists and Human Rights Defenders in the Digital Space: Risks and Threats, Syrian Female Journalists Network*. Stichting Female Journalists Network https://media.sfjn.org/en/digital-safety-is-a-right/.

Yet, time and again, platforms and their algorithms fail to act proactively on TFGBV, contributing to a hostile environment for women and LGBTQIA+ people, and leading to their further exclusion from political participation and civic life.

In an interview, Elswah, attributed these shortcomings to three factors: sources of the training data, annotation of the data, and the way the models are built. Many dialects spoken in the region are considered low resource, meaning that there is not enough quality data to adequately train algorithmic systems on. The annotation process of the data also presents opportunities for errors. According to her:

> "So, when you get the data, you need some people to annotate it, categorize it, to label it so you can actually start processing it and start building your model and the annotators. It's such a boring work, and it's usually underpaid, and it comes with bias because the people who annotate are human beings and if you are hiring people who are all male, for example, and you're asking them to annotate data there, they will show some bias."

Platforms' solutions for detect harmful content in languages other than English has been to deploy Multilingual Language Models (MLMs).[48] Through training on data from multiple languages at the same time, MLMs "infer connections between languages, allowing them to uncover patterns in higher resourced languages and apply them to lower resourced languages."[49] However, these models have not necessarily improved content moderation in MENA in such a way that they are ensuring adequate and timely detection and removal of harmful content while at the same time ensuring freedom of expression and information is not severely impacted by erroneous removals and censorship. A 2023 study by CDT identified four limitations in these models: they often reply on machine-translated text that contain errors and do not reflect native languages, do not work well in all languages, and fail to consider and reflect the contexts of local language speakers when problems, and when problems arise in these models, they are hard to identify and fix.[50]

Further, Generative AI, and its use to generate sexual abuse deepfake, which disproportionately target women, will make it digital spaces even less safe. For instance, In Iraq, in the 2021 parliamentary elections, one woman candidate, was blackmailed using a sexual abuse deep fake, forcing her to drop out of the race.[51] Yet, technology companies are not taking adequate steps to address the shortcomings of their algorithms, whether those that curate and recommend content or moderate. In fact, with mass layoffs affecting their trust and safety teams,[52] reliance on algorithms, without proper oversight from human moderators, is expected to increase. risking exacerbating biases and inaccuracies in content moderation in MENA's languages and dialects.

"They [tech companies] are laying them [human moderators] off while touting their machine learning capabilities, they have decided this is the future. If you go into conversations in the US with content moderators, they don't want to do this work anymore…internally the content moderators themselves want an automated structure, they don't want the trauma of looking at this horrible stuff," Azza El Masri, who is pursuing her doctoral studies in Journalism and Media at the University of Texas, said in an interview. She emphasized the need for a change in strategy from civil society groups that prioritize the machine learning capabilities of these platforms.

48 Nicholas, G. and Bhatia, A. (2023). The Dire Defect of 'Multilingual AI Content Moderation. *Wired*. https://www.wired.com/story/content-moderation-language-artificial-intelligence/

49 Nicholas, G. and Bhatia, A. (2023). Lost in Translation: Large Language Models in Non-English Content Analysis". *Center for Democracy and Technology.* https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf.

50 *Ibid.*

51 Al-Kaisy, A. (2022). *Online violence towards women in Iraq*. Elbarlement. https://elbarlament.org/wp-content/uploads/2022/03/Aida-2.pdf

52 Motyl, M. and Ellingson, G. (2024). The Unbearably High Cost of Cutting Trust & Safety Corners. *Tech Policy Press*. https://www.techpolicy.press/the-unbearably-high-cost-of-cutting-trust-safety-corners/

Algorithmic gatekeepers' suppression of SRHR content

There are multiple barriers that prevent populations in MENA countries from fully enjoying their Sexual and Reproductive Health and Rights (SRHR), although the situation differs from country to country.[53] Taboos, stigma, and cultural sensitivities still largely surround SRHR including abortion and contraception, sexually transmitted diseases and infections, and the ability to make informed decisions about one's body.[54] Access to information that is factual, inclusive and non-stigmatizing is essential. Yet, in the absence of sex education programs and with the lack of open conversation within family settings about reproduction and sexuality, "young people resort to the Internet, online pornography, and peers as sources of SRH information, which are often inaccurate and potentially harmful to equitable gender norms."[55] It is once again worth noting that while the internet is indeed a key resource, the digital divide and gender digital divide remains a reality in some countries. Additionally, certain governments practice censorship, on political, social and religious grounds,[56] affecting what content users can access, including SRHR content. Beyond these two hurdles, algorithmic gatekeepers have also been shown to suppress SRHR content or recommend content that is non-factual or stigmatizing.

In a 2018 article for Jeem, a regional feminist media organization, author Salma Mohamed, recounted her experience as a teen looking up information on the internet about sexual health and sexuality in Arabic, and getting results filled with health misinformation, stigma, and religious edicts.[57] Years later, as a medical student it occurred to her to start looking up the same information in English, and she ended up receiving totally different results that did not perpetuate stigma. Once again here, the algorithmic systems of digital platforms–in this case search engines–are reflecting the data with which they are being fed.

Additionally, content creators, health practitioners, activists and civil society groups posting on social media about SRHR, including to raise awareness, constantly face censorship and removal of their content, ads, and accounts. Those who post in Arabic were more likely to face censorship than in English, creating another obstacle to accessing essential information and resources for those who only speak Arabic.[58] Platforms rely heavily on automation to moderate content and ads, and research by digital rights organization SMEX documented several examples of educational SRHR content being taken down for violating platforms' policies on "adult content," raising questions as to what extent the algorithms understand the context in which this content is being posted, particularly in Arabic.[59]

According to researcher Mira Nabulsi who studied the restrictions SRHR content in the region faces, several creators she interviewed believed Meta was helping "maintain a very conservative status quo in our societies that censors female bodies and important information pertinent to people's rights to make decisions about their lives and futures."[60]

53 As an example, Tunisia is the only country in the region where abortion is legal on demand during the first trimester. In the rest of the region, access and restrictions vary. While all countries permit abortion when the pregnant woman's life is in danger, some countries permit it when the woman's physical and / or mental health is in danger, in cases of fetal impairment, or rape [see: Maffi I, Tønnessen L. The Limits of the Law: Abortion in the Middle East and North Africa. Health Hum Rights. 2019 Dec;21(2):1-6. PMID: 31885431; PMCID: PMC6927385.]

54 Oraby D. Sexuality Education for Youth and Adolescents in the Middle East and North Africa Region: A Window of Opportunity. Glob Health Sci Pract. 2024 Feb 28;12(1):e2300282. doi: 10.9745/GHSP-D-23-00282. PMID: 38290752; PMCID: PMC10906548.

55 *Ibid*.

56 Freedom House (2024). *Internet Freedom in the Middle East Remained Restricted in 2024*. https://freedomhouse.org/article/fotn-2024-middle-east-release

57 Mohamed, S. (2018). كيف ساهم المحتوى العربي المغلوط في تشويه حياتنا الجنسية؟ [How misinformation in Arabic contributed to distorting our sexual lives., *Jeem*. https://jeem.me/bodies/116.

58 SMEX (2024). From Sharing to Silence: Assessing Social Media Suppression of SRHR Content in WANA. *SMEX*. https://smex.org/from-sharing-to-silence-assessing-social-media-suppression-of-srhr-content-in-wana/.

59 *Ibid*.

60 Nabulsi,M. (2024). Navigating Taboos: Exploring social media policies and SRHR content restrictions in WANA". *SMEX*. https://smex.org/wp-content/uploads/2023/03/MiraNabulsi-SRHR-Mariam-al-Shafei-Fellowship-2023.pdf.

In this sense, while the algorithms are feeding on stigmatizing data surrounding SRHR, platforms are further exacerbating those stigmas and inequalities by adopting a biased content moderation towards the region and punishing those who challenge existing taboos and norms. This is clearly reflected, for instance, in platforms' advertising policies, which seem to be grounded in the local laws and/or social conservatism of MENA countries, resulting in additional censorship and restrictions. For instance, X bans "promoting non-prescription contraceptives" in many MENA countries.[61] YouTube, on the other hand, does not allow "ads related to birth control or fertility products" in 23 countries, most of which in MENA (17 countries).[62] Meta has an overall ad policy that allows "ads promoting sexual and reproductive health or wellness products or services, such as contraception and family planning" as long as they do "not focus on sexual pleasure."[63] TikTok has similar policies.[64]

The drawbacks of automation on women's already precarious participation in economic life

AI risks replacing jobs that are typically held by women such as secretaries, accountants, bookkeepers, and administrative assistants. While the negative impacts of automation will be most felt in high-income countries,[65] in middle-income countries, some jobs, such as call center work, will still be exposed to the risk of automation. In the region, call centers employ many people in middle income countries like Tunisia66 and Morocco,[67] including women.

Nagla Rizk, professor of economics at the American University of Cairo School of Business, and founding director of the Access to Knowledge Development Center, said in an interview that "the higher the skill level the more likely it will be enabled by technology, as you go down the skill structure, especially the medium skills, anything that has repetition is likely to be replaced by a machine."

AI's impacts on women's participation in the workplace are the result of deeply rooted biases that encourage women or limit their work to certain roles and jobs. With AI automating some of these functions and jobs that they typically hold, it risks eroding women's participation in the workplace, particularly with women not having equal access to opportunities as men to catch up with the new demands of a changing job market because of the amount of unpaid labor and care work they have to do on a daily basis compared to men. Unequal gender norms in MENA are still prevalent and women are still largely expected to be the caregivers and conduct household activities or to limit paid work to certain roles and certain sectors.[68]

61 "X Business. Healthcare". acceso del 15 de noviembre de 2024. https://business.x.com/en/help/ads-policies/ads-content-policies/healthcare.

62 "Healthcare and medicines". Políticas de Google sobre publicidad. Acceso del 15 de noviembre 2024. https://support.google.com/adspolicy/answer/176031?hl=en&ref_topic=1626336&sjid=13573517111321317776-EU#zippy=%2Ctroubleshooter-birth-control.

63 Meta. "About Meta's Health and Wellness advertising policy". Acceso del 15 de noviembre de 2024. https://www.facebook.com/business/help/2489235377779939?id=434838534925385

64 TikTok. "Adult Content". Políticas de TikTok sobre publicidad. Acceso del 1 https://ads.tiktok.com/help/article/tiktok-ads-policy-adult-content y "Healthcare and Pharmaceuticals," Políticas de TikTok sobre publicidad, acceso del 15 de noviembre del 15. https://ads.tiktok.com/help/article/tiktok-ads-policy-healthcare-pharmaceuticals

65 Gmyrek, P., Berg, J., Bescond, D. (2023). *Generative AI and jobs: A global analysis of potential effects on job quantity and quality*, Documento de trabajo de la OIT 96 (Ginebra, OIT). https://www.ilo.org/publications/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and

66 Ahmed, S. (2024). Le télémarketing: Un secteur négligé malgré ses contributions cruciales à l'économie tunisienne, *La-Presse*. *https://lapresse.tn/2024/09/08/le-telemarketing-un-secteur-neglige-malgre-ses-contributions-cruciales-a-leconomie-tunisienne/*

67 TRTFrançais (2024). Intelligence Artificielle, le grand remplacement dans les centres d'appels marocains?. *TRT Français*. https://www.trtfrancais.com/actualites/intelligence-artificielle-le-grand-remplacement-dans-les-centres-dappels-marocains-17699021

68 Nazier, H. (2019). *Women's Economic Empowerment: An Overview for the MENA Region*. Instituto Europeo del Mediterráneo. https://www.iemed.org/publication/womens-economic-empowerment-an-overview-for-the-mena-region/.

The gender digital divide in some cases also prevents women from developing their digital capacities so that their skills stay relevant to a changing job market. 69 In MENA, women are 12% less likely to use the internet than men because they are unable to afford internet access, lack digital skills or as a result of gender norms that limit their presence and access to the internet.70 Additionally, even when women have access, they are often limited in the time they can allocate to engage with ICTs and further develop their skills given the domestic labour that mostly falls on their shoulders.[71]

Another challenge is the exacerbation of discrimination against women in AI tools deployed in recruitment and in the online gig economy.

As more employers and recruitment agencies[72] [73] [74] rely on AI solutions and career platforms that deploy AI like LinkedIn and Bayt.com to sift through applications, select interviewees and eventually hire people, there are concerns that this will reduce women's chances of getting hired.

According to Sarah Cupler, a PhD candidate at the University of Melbourne researching police use of automated decision-making tools: "Data is highly influenced by the collection process and how it is labelled, there can be a lot of biases in data and reflect historic discrimination…If a woman is more likely to have to quit working due to economic, work, societal pressures once pregnant, the algorithm could see that as women are less likely to be able to succeed. AI primarily when used uncritically can reflect back society as it is but make it seem objective and neutral and perpetuate these problems we have."

"We already have this problem of bias against women in employment. Any AI system in the region will be fed with such data and will reflect that problem. I am certain that this is the case, with a few exceptions, where there is awareness in a particular institution about bias…and if there was real awareness, we would have seen this reflected in traditional recruitment mechanisms," Naim said. She further explained that these biases are compounded by women's lack of representation in AI companies, including those that provide AI solutions. In fact, while in the region, more women have been graduating in STEM fields, their representation in the workforce remains disproportionate75 and women also struggle to reach higher and executive level positions. This "prevents them [women] from taking on higher positions such as of managers, CEOs, and executive positions that have broad influence within companies and institutions, so when they are prevented from reaching these positions, their influence and existence and the power they hold in these positions are taken away from them, and this will contribute to the continuity of biases.''

In online gig work or platform work, algorithms are also exacerbating biases.

Given the high unemployment rates particularly among women and youth in MENA, platform work offers opportunities for many to participate in economic life and generate inco-

69 UNESCO, OCDE, BID (2022). *The Effects of AI on the Working Lives of Women*.
70 Traidi, A. (2024). Gender digital divide: The new face of inequality in the MENA region. *Global Campus on Human Rights.* https://gchumanrights.org/gc-preparedness/preparedness-gender/article-detail/gender-digital-divide-the-new-face-of-inequality-in-the-mena-region.html.
71 Rizk, N. (2020). Artificial Intelligence and Inequality in the Middle East: The Political Economy of Inclusion, in Dubber, M. D.; Pasquale, F. y Das, S. (eds), *The Oxford Handbook of Ethics of AI*. https://doi.org/10.1093/oxfordhb/9780190067397.013.40, acceso del 15 nov. 2024.
72 Maharat. Révolutionnez votre recrutement avec l'IA. Acceso del 16 de diciembre de 2024. https://maharat.ma/
73 Look Up Tunisie. Les Nouvelles Technologies dans le Recrutement. Acceso del 16 de diciembre de 2024. https://www.lookuptunisie.com/les-nouvelles-technologies-dans-le-recrutement/
74 Kader. Get to know Kader. Acceso del 16 de diciembre de 2024. https://www.kaderapp.com/en/about-us
75 Ferrant, G. and Lunati, M. (2023). *The potential of digitalisation for women's economic empowerment in MENA countries*, en Joining Forces for Gender Equality: What is Holding us Back?, Publicaciones de la OCDE, París, https://doi.org/10.1787/28736eeb-en.

me.[76] However, this type of work is "heavily gendered"[77] and perpetuates deeply-rooted gender norms and existing biases against women.[78] For example, domestic gig work, such as cleaning and childcare, remains to be dominated by women.[79] In addition, its illusory promise of flexibility encourages women to take on gig work so that they can at the same conduct unpaid household work and care work, further entrenching gender inequality in unpaid work.[80]

One concern cited by women drivers in Egypt in a 2018 study is that rating systems and existing screening procedures do not effectively mitigate safety concerns, particularly when it comes to risks posed by male passengers to women drivers.[81]

Additionally, women face pay inequality. Rizk who researched women working in rides-haring and delivery apps in Egypt gave as an example how algorithms penalize women in bonuses, without considering their socio-cultural realities and contexts:

> "For ride sharing—this is information we got from people on the ground—they [ride sharing apps] base the bonuses on the number of hours put at work. This is where it gets dangerous as the algorithm reflects and amplifies biases on the ground. So, if your culture allows women to work fewer hours than men because they have to take care of the kids, do the housework... if the algorithms and the machine reflect what's happening on the ground, it is amplifying it. Going back to ridesharing, immediately women are going to be excluded from bonuses."

Smart cities, gendered surveillance and bodily integrity

The deployment of automated surveillance capabilities will only risk exacerbating the threats surveillance poses in the region.[82]

"Now we're moving into an ability to do mass surveillance much easier and an ability to process that data much more quickly, which would have a huge impact on civil society," Cupler said.

Governments in the region have been expanding their AI capabilities for surveillance purposes. Outside Israel, the Gulf region has shown the most interest and biggest investment in the deployment of facial recognition and other surveillance tech.[83] Qatar, for instance, deployed a vast network of CCTV cameras equipped with facial recognition capabilities during the 2022 FIFA World Cup.[84] In the United Arab Emirates (UAE), Dubai police partnered

76 Rizk, N. (2020). Artificial Intelligence and Inequality in the Middle East: The Political Economy of Inclusion, in Dubber, M. D.; Pasquale, F. and Das, S. (eds), *The Oxford Handbook of Ethics of AI*. https://doi.org/10.1093/oxfordhb/9780190067397.013.40, acceso del 15 nov. 2024.

77 Siddiqui, Z. and Zhou, Y. (2021). How the platform economy sets women up to fail. *Rest of World*. https://restofworld.org/2021/global-gig-workers-how-platforms-set-women-up-to-fail/

78 Al-Kaisy, A. (2021). *Bias In, Bias Out: Gender and work in the platform economy*. IDRC. https://idl-bnc-idrc.dspacedirect.org/items/7d8e2f97-b1dd-49ad-9843-a0480f5f80eb.

79 Fairwork (2022). *Domestic Platform Work in the Middle East and North Africa*, Fairwork: https://fair.work/en/fw/publications/domestic-platform-work-in-the-middle-east-and-north-africa/

80 Interview with Nagla Rizk.

81 Rizk, N. et al. (2018). A Gendered Analysis of Ridesharing: Perspectives from Cairo, Egypt. En: *Urban Transport in the Sharing Economy Era*. CIPPEC. https://www.cippec.org/wp-content/uploads/2018/09/UrbanTransport-completo-web_CIPPEC.pdf

82 Business and Human Rights Resource Center (2024). *Keeping watch: Surveillance companies in Middle East & North Africa*. https://www.business-humanrights.org/en/from-us/briefings/mena-surveillance-2024/

83 Cupler, S. (2023). "*A Brief Overview of AI Use in WANA*". SMEX. https://smex.org/a-brief-overview-of-ai-use-in-wana/

84 Zidan, K. (2022). The Qatar World Cup Ushers in a New Era of Digital Authoritarianism in Sports. *The Nation*. https://www.thenation.com/article/society/qatar-world-cup-surveillance/

with SAS, a vendor of AI solutions, for the provision of predictive policing solutions[85] and in Abu Dhabi, police rely on machine learning solutions and facial recognition to predict crime and direct patrol cars to areas considered "high risk".[86] Other MENA countries are showing interest too. The municipality of Greater Amman, in the capital of Jordan, announced in 2023 plans to start using facial recognition technology to "help to improve security, reduce crime, and make the Capital more efficient."[87] Development of smart cities is proliferating with many governments planning to invest in the sector.[88] Egypt's New Administrative Capital (NAC), for instance, will be equipped with a network of 6,000 surveillance cameras, manufactured by U.S. company Honeywell, that will feed into a command and control center that runs "sophisticated video analytics to monitor crowds and traffic congestion, detect incidents of theft, observe suspicious people or objects, and trigger automated alarms in emergency situations."[89] In addition to the cameras, residents will be tracked using mobile phone trackers, digital check points and digital control gates in public transport stations. According to Waisová: "The inhabitants of the NAC have only a limited possibility of living authentically and experiencing a natural and organic development of society. The NAC has become an instrument of segregation, exclusion and a source of social, political and economic inequality."[90]

Automation exacerbates surveillance and its impacts on women ad LGBTQIA+ people, who already face high levels of scrutiny because of their gender, gender identity / expression or sexual orientation. With AI making it easier for governments to track and collect data, and in a context that lacks robust privacy protection and independent judicial oversight,[91] their bodily integrity, and ability to freely move, exercise their freedoms of thought, opinion, and simply make decisions about their lives and bodies will be further made harder under the constant watchful eyes of the State. In Iran, evidence emerged of the government's use of facial recognition, web traffic analysis, and geolocation and other AI tools to police and enforce mandatory hijab rules on women and crack down on women's rights movement.[92]

It is not hard to imagine more evidence emerging in the future of such technologies being deployed to further control women, particularly in countries that have "male guardianship" policies, places restrictions on women's movements and freedoms, such as to travel or obtain a passport, without the approval of so-called male guardians, usually their husbands if they are married, or a father, brother, uncle, grandfather, or even a son in some cases.[93] For example, in Jordan, Kuwait, Qatar, and Saudi Arabia, male guardians and other family members can report women to the police for being "absent" from their homes. In Bahrain, Iran, Kuwait, Oman, Qatar, Saudi Arabia, and the UAE, women at state universities cannot go on field trips or stay at or leave campus accommodations without the permission of their male guardians.[94]

85 Cupler, S. (2023). A Brief Overview of AI Use in WANA. *SMEX*. https://smex.org/a-brief-overview-of-ai-use-in-wana/.
86 Dawood, A. (2021). AI looks at historical data to predict future crimes for UAE's police force. *Mashable Middle East.* https://me.mashable.com/tech/15800/ai-looks-at-historical-data-to-predict-future-crimes-for-uaes-police-force
87 Alakaleek, H. (2023). Facial recognition technology usage in Jordan. *Jordan News.* https://www.jordannews.jo/Section-36/Opinion/Facial-recognition-technology-usage-in-Jordan-3121
88 Belaïd, F., Amine, R., Massie, C. (2024). Smart Cities Initiatives and Perspectives in the MENA Region and Saudi Arabia. In: Belaïd, F., Arora, A. (eds) *Smart Cities. Studies in Energy, Resource and Environmental Economics*. Springer, Cham. https://doi.org/10.1007/978-3-031-35664-3_16
89 Thomson Reuters Foundation (2023). FEATURE-CCTV cameras will watch over Egyptians in new high-tech capital. *Reuters*. https://www.reuters.com/article/business/media-telecom/feature-cctv-cameras-will-watch-over-egyptians-in-new-high-tech-capital-idUSL8N33I0DO/
90 Waisová Š. (2022) The Tragedy of Smart Cities in Egypt. How the Smart City is Used towards Political and Social Ordering and Exclusion. *Applied Cybersecurity & Internet Governance*. (1):1-10. doi:10.5604/01.3001.0016.0985.
91 Access Now (2021). "*Exposed and Exploited: Data Protection in the Middle East and North Africa*". https://www.accessnow.org/wp-content/uploads/2021/01/Access-Now-MENA-data-protection-report.pdf.
92 George, R. (2023). The AI Assault on Women: What Iran's Tech Enabled Morality Laws Indicate for Women's Rights Movements. *Council on Foreign Relations*. https://www.cfr.org/blog/ai-assault-women-what-irans-tech-enabled-morality-laws-indicate-womens-rights-movements
93 Human Rights Watch (2023). Trapped: How Male Guardianship Policies Restrict Women's Travel and Mobility in the Middle East and North Africa*. Human Rights Watch*. https://www.hrw.org/report/2023/07/18/trapped/how-male-guardianship-policies-restrict-womens-travel-and-mobility-middle.
94 *Ibid*.

AI and smart city tech will only make it easier to track women's movements and activities, which can further encroach on their freedoms and possibly put their safety in danger (for instance, when domestic abuse victims are tracked and forcibly returned to their abusers). As Waisová wrote about the NAC in Egypt, inhabitants will "have only a limited possibility of living authentically".[95]

The automation of occupation and genocide: the devastating impacts on Palestinian women and children

During Israel's ongoing genocide in Gaza, Israeli military has been deploying AI systems to generate targets for killing and commit homicide (massive destruction of homes). One of the tools, called "Where's Daddy", is used to track targets and bomb them once they arrive at their family residences. Another tool, called "Lavender", marked tens of thousands of Palestinians in Gaza as "suspects", and according to an investigation by +972 Magazine and the Local Call, the system was known to the military "to occasionally mark individuals who have merely a loose connection to militant groups, or no connection at all."[96]

"The scale of destruction that we have seen in Gaza is only possible because AI technology made it much faster to make a decision", Cupler noted.

This has had devastating impacts on civilians in Gaza, with nearly 70% of those who died in the conflict are women and children according to UN data.[97] Analysis by Oxfam in September 2024 found that "More women and children have been killed in Gaza by the Israeli military over the past year than the equivalent period of any other conflict over the past two decades,"[98] underscoring the impact of the violence on women and children as a result of Israel's dehumanization of Palestinians, lack of due diligence in war conduct, including in the use of automated systems, to target civilian infrastructure such as schools, homes, hospitals, shelters. Umaiyeh Khammash, director of Oxfam partner Juzoor, which is supporting hundreds of thousands of people in more than 90 shelters and health points across Gaza,[99] noted how women in Gaza are "bearing a double burden": "Many have suddenly become the heads of their households, navigating survival and care in the midst of destruction. Pregnant and breastfeeding mothers have faced immense difficulties, including from the collapse in healthcare services".

At checkpoints in the occupied West Bank, Israel deploys facial recognition as part of a vast network of surveillance cameras that scan Palestinians' faces, add them to surveillance databases without their consent to keep a constant tab on them "part of a deliberate attempt by Israeli authorities to create a hostile and coercive environment".[100]

Any Palestinian attempting to navigate these checkpoints faces a repressive reality that is fraught with "extensive periods of waiting, invasive interrogation and identity checks, and

95 Waisová Š. (2022) The Tragedy of Smart Cities in Egypt. How the Smart City is Used towards Political and Social Ordering and Exclusion. *Applied Cybersecurity & Internet Governance*. (1):1-10. doi:10.5604/01.3001.0016.0985

96 Abraham, Y. (2024). 'Lavender': The AI machine directing Israel's bombing spree in Gaza. +972 *Magazine*. https://www.972mag.com/lavender-ai-israeli-army-gaza/.

97 Farge, E. (2024). Gaza women, children are nearly 70% of verified war dead, UN rights office says. *Reuters*. https://www.reuters.com/world/middle-east/nearly-70-gaza-war-dead-women-children-un-rights-office-says-2024-11-08/

98 Oxfam (2024). More women and children killed in Gaza by Israeli military than any other recent conflict in a single year – Oxfam. Oxfam. *https://www.oxfam.org/en/press-releases/more-women-and-children-killed-gaza-israeli-military-any-other-recent-conflict.*

99 *Ibid*.

100 Amnistía Internacional (2023). Israel/OPT: Israeli authorities are using facial recognition technology to entrench apartheid. Amnistía. https://www.amnesty.org/en/latest/news/2023/05/israel-opt-israeli-authorities-are-using-facial-recognition-technology-to-entrench-apartheid/

the constant threat of violence."[101] For women, these checkpoints further represent "highly gendered impositions of (im)mobility, embodied experience and relations of care," authors Mark Griffiths and Jemima Repo argued in a paper exploring the gendered dimensions of Israel's checkpoints on Palestinian women.[102] In fact, they found that women's ability to cross these checkpoints are limited to their roles as caregivers (i.e. if they are accompanying family members who are getting medical treatment) or for religious reasons, further reinforcing existing gender norms that position men as heads of households and breadwinners and women as caregivers.

Given Israel's reliance on AI systems at checkpoints, those biases and forms of discrimination will only be exacerbated. There is thus a need to investigate more the impacts of its use of AI on women and girls in Palestine and elsewhere as the genocide in Gaza continues unabated, and as Israel expands its attacks on Palestinians in the West Bank other and countries including in Lebanon and Syria.

Conclusion and recommendations

This chapter explored the multidimensional impacts of AI on gender justice in the MENA region. It specifically looked at the roles of Generative AI, bots, and algorithmic systems deployed by social media platforms in spreading gender-based violence and harmful stereotypes against women and LGBTQIA+ people in the region. In the meantime, women journalists and human rights defenders, feminist activists, LGBTQIA+ communities, and others seeking to counter these narratives, are faced with censorship and further violence due to MLMs that that produce inaccuracies and do not work well in MENA's diverse contexts, languages and dialects.

For El Masri, one solution could be to put in more resources in building Small Language Models (SLMs): "they are just as strong and in fact carry more context, you can create more parameters for context in an SLM than you can in an LLM [Large Language Model]. You can also do it in a collaborative way." Investing in such localized and community-driven solutions, in an inclusive way that ensures the participation of women, LGBTAIQ+ people, minorities, and people of diverse expertise can help address the content moderation harms of these platforms, particularly as these harms will likely exacerbate with more advanced technological development and reduced human oversight.

As automation threatens replacing jobs or job functions that tend to be repetitive, there are concerns that this will disproportionately impact jobs typically held by women. To prevent the erosion of women's participation in the workplace, it is essential for different stakeholders from governments, private sector, local and international organizations to prioritize the upskilling of women and all those working in jobs at risk of automation. Beyond designing and launching upskilling programs, there is also a need to bridge the gender digital divide and counter existing gender norms that place on women the responsibilities of unpaid care work limiting their possibilities to catch up with a changing job market. The latter will be an uphill battle as those norms are entrenched, however, by working with women, local women's groups and women's rights organizations it is possible to design upskilling programs that consider the needs and realities of women. Most importantly, employers need to be incentivized so that their employees can dedicate certain working hours per week or month to upskilling.

101 Griffiths, M. y Repo, J. (2021). Women and checkpoints in Palestine. *Security Dialogue*, 52(3), 249-265. https://doi.org/10.1177/0967010620918529
102 *Ibid.*

In MENA, where many countries have high unemployment rates, particularly among the youth and women, platform work provides a lifeline for many, including the opportunity for women to participate in economic life and generate their own income. But this type of work remains divided across existing gender norms and the algorithms of gig platforms often end up reflecting existing biases against women. It is thus essential to document more the impacts of these apps and their algorithms on workers from a gender perspective. There is also a need for more initiatives that proactively involve gig workers to understand their concerns, needs, and use that knowledge to push gig platforms to change their policies and address biases in algorithms.

Finally, governments' deployment of AI, particularly facial recognition, tracking technologies, automated decision-making systems pose serious threats to human rights, the civic space, and civilians. The most severe of these risks have emerged from Israel's deployment such systems in its occupation of Palestinian territories, including its ongoing war in Gaza. There is an urgency to regulate these technologies based on international human rights law, yet States may choose to disregard any such measures. International governance should thus be combined with pressure on technology companies that are complicit in these violations such as providers of facial recognition and surveillance technologies and companies that provide cloud computing and machine learning services such as Google and Amazon Web Services.[103] For example, pushing for export regulations to restrict the sale of these technologies such as providers of facial recognition and the boycott of companies that disregard human rights. These tactics should be adopted beyond the context of Israel's war on Gaza, since other MENA governments are increasing their investments in AI-enabled surveillance technologies, particularly facial recognition and smart city tech, to step up their control and oppression. AI will make it easier for States to track people, collect more data about them and analyse it. This will make it even harder to escape the State's watchful eyes, and women and LGBTQIA+ people, who already face high levels of scrutiny and discrimination, will be disproportionately impacted. Their most basic freedoms will be subjected to constant monitoring and tracking, and with that they risk losing whatever little margins of freedoms they have had so far to live as freely and as authentically as they could.

103 Fatafta, M. and Leufer, D. (2024). Artificial Genocidal Intelligence: how Israel is automating human rights abuses and war crimes. Access Now. https://www.accessnow.org/publication/artificial-genocidal-intelligence-israel-gaza/

Summary: Using the Different Threads to Weave a New Model of Artificial Intelligence

Carlos Bajo Erro
Oxfam Intermón
Translation: Teri Jones-Villeneuve

A politically charged technology

Artificial intelligence (AI) is the driving technology behind some of the most innovative medical research on early, non-invasive and safe diagnosis of deadly diseases as well as efforts to prevent and deal with the ever more frequent and violent extreme climate events. But it is also the technology behind superfluous apps that create supposedly amusing memes, cute cat videos and chatbots that operate as personal assistants, but which cannot guarantee accurate instructions or user safety. The same goes for the tools used to increase the credibility of fake news with graphic images or manipulated videos, as well as those that make use of sexual content that at times includes unauthorized images. This diversity makes it substantially more complicated to adopt a clear stance regarding this technology, which is fuelling a publicity boom while becoming both a vital source of hope and trend all at once. However, perhaps it is not about taking a stance on a specific technology so much as analysing and taking decisions based on its uses and consequences.

AI is now a part of the international production chains stemming from the globalization process over recent decades. As such, if we created a map to connect the points around the world where any process within the AI life cycle occurs, very few places would remain disconnected. Such a map would show the relationships between the places where materials are extracted and those where microprocessors are built, or where electronics with very short life cycles from data centres are disassembled; the places from which data are extracted to design models or where data centres are located; the places where models are trained or where the people who label and refine data live; and more broadly, the people around the world who do essential microjobs to build AI-based tools.

This map could also include the locations of people affected by these developments. This new layer would show the people whose lives come into contact with the algorithmic systems involved in managing their work, in providing public services or in offering entertainment. It would show those who participate in the design, development or implementation of the tools, from the most precarious jobs to top management; those who provide the data (with or without their awareness) that are used by the training processes – the creators of written, verbal, graphic or audiovisual work as well as users of digital tools or simply people whose actions have left traces in a database. Finally, it would also show those who experience the indirect effects of some of the processes, from recipients of a vaccine discovered through health research using AI to people living near a data centre who are deprived of water or electricity that is prioritized for digital infrastructure use.

Obviously, AI did not create today's global industrial mechanisms, but it is driving their development to the full through long supply chains, process offshoring and microfragmentation of jobs, in addition to efforts to universalize sales of some of its products. Nor did AI invent the production practices that some companies use, such as extractivism or intensive lobbying. Such practices create an awkward situation for actors in the digital sector who champion fair and equitable use of AI that respects the fundamental rights of individuals in all places, without worsening inequality, while contributing positively to the collective well-being.

The impact of AI on well-being

This publication attempts to shed light on some of the places where the expansion of AI is deepening inequality by worsening some long-standing areas of discrimination or creating new issues. Given that AI is expanding into ever more areas of our lives, the focus will be on dimensions where its influence is considered to be the most extensive or significant.

In addition to making the traces of AI visible in the various spheres that impact human well-being, these different dimensions convey the complexity of the phenomenon and its effects when considered all together. These effects can be seen not only in the multiple layers of our lives touched by this expanding technology, but also in the intersectional relationships and interactions between many of these layers.

Argentinian economist Sofia Scasserra calls attention to the reproduction of age-old patterns and dynamics in the economic disparities that are created or worsened by the current AI development model. The promise of a new world created through the expansion of AI has not dismantled the historic international division of labour, where the periphery countries are the providers of commodities that have spurred global economic growth and the final consumers of products that are processed in core countries, which profit from the added value. Meanwhile, production for AI development requires a level of technical modernization and investment that is difficult for periphery countries to achieve. As a result, this production, which Scasserra calls "industrial AI" – in reference to the processes to transform massive amounts of data into large models – is an asset that remains under the control of a small group of countries in the Global North. All that is left for the periphery countries is what Scasserra refers to as "artisanal AI": more modest and less lucrative development projects with a highly local reach that are always built upon industrial AI.

The second role of the Global South in this development model is to provide the raw materials necessary for the industrial process. The commodities in question are the data and labour that are stripped largely from the Global South in familiar extractivist and neocolonial patterns. Scasserra uses a very suggestive image that also serves as a clear warning: "We cannot allow the pillaging of our resources for another century. We cannot be another Potosí."

Using another striking image – "It's not a cloud; it's an industrial warehouse" – Ana Valdivia reiterates the importance of narrative in untangling the eco-social impact of this technology's development. Narratives, especially those employed by the tech industry, play a fundamental role when discussing the increasingly enormous environmental impact of AI. By bringing to light the negative effects for the planet and life itself, the narratives that are portrayed as fact must be picked apart and shown for the mere promises they truly are. For example, claims that AI would be a vital ally in fighting climate change have so far resulted only in voracious consumption of natural resources, energy, water and land.

Similarly, such narratives have bolstered an image of immateriality, obfuscating the material reality of the digital infrastructure that underpins AI's computing power. What happens in

between a prompt typed into an AI-based tool and its output is not magic. Instead, it is a series of electronic processes running through thousands of kilometres of fibre-optic cables protected by plastic and metallic coverings or tonnes of metal orbiting the Earth that arrive in huge concrete office buildings filled with servers and processors requiring massive amounts of electricity to run 24/7 along with water to keep them cool, all while local communities compete for the same resources. Estimating the actual amounts of resources consumed can be extremely difficult because most of the companies baulk at providing clear, complete and transparent information.

The current path of today's AI development model clashes directly with the most basic requirement of sustaining life when the complicity between the tech companies and the weapons industry is considered. This conflict of interests is even worse when one considers that it is actually consumer technology that is increasingly being deployed and adapted for military use to increase the lethal capacity of armed forces. Examining the issue from this angle conveys a reality that may seem redundant, but which reflects an unsustainable dehumanization of war. This dehumanization is a result of the proliferation of autonomous weapons and tools that act without the need for human intervention to end human life.
Once again, narratives are deployed in this sinister collaboration between the military–industrial complex and the tech sector to justify an outcome that runs contrary to the principles of sustaining human life. In this case, the argument for this technology revolves around the objectivity and precision offered by an unwavering, unerring robot that analyses and acts. However, the results to date are by no means surgically precise and are rife with calculation errors and consequences from which there is no coming back. It is more difficult to counter another of the main arguments – that of cost-cutting. However, there is a need for nuance, because the cost reductions are purely economic – the human costs are continually rising in the current armed conflicts where these lethal tools are being tested. This complicity is a source of tension for the tech sector in terms of its human dimension. The employees who remember the professional codes of ethics of some of these companies that have said that they had red lines they would not cross when it came to their own research into and development of AI, one of which is the principle that they would never put human life in harm's way.

It is undeniable that generative AI is being widely used in content creation and knowledge production. Here again, the global disparity in technology production and the concentration of these capacities in the hands of a few actors is becoming clear. Most of these actors are corporations in the Global North that are consolidating their wealth and power, adding yet another layer to the dimensions of inequality. Language forms the foundation of the AI models, and this concentration of power and wealth has made English the pre-eminent language in the development of these tools. The warning in this case is clear and resounding: failing to include linguistic and cultural diversity in this process translates to impoverishment of the knowledge that is being produced and creates an obstacle for many communities to see themselves reflected in this new knowledge that will play a major role in the future (and is already doing so now).

However, including cultural and linguistic diversity does not mean simply making these large models available in different languages – which would only be a profit-making move to increase sales – or training the models on content and data produced in different languages. As Pelonomi Moiloa explains, including cultural and linguistic diversity in AI involves developing models from the ground up, based on different linguistic structures so that the entire model reflects the specific characteristics of each context and draws from the richness this diversity has to offer.

Beyond content production, task management (especially labour management) is one of the most extensive functions of tools based on AI or algorithmic mechanisms. Algorithmic management in the platform-based economy has a major impact in the world of work. To understand its reach, the consequences must be made clear. Once again, carefully craf-

ted narratives portray this phenomenon as a way to create new work opportunities or offer positive benefits such as flexibility and autonomy. This narrative obscures the conflicts that have emerged between these working conditions and supposedly established labour rights. Research shows that the opaqueness of the algorithms puts workers in a helpless position as they try to earn a decent living under difficult working conditions and end up forced to accept precarious jobs that run contrary to basic labour rights. Long working hours, having to accept jobs without knowing the pay and arbitrary allocation of jobs go hand in hand with hypersurveillance and threats to privacy. Villarreal and Pérez de la Mora compare this situation to trying to hit the jackpot, as if the workers were playing a game of chance while the algorithm gives them basic jobs to keep them interested and working.

Public services, and more specifically, the administration of social protection tools, have also turned to algorithmic management. Time and time again, fact must be separated from fiction in the narratives being told. These tools are implemented with the argument that they will increase efficiency, trim unnecessary costs and optimize regrettably limited resources. In practice, many of these tools (at least, those that have gained attention for their glitches and serious consequences) are used exclusively for fraud detection. They have been shown to have racist and chauvinistic biases in their reasoning, which places suspicion on the vulnerable individuals who access these social protection mechanisms.

It should be noted that awareness has grown about these types of popular algorithms after problems were detected in how they work following research by civil society organizations. This provides some idea of the opaque climate that crumbled when the consequences of these errors were especially serious. Broadly speaking, the way these mechanisms operate increases the helplessness of the people being managed by these systems and who belong to vulnerable groups – getting help from an actual human is complicated, and often the machine is assumed to be infallible.

Individuals' ability to effectively exercise their rights to participate in the political process also appear to be affected by the features of some AI-based tools that can corrupt democratic processes. Research by various civil society organizations such as AI Forensics has detected risky scenarios for democracy in Europe. Among the different situations are three that appear to have been put to the test. The first is the inaccurate information produced by some chatbots, specifically in electoral contexts, and which amplify narratives that constitute a systemic risk for democracy. The second is the insufficient content-moderation efforts depending on the languages in which such content is posted. And finally, the third is the use of generative AI tools to produce content that boosts fake news to increase its credibility and reach.

Two of the most extensive and deeply entrenched areas of discrimination are race, ethnicity or place of origin on the one hand, and gender on the other. It is important to remember that structural conditions underpin the negative impacts of AI-based tools.

In the case of racism, practices have been identified that are only normalized in situations where there are a majority of racialized people, such as with the use of tools that are practically only used for migration management, because in other contexts the threat to basic rights makes their use unthinkable. In other cases, real-world use of applications with unproven effectiveness or with recurring problems has been especially harmful for racialized people – as in the use of algorithmic mechanisms for security or public service management, and especially with social protection tools, where people see a direct impact on their ability to exercise their rights. Additionally, the political economy that dictates the AI life cycle, from the extraction of critical materials to the implementation of tools, has been marked by the racist history of the disparities between the Global South and Global North. And in the case of gender discrimination, generative AI and the algorithmic systems used by social media increase the spread of gender-based violence, misogynistic content, hate and stereotypes against women and LGBTQIA+ communities. Research such as that by Afef

Abrougui shows how these circumstances make social and political participation challenging in more hostile contexts and hamper debate on sexual and reproductive rights, which promotes continued discrimination.

Similarly, a gender-based impact in the workplace has been detected, where AI tends to take over jobs that are typically done by women (such as administrative positions), thus creating economic uncertainty for women and reinforcing the traditional gender-based division of labour. In the political sphere, there have been cases of sexual violence with deepfakes involving politicians, some of whom have even been forced to end their careers. The effects are worse for women in situations where hypersurveillance is used to support smart cities and strengthen other population control measures. The same occurs in cases of armed conflict, where the use of AI-based tools is especially felt by women, children and LGBTQIA+ people and surveillance tools can have devastating effects.

Multidimensionality and intersectionality

Although each chapter of this book has focused on a specific dimension, most of the analyses and assessments deal with the multidimensionality and intersectionality of different layers, the relationships between these areas and how many of them mutually reinforce, amplify or complement each other, and how they interact to show the complexity of the phenomenon. A simplistic or superficial approach is incapable of providing any meaningful insights into the impact of AI in the lives of individuals and their well-being. Determining the influence of AI in the various areas of discrimination (those that are specifically mentioned here and many others) and inequality requires time and thorough analysis.

These chapters make clear, for example, that the negative impact of AI-based tools for migrant women in Europe involves the multiple dimensions and intersection of race and gender, along with threats to democracy, which goes back to the contribution to the weapons industry and possibly social protection mechanisms. Obviously, AI could have positive impacts as well, but these do not worsen inequality and are not subject to the same concern. Similarly, the driver of a rideshare platform in the suburbs of Mexico City could experience the intersectionality of the dimensions related to decent work and algorithmic management of labour, as well as the environmental impact of the data centres being built nearby and global economic disparities. The list goes on. It could be interesting to imagine random profiles around the world and consider which of these dimensions of the impact of AI affect their lives, from a civilian in Ukraine to a farmer in South-East Asia or a musician who raps in a national language in Africa.

There are also areas of inequality that this document does not cover in depth. For example, the impact of AI on human bodies has not been covered in detail through a non-ableist, queer, feminist, anti-colonial and anti-racism lens, although aspects of these relationships are touched upon throughout the text.

What if AI followed a different development model?

The ideas about the different dimensions of AI's impact on individual well-being shared throughout this book highlight the threats and negative effects of the technology, but do not seek to change the technology itself. Efforts have been made to suggest alternatives and solutions that make one thing clear: the problem is not AI, but its specific development model that has become hegemonic.

Most of the suggestions call into question the principles that widen the gaps in inequality, but the researchers and experts provide ideas throughout the chapters to fix the problems with the model.

Some of these ideas include leveraging the regionalism that Latin America has drawn from to address other challenges and the tradition of protecting common goods as areas based on which more egalitarian governance mechanisms could be established. Given the evidence that the current model is environmentally unsustainable, one suggestion is to critically examine the usefulness of algorithmic systems – in other words, determine which of the algorithms and AI-based tools offer a genuine benefit for human life compared to their high costs. This would be a more or less immediate solution until the technology is addressed from a more comprehensive perspective that places human life and sustainability ahead of profits.

Regarding other threats, suggestions include strengthening and improving regulatory mechanisms, and more specifically, making them more efficient given the flexibility AI tools have shown to fly under the radar and evade control. With this in mind, the authors agree on the need to improve regulation to set up guardrails for knowledge transfer and interactions between the tech and weapons industries, as well as to guarantee that international humanitarian law covers new modalities of armed conflict. The situation is similar with labour rights, where regulation is needed to ensure that gains are not lost and respond to the threats stemming from deliberate attempts to overhaul the labour market.

With regard to threats to democracy, there appear to be sufficient mechanisms to formally guarantee its protection, but the facts show that there are some doubts about their effectiveness or capacity for application. The researchers identified the need for improved vigilance and greater leeway for intervention in cases where abuses have occurred with these tools. With going into further detail, they caution that while the European Union's Digital Services Act (DSA) requires large platforms and search engines to assess the risks associated with their services, they do not always do so.

Community capacity-building and promoting participation also appears to be a counterweight to some of the negative impacts. Introducing linguistic diversity would also be an ideal alternative, and as Moiloa notes, it would open the door to a paradigm that benefits not only those who speak underrepresented languages but also global technological ecosystems by providing diverse and sustainable models for the future. Something similar occurs in the fight against racial discrimination. Communities must have the skills they need to produce tools that are suited to their realities as well as the resources to do so, which requires redistributing resources and eliminating the monopolized access to financing held by a handful of actors. With regard to the threat from the complicity between the tech and military industries, communities also have a key role to play. In this case, the communities of workers could become stronger to act as the first level of containment of malicious use.

Finally, algorithmic transparency is also presented as a frequent antidote to the very real threats: transparency in the relations between the tech and weapons industries, in the algorithmic management of work, and in the tools that manage social protection.

The researchers have put forward a formula that includes creativity to change the paradigm and imagine new ways of governing; regulation to guarantee rights are respected and ensure that the general public do not lose control of the technology to special interests; participation to diversify perspectives and foster respectful development that meets people's different needs; and transparency to ensure continuous citizen oversight and build a model that is sustainable over time.