AI & EQUALITY

Re-visions of Now and Future III

Community Publication #3

January Term 2025

This volume brings together nineteen essays written by participants of the AI & Equality Human Rights online course, held during the 2025 January J-Term.

AI & EQUALITY
A WOMEN AT THE TABLE INITIATIVE

Table of contents

AI Governance & Procurement

AI, Labor, and Responsibility

Biographies of Contributors

Introduction

Welcome to the third edition of ***Re-Visions of Now and Future***, an AI & Equality community publication. This volume brings together nineteen essays written by participants of the AI & Equality Human Rights online course, held during the January J-Term 2025. This free, open-access course invited thinkers, practitioners, and activists from across disciplines to interrogate the ways AI systems entrench or challenge existing inequalities. Through rigorous engagement with human rights frameworks and critical questions of power, fairness, accountability and governance, participants produced essays that push beyond mainstream AI discourse to address its structural, political, and deeply human dimensions.

The response to the course was striking—138 registrations from 48 countries spanning North America, South America, Europe, Africa, and Asia. The participants came from diverse academic and professional backgrounds, with social sciences and international relations/political science as dominant disciplines. While many were engaged in formal education at the undergraduate, master's, or PhD levels, others brought extensive professional expertise, ensuring that the discussions were shaped by a mix of scholarly research and real-world experience. This convergence of perspectives is reflected in the essays, which do not merely analyze AI's impact in abstract terms but interrogate its role in reinforcing or dismantling structural inequalities.

Participants wrestled with foundational questions: How do AI systems perpetuate historical patterns of discrimination? Can data feminism provide a radical framework for resisting algorithmic oppression? What forms of governance and procurement can ensure ethical AI adoption, particularly in humanitarian and development contexts? How does AI reshape labor, human rights, and personal agency? The essays in this collection explore these issues through themes of privacy, surveillance, bias, labor, and governance—challenging dominant narratives and offering visions for more just AI futures.

This volume critically engages with AI's role in shaping power, inequality, and justice across multiple domains. The nineteen essays in this edition take a variety of forms, reflecting the diverse backgrounds, experiences, and analytical approaches of their authors. Some are extensive research papers while others are short reflective pieces that bring personal experience, activism, or professional insight into conversation with the course's themes.

Several essays interrogate the ethical dilemmas of machine learning, from the potential of large language models to promote fairness in humanitarian contexts to the pressing need for machine unlearning as a human rights imperative. Others expose the privacy and data collection risks embedded in AI-driven recommendation systems, highlighting how opaque algorithms disproportionately impact marginalized communities. Across these discussions, the authors emphasize the urgency of accountability, transparency, and ethical AI development, resisting the pervasive assumption that technological progress is inherently neutral or beneficial.

A strong feminist critique runs through many contributions, focusing on AI's entanglement with gendered oppression and digital inequalities. Authors explore the potential of data feminism as both a theoretical lens and a practical tool to combat tech-facilitated gender-based violence, biased content moderation, and systemic discrimination in algorithmic design. Other essays shift the focus to AI's governance, procurement, and labor implications, exposing how AI reshapes global work structures—often at the expense of workers in the Global South. From the ethical responsibilities of NGOs adopting AI to the erosion of human agency under algorithmic management. Taken together, the collection forms a powerful call for resistance, transformation, and justice in AI.

— Amina Soulimani, Editor

AI & EQUALITY

<AI & Equality> is a community committed to establishing and promoting a human rights-based approach to AI that centers equality & inclusion at the core of the code.

All authors are members of our online community. This open-to-anyone global community aims to connect individuals from all backgrounds, regions, and disciplines to work towards a collective goal of AI and emerging technologies based on human rights principles and the dignity of all.

We continue to provide a platform and space for sharing the thoughts, ideas, and work of our community through conversation and publication.

Join us at community.aiequalitytoolbox.com.

AI & EQUALITY

LLMS, Machine Learning, and Privacy

Utilizing Large Language Models to Enhance Fairness and Accountability to Affected Populations in Humanitarian Protection Programs

Sezen Yalcin, Turkey
A humanitarian and development professional with expertise in child protection, community-based protection, and gender-based violence in emergency settings.

The rise of artificial intelligence in humanitarian action offers great potential to improve the quality, efficiency, and accountability of crisis responses. Large Language Models (LLMs), in particular, have shown significant promise in enhancing various areas of humanitarian programming. On the other hand, it's crucial to note that the success of these technologies relies heavily on the use of transparent, complete, and bias-free datasets (UNHCR 2022).

Aid agencies are now rapidly developing AI applications to increase the reach and speed of their responses and improve the overall quality of their support. While researching the AI models developed to enhance the efficiency of humanitarian assistance, I came across several AI-supported applications designed to enhance service delivery in the humanitarian sector, such as Signpost, AI Chatbots for MHPSS**,** Displacement Tracking Matrix**,** and U-Report chatbot. Despite these advancements, there appears to be a gap: no AI application focuses exclusively on Accountability to Affected Populations (AAP). AAP is a key component of human-rights based humanitarian action and it is agreed as a common standard as emphasized by the Core Humanitarian Standard on Quality and Accountability (CHS). It is a cornerstone of principled humanitarian action, ensuring that affected communities are meaningfully engaged, their feedback is collected and utilized, and programming is adapted accordingly. To prevent humanitarian aid from being used for political purposes, and to ensure that it is carried out solely based on humanitarian needs, it is crucial that humanitarian aid is conducted in line with the principle of accountability (The HAP and humanitarian accountability, 2023).

Developing an AI-driven tool dedicated to AAP could address this unmet need, enhancing inclusivity, trust, and the relevance of humanitarian responses. Such a tool could facilitate

accessible and anonymous feedback channels, enable real-time data analysis for actionable programming ideas, and ensure that affected populations are at the center of decision-making processes. This paper explores the potential of LLMs to enhance AAP in humanitarian programming, focusing its limitations and ethical considerations with a fairness perspective.

AAP is a fundamental principle that ensures humanitarian efforts are people-centered, inclusive, and dignified. It emphasizes the agency of affected communities in accessing essential rights and services, thereby fostering participation and trust. However, collecting feedback and complaints from vulnerable populations poses significant challenges due to the sensitivity of issues raised, fear of retaliation, or concerns that feedback may negatively affect access to aid.

LLMs can offer innovative solutions to these challenges by enabling anonymous and efficient communication channels. LLM supported chatbots can create safe and confidential platforms for beneficiaries to share their experiences, grievances and suggestions without fear of retaliation. Such systems promote open dialogue, enhance trust, and strengthen the feedback loop, ultimately reinforcing the principles of AAP (UNHCR 2020). By ensuring that affected populations have a meaningful voice in shaping interventions, these technologies contribute to more accountable and responsive humanitarian programming.

One of the primary challenges in implementing AI-driven feedback systems lies in ensuring fairness, a fundamental principle of humanitarian action rooted in non-discrimination. Critical considerations include determining who will have access to these AAP mechanisms and whose feedback on humanitarian programs will be prioritized. Addressing these questions is essential to prevent the exclusion of marginalized voices and ensure equitable representation. Another vital aspect of AAP is that collected feedback and complaints must effectively inform program implementation. For this reason, the role of LLMs becomes particularly significant. How the model filter, analyze, and prioritize the data they gather is influenced not only by how it is trained but also by the contextual nuances of the humanitarian setting. Determining whose feedback is deemed actionable and how it shapes programmatic decisions are complex, context-specific challenges that underscore the importance of ethical and inclusive AI design.

To address the aforementioned considerations, the AAP model must be trained on localized data to capture the specific cultural, social, and contextual features of the affected populations. This approach ensures that the model's recommendations and feedback processing accurately reflect the needs and realities on the ground, thereby enhancing its relevance and effectiveness. Equally important is that beneficiaries perceive the model as a trustworthy and reliable interlocutor. To foster this trust, the model should prioritize

empathetic and culturally sensitive responses, encouraging open and honest feedback. This approach helps to avoid perceptions of tokenism or the impression that the process is merely a "tick-the-box" exercise, thereby reinforcing the credibility and impact of the feedback mechanism.

I will draw on insights from the *Defining Fairness & Algorithmic Fairness Metrics* discussions we conducted during the AI & Equality course. To ensure that an LLM-supported AAP model effectively meets the needs and requirements of affected populations in an equitable manner, a group fairness metrics approach seems to be suitable to adopt. Group fairness is grounded in the principle that distinct demographic groups should be treated equitably. For example, in the context of a refugee crisis resulting from conflict, the AAP model must be capable of understanding, classifying without bias, and deriving actionable programmatic improvements from the feedback of refugees, internally displaced persons (IDPs), and affected host communities. To design an AAP system aligned with the needs of affected groups, the model must adhere to the principles of group fairness metrics.

Nevertheless, it is important to recognize the limitations of relying exclusively on group-level metrics, which may overlook the diverse needs within these groups. If a model is trained primarily on the general needs of a specific group—such as refugees, IDPs and host community members—it may fail to address the concerns of marginalized subgroups within that population. These could include transgender individuals, people living with HIV/AIDS, persons with disabilities, or members of minority ethnic groups. Such omissions risk leaving critical and context-specific needs unaddressed. Conversely, an individual fairness perspective emphasizes recognizing the unique characteristics and experiences of individuals within a dataset. While this approach ensures that decisions account for personal experiences, it risks disregarding the collective experiences stemming from group membership. For instance, in the case of refugees, IDPs, or host communities, individual fairness may fail to incorporate the shared experiences and community affiliations that influence beneficiaries' feedback on humanitarian programs. Balancing these two approaches is therefore essential to create a robust and inclusive AAP model.

The integration of LLMs into humanitarian protection programming offers significant opportunities to enhance AAP. To ensure that an AAP model aligns with principles of fairness and human rights, several key considerations must guide its development. First, datasets used to train the model must represent diverse perspectives, including marginalized subgroups within affected populations. This inclusivity is crucial to avoid reinforcing existing inequities and to ensure the model delivers equitable outcomes. Additionally, LLMs should be tailored to the specific cultural, linguistic, and social contexts of humanitarian crises, enhancing their relevance and effectiveness in addressing the unique needs of affected communities.

Fairness metrics must be applied throughout the AI lifecycle to evaluate outcomes at both group and individual levels, promoting equity while addressing the diverse needs of different demographic groups. Furthermore, the active involvement of affected populations in the design and evaluation of LLM-based tools is essential. This participatory approach ensures that the tools are responsive to community needs, fosters trust, and strengthens the credibility of the model. By integrating these principles into the development process, AAP models can more effectively uphold fairness, inclusivity, and the core values of rights-based humanitarian action.

References

Core Humanitarian Standard (n.d.). *Core Humanitarian Standard on Quality and Accountability*. Available at: https://www.corehumanitarianstandard.org

United Nations High Commissioner for Refugees (2022). *AI Applications in Humanitarian Response: Enhancing Efficiency and Accountability*. Geneva

United Nations High Commissioner for Refugees. *Accountability to Affected People Operational Guidance Toolkit*. September 2020. Geneva

The HAP and humanitarian accountability, Agnès Callamard, May 2003
https://odihpn.org/magazine/the-hap-and-humanitarian-accountability/

Machine Unlearning: A Human Rights Imperative

Duarte Silva, Portugal

A final-year Master's student in Mathematics and Computer Science. Duarte holds a Bachelor in Applied Statistics.

Artificial Intelligence (AI) has the potential to revolutionize various sectors, from healthcare to finance. However, as we navigate the promises of AI, it is crucial to address its ethical implications, particularly concerning human rights. One of the emerging concepts in AI governance is Machine Unlearning, a technique aimed at selectively removing data from trained models to comply with privacy rights, correct biases, or mitigate harm. This concept is especially relevant in the context of human rights and AI ethics, as explored throughout the course.

This essay examines Machine Unlearning (MU) through the lens of the Universal Declaration of Human Rights (UDHR) and AI fairness frameworks. It highlights how the inability to erase or rectify learned biases can perpetuate discrimination, undermine privacy, and reinforce systemic injustices. Drawing from course materials, I will briefly explore how Machine Unlearning can be a necessary tool to align AI with human rights principles.

The right to privacy (Article 12 of the UDHR) and the right to non-discrimination (Article 2) are two core human rights that can be compromised by AI systems. Traditional AI models, once trained, retain learned information indefinitely, making it difficult to remove specific data points when necessary. This becomes problematic in cases of biased training data, inaccurate profiling, and consent withdrawal under privacy laws like the GDPR.

Machine Unlearning provides a solution by enabling AI models to forget specific data, reducing the risk of harm. It can be particularly impactful in scenarios such as:

1. **Data Privacy and the Right to Be Forgotten**: Many jurisdictions, particularly the European Union's General Data Protection Regulation (GDPR), enshrine the right to request the removal of personal data from databases. However, if an AI model has already been trained on this data, deletion from the raw dataset is insufficient. MU techniques allow AI models to comply with legal and ethical mandates effectively.

2. **Mitigating Algorithmic Bias**: Throughout the course, we examined bias entry points in the AI lifecycle. Biases embedded in training data can lead to unfair outcomes in

hiring systems, credit assessments, and predictive policing. When biases are detected post-deployment, Machine Unlearning provides a mechanism to correct these injustices by removing problematic data or adjusting learned representations without requiring full retraining. Several cases illustrate the dangers of biased AI systems:

- **Mortgage Approval Bias**: Martinez reported that AI-driven mortgage approval systems unfairly denied loans to Black and Latino applicants at higher rates than white applicants, despite similar financial backgrounds. Machine Unlearning could help address such disparities by removing biased data points from training sets;
- **Housing Discrimination**: Block highlighted how biased algorithms reinforced racial and socioeconomic barriers in housing applications, leading to systematic exclusions. Machine Unlearning offers a corrective measure by identifying and removing discriminatory patterns;
- **Algorithmic Scandals in Policymaking**: the Dutch government child benefits scandal exemplifies the severe consequences of algorithmic bias, where flawed automated systems falsely accused thousands of families, many from minority backgrounds, of fraud. Such cases underscore the urgent need for Machine Unlearning to rectify injustices post hoc;
- **Criminal Justice Disparities**: Zilka and others explored how risk assessment tools used in criminal justice perpetuate racial disparities due to biases in historical enforcement data. Machine Unlearning presents an opportunity to mitigate such biases by eliminating discriminatory patterns from AI decision-making processes.

3. **Protecting Vulnerable Groups**: AI systems that perpetuate discriminatory outcomes can disproportionately harm marginalized communities. If an AI-driven hiring tool systematically excludes candidates based on gender or ethnicity due to biased historical data, Machine Unlearning can help remove these biases without requiring an entirely new model.

Practical Applications and Challenges

While Machine Unlearning looks promising, it also presents technical and ethical challenges. One major challenge is ensuring certifiability, verifying that a model has truly forgotten the specified data. Techniques such as Sharded, Isolated, Sliced and Aggregated Training (SISA) or Amnesiac Unlearning are emerging solutions that offer verifiable deletion. However, these methods are computationally expensive and may not always be practical for large-scale AI systems. Moreover, decisions on what to forget create ethical dilemmas. Who decides which data should be removed? What if removing certain data introduces new biases? These are

critical questions that require interdisciplinary collaboration between AI practitioners, policymakers, philosophers and human rights experts.

Despite being a relatively new concept - introduced a decade ago by Cao & Yang - the field has seen rapid advancements. Recent research has expanded its applicability beyond classification tasks to large language models (LLMs) and fairness-oriented AI:

- **Safe-CLIP** (Popi et al., 2024): demonstrates how Machine Unlearning can successfully remove NSFW concepts from vision-and-language models;
- **The WMDP Benchmark** (Li et al., 2024): Explores methods for reducing the risks of large language models (LLMs) empowering malicious actors in developing biological, cyber, and chemical weapons;
- **Fair Machine Unlearning** (Oerstling et al., 2024): Investigates techniques for data removal that simultaneously mitigate disparities in AI outcomes.

For a broader understanding of the challenges ahead, Barez and fellow researchers outline open problems in Machine Unlearning for AI safety, while Nguyen et al. provide a comprehensive survey of the state of the art in this evolving domain.

Machine Unlearning is not merely a technical solution; it is a human rights necessity. As AI continues to shape our societies, ensuring that models can forget harmful, biased, or privacy-sensitive information is fundamental to creating ethical AI systems. Throughout this course, we explored how AI can both infringe upon and uphold human rights. Machine Unlearning serves as a tangible step toward aligning AI development with the principles of fairness, privacy, and accountability.

By fostering interdisciplinary collaboration and integrating these principles into AI governance, we can create more just and equitable AI systems. As we move forward, the question is not whether Machine Unlearning should be implemented, but how we can ensure it is done effectively, ethically, and equitably.

References

Barez, F., Fu, T. et al. (2025) 'Open Problems in Machine Unlearning for AI Safety', *arXiv*. Available at: https://arxiv.org/abs/2501.04952.

Block, Bill (2022) 'How biased algorithms create barriers to housing', *ACLU Washington*. Available at: How biased algorithms create barriers to housing.

Bourtoule, L., Chandrasekaran, V. et al. (2022) 'Machine Unlearning', *arXiv*. Available at: https://arxiv.org/abs/2209.02299.

Cao, Y., Yang, J. et al. (2015) 'Towards Making Systems Forget with Machine Unlearning', *IEEE*. Available at: https://ieeexplore.ieee.org/document/7163042.

Heikkilä, Melissa (2022) 'Dutch scandal serves as a warning for Europe over risks of using algorithms', *Politico*. Available at: Dutch scandal serves as a warning for Europe over risks of using algorithms.

Graves, L., Nagisetty, V. and Ganesh, V. (2020) 'Amnesiac Machine Learning', *arXiv*. Available at: https://arxiv.org/abs/2010.10981.

Li, N., Pan, A. et al. (2024) 'The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning', *arXiv*. Available at: https://arxiv.org/abs/2403.03218.

Martinez, E., Kirchner, L. (2021) 'The secret bias hidden in mortgage approval algorithms'. *The Markup*. Available at: The Secret Bias Hidden in Mortgage-Approval Algorithms.

Nguyen, T.T., Huynh, T.T. et al. (2024) 'A survey of machine unlearning', *arXiv*. Available at: https://arxiv.org/abs/2209.02299.

Oesterling, A., Ma, J. et al. (2023) 'Fair Machine Unlearning: Data Removal while Mitigating Disparities', *arXiv*. Available at: https://arxiv.org/abs/2307.14754.

Poppi, S., Poppi, T. et al. (2023) 'Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models', *arXiv*. Available at: https://arxiv.org/abs/2311.16254.

Zilka, M., Fogliato, R. et al. (2023) 'The progression of disparities within the criminal justice system: Differential enforcement and risk assessment instruments', *arXiv*. Available at: https://arxiv.org/abs/2305.07575.

Navigating the Ethical Landscape: Privacy and Data Collection Concerns in recommendation systems

Sadia Tabassum, Bangladesh
UMSAILS Scholar with a Master of Laws (LL.M) from the University of Asia Pacific, Dhaka, in affiliation with the UNESCO Madanjeet Singh South Asian Institute of Advanced Legal and Human Rights Studies (UMSAILS). Sadia also holds a Bachelor of Laws (LL.B. Hons.)

The swift progress of artificial intelligence (AI) and recommendation systems (RS) has raised significant ethical and privacy issues. While these technologies provide tailored content and improve user experiences, they also bring serious threats to privacy, autonomy, and basic human rights. This essay examines the ethical dilemmas linked to AI-driven recommendation systems, concentrating on the challenges of transparency, bias, privacy breaches, and the diminishing of individual autonomy.

Recommendation systems can exhibit biases stemming from the training data and the biases of their developers, which may result in discrimination. Profiling techniques that analyze personal data and categorize individuals based on their behavior can reinforce existing social inequalities (Milano, 2020). For example, if an AI system is trained on biased data, it might produce recommendations that unfairly disadvantage specific demographic groups. This concern is heightened when sensitive personal information, such as race, gender, and socioeconomic status, is involved, as it can lead to unjust treatment and perpetuate current social disparities (Noble, 2018).

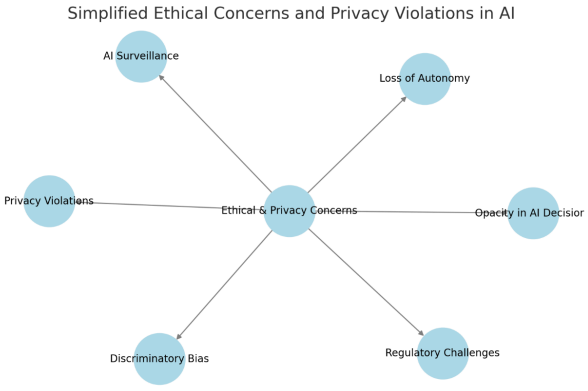


Fig 1: Simplified figure showing ethical concerns and privacy violation in AI

Opacity in System Decisions

AI systems, especially those utilizing recommendation algorithms, frequently function as "black boxes," meaning their decision-making processes are not transparent, even to their creators. This lack of transparency raises significant ethical issues, as it conceals the underlying values and biases within these systems, making it challenging to ensure they adhere to ethical standards and respect fundamental rights (Milano, 2020). The ambiguity surrounding how these algorithms operate hinders users from understanding the reasons behind specific recommendations, limiting their ability to contest or question automated decisions. This problem is particularly pertinent in social contexts where AI learns from user interactions, resulting in unpredictable outcomes that are hard to audit (Pasquale, 2015).

Privacy, Data Protection Violations and the erosion of Human Autonomy

AI-driven recommendation systems gather and process large volumes of personal data, which raises serious privacy concerns. These systems often deduce highly sensitive information, such as emotional states, religious beliefs, and political preferences, through automated profiling. In fields like healthcare and finance, where data sensitivity is particularly high, the risk of privacy violations becomes even more pressing. Additionally, many users remain unaware of the extent to which their data is being collected and analysed (Zuboff, 2019).

One concerning example is the use of AI to predict a person's gender, job, and future movements based on cell phone data. Researchers have shown how machine learning (ML) models can analyze publicly available information to infer private details, often without users' explicit consent. This level of data processing infringes on the fundamental right to privacy and highlights the urgent need for strict data protection regulations (Crawford & Schultz, 2014).

Recommendation systems shape users' choices by curating content based on their assumed preferences. While this personalization can improve user experience, it also raises ethical questions about autonomy and self-determination. Users frequently do not have a real chance to make informed decisions due to the manipulative nature of algorithm-driven content delivery. The consent mechanism for data collection is often flawed, as users are typically pressured to agree to data processing in return for access to services. Moreover, AI-driven profiling impacts users' personal identities by influencing their online experiences. Many recommendation systems use collaborative filtering, which creates collective rather than individual profiles. The lack of transparency in profiling further complicates this issue, as users remain unaware of how their data is being used to shape their digital experiences (Dwork & Mulligan, 2013). This can lead to misclassification, where users are grouped in ways that do not reflect their actual social attributes or self-identification.

AI-Powered Surveillance and the End of Anonymity

The rise of AI-powered surveillance has resulted in significant privacy infringements. Both governments and corporations are increasingly utilizing AI tools to monitor individuals, often without their knowledge or consent. For instance, facial recognition technology is being used in public areas, diminishing anonymity and fostering a sense of perpetual surveillance. In China, the rollout of centralized CCTV systems equipped with facial recognition has sparked worries about authoritarian governance and widespread monitoring. Likewise, in the United States, almost half of all adults are now part of law enforcement facial recognition databases (Smith, 2020).

The OECD principles regarding AI and human rights stress the importance of developing privacy-preserving technologies and establishing safeguards against mass surveillance (OECD, 2019). Due to extensive implementation of AI surveillance disproportionately impacts marginalized groups, who frequently become the main targets of these monitoring initiatives. The anxiety of being observed can stifle freedoms of expression and association, resulting in a chilling effect on democratic engagement. Moreover, the rationale for AI surveillance, framed as a means of preventing crime and ensuring public safety, does not sufficiently justify the infringement on the fundamental right to privacy.

Legal and Ethical Implications

Privacy is acknowledged as a fundamental human right in various legal frameworks, including the Universal Declaration of Human Rights (UDHR) and the General Data Protection Regulation (GDPR). Data protection laws assert that individuals should maintain control over their personal data, and organizations are required to uphold transparency and accountability in their data handling practices. However, the swift integration of AI technologies has outstripped regulatory responses, resulting in enforcement and protection gaps (European Commission, 2021). It is essential for governments and organizations to establish more rigorous oversight mechanisms to ensure that AI systems adhere to ethical standards and human rights principles. This includes improving transparency and accountability in AI operations.

The ethical issues surrounding AI-driven recommendation systems and surveillance technologies emphasize the pressing need for enhanced data protection and privacy laws. The lack of transparency in AI systems, along with biases, privacy infringements, and threats to personal autonomy, highlights the difficulties in ensuring ethical AI use. As AI technology progresses, it is crucial for all stakeholders to focus on ethical principles and the protection of human rights to avoid the misuse of personal information and to uphold individual freedoms. Tackling these challenges calls for a joint effort among governments, technology firms, and civil society to establish a more transparent and responsible AI environment.

References

Crawford, K., & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review, 55*(1), 93-128.

Dwork, C., & Mulligan, D. K. (2013). It's Not Privacy, and It's Not Fair. *Stanford Law Review Online, 66*, 35-40.

European Commission. (2021). Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence. Retrieved from https://ec.europa.eu

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society. *Mind & Society, 17*(1), 1-33.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Milano, S., Taddeo, M. and Floridi, L. (2020) Recommender systems and their ethical challenges - ai & society, SpringerLink. Available at: https://link.springer.com/article/10.1007/s00146-020-00950-y#citeas \

OECD. (2019). Recommendation of the Council on Artificial Intelligence. Retrieved from https://www.oecd.org

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

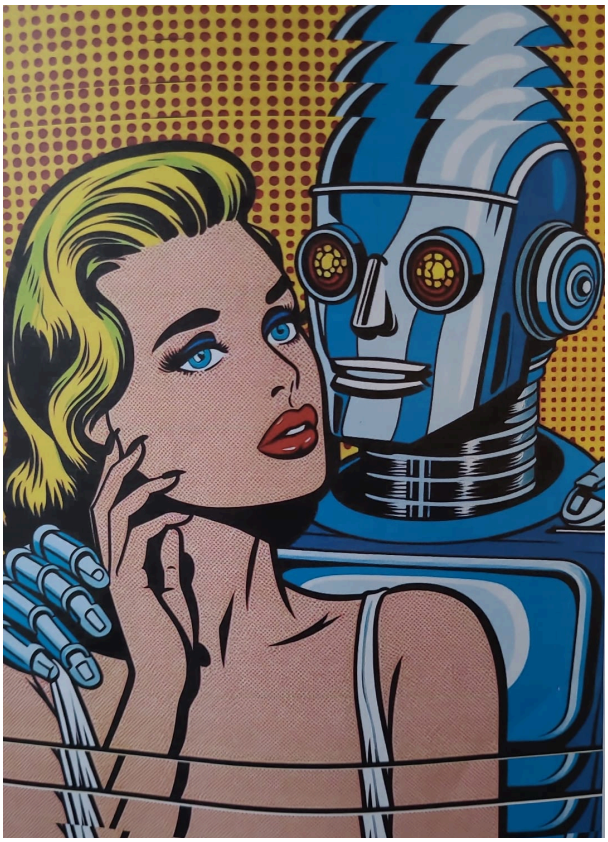
Smith, A. (2020). The Rise of AI Surveillance. *Journal of Privacy and Security, 12*(4), 45-62.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Data Feminism and Gender Equality

The Master's Techniques on how to Dismantle AI Suppression: A Critical Examination of AI-Human communications

Freyja van den boom, The Netherlands
A transdisciplinary practice-based postdoctoral researcher and speculative socio-legal designer. They hold a PhD from Bournemouth University.



Audre Lorde's powerful assertion that "the master's tools will never dismantle the master's house" serves as an urgent reminder of the limitations inherent in existing systems of power and our ways to address the harms from the consolidation of power we see happening in AI.

As artificial intelligence (AI) technologies, particularly large language models (LLMs), become increasingly integrated into our daily lives, it is crucial to critically examine their impact on marginalized communities, especially women. A notable example of AI's potential to cause harm is the case of an AI chatbot that, when asked about women's rights, provided responses that were not only dismissive but also perpetuated harmful stereotypes. Such incidents highlight the urgent need for a deeper understanding of the risks and harms associated with AI-human interactions.

This article aims to introduce Berit Ås's seven Male Master Suppression Techniques as a lens through which we can evaluate human-LLM interactions. By identifying these suppression techniques, we can better understand the specific harms women face in AI contexts and explore ways to mitigate them. Furthermore, we will discuss the role of speculative design in challenging dominant narratives and advocating for transformative change, ensuring that we are not just involved but empowered with the knowledge we need in order to make well informed decisions about the development, deployment, and use of AI technologies that increasingly affects every aspect of our lives.

Why do we need to provoke attention?
Mansplaining neural networks and pink tax recommender systems

The integration of AI systems into various aspects of life presents significant risks, particularly concerning power consolidation and the lack of female representation in AI development. Research has shown that AI systems can perpetuate existing biases, leading to harmful outcomes for women. For instance, biased hiring algorithms have been documented to favor male candidates, while voice assistants often respond to female users in condescending ways (Duffy, 2021). The absence of women in AI development teams exacerbates these issues, as the perspectives and experiences of half the population are often overlooked (West, Whittaker and Crawford, 2019).

Regular audits of AI systems can help to identify and rectify biases in algorithms and datasets (Raji and Buolamwini, 2019). There are other ways by which we may be able to ensure that the benefits outweigh the risks: ensuring that AI development teams are diverse and inclusive to better represent the needs and experiences of all users (West et al., 2019); implementing of transparency mechanisms that allow users to understand how AI systems make decisions, and hold developers accountable for any biased outcomes (Binns, 2018).

We adopt speculative design as a proven method to critique dominant narratives and envision alternative futures. By imagining scenarios where women are not only included but empowered in AI development, we can challenge the status quo and advocate for transformative change. This approach encourages critical discussions about the implications

of AI technologies and the need for inclusive practices that prioritize the voices of marginalized communities.

We developed a set of artefacts including the ***AI discussion cards*** based on the suppression techniques and co-designed a series of *AIFutures zines*. Zines in particular have a rich history in feminist activism, as grassroots tools for marginalized voices to share their experiences, ideas, and critiques of dominant narratives. Our aim with this approach is to help people be aware of the risks while engaging with AI. The aim is not to prevent people from using AI but to benefit from its potential to improve their lives in responsible and equitable ways.

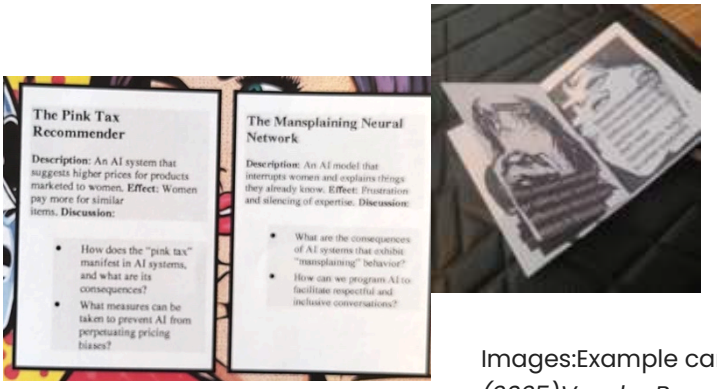
Here we will briefly introduce our project using speculative design and the master suppression techniques as a novel lens to assess the limitations of AI systems and encouraging critical engagement with AI-generated content.

Mastering the AI Suppression Techniques

Berit Ås, a Norwegian social psychologist and feminist, identified seven Male Master Suppression Techniques that illustrate how power dynamics operate in social interactions. These techniques are:

1. *Making Invisible: Ignoring or excluding women's contributions.*
2. *Ridicule: Mocking or belittling women's opinions and experiences.*
3. *Withholding Information: Deliberately keeping relevant information from women.*
4. *Double Binding: Placing women in situations where any action they take will be criticized.*
5. *Heaping Blame and Putting to Shame: Blaming women for mistakes or failures.*
6. *Objectification: Treating women as objects rather than individuals with agency.*
7. *Threat of Force: Using intimidation or coercion to control women's behavior.*

These techniques provide a useful framework for analysing how AI systems may reinforce existing gender inequalities and contribute to the marginalization of women. Using the seven techniques we mapped the risks from LLM's and developed a set of scenarios to illustrate how they may manifest in our interactions. We developed a set of cards and a zine (available upon request) to provoke discussion about the impact of LLM's, the way they are designed to communicate and what are the known and unknown consequences.



Images:Example cards and the *To Master AI- zine (2025)VandenBoom*

Adaptation to AI:

These techniques can be applied to human-AI interactions, particularly focusing on interactions between LLMs and women. For instance, LLMs may inadvertently reinforce these suppression techniques through biased training data and lack of diverse representation in their development.

1. **Making Invisible:**
   An AI assistant consistently ignores a woman's commands while promptly responding to a male user's requests.
   For example: a virtual assistant may fail to recognize a woman's voice or input, leading to frustration and feelings of invisibility.
   Countermeasure: Implement algorithms that ensure diverse voices are recognized and value

2. **Ridicule:**
   A chatbot responds to a woman's inquiry with sarcasm or belittling remarks.
   For example: a scenario where a woman asks for help with a technical issue, and the AI responds with a mocking tone, undermining her confidence.
   Countermeasure: Train AI systems using diverse datasets that respect and uphold the dignity of all users.

3. **Withholding Information**:
   An AI system provides incomplete or biased information to women.
   For instance, a health-related chatbot may offer less comprehensive advice to female users, perpetuating health disparities.
   ***Countermeasure**: Develop transparency mechanisms in AI models to ensure equitable access to information.*

4. **Double Binding**:
   An AI-driven hiring tool places women in a no-win situation, penalizing them for traits that are considered positive in men.
   For example, a woman may be criticized for being assertive in an interview, while the same behavior is praised in male candidates.
   Countermeasure: *Regularly audit AI systems for bias and involve diverse stakeholders in the development process.*

5. **Heaping Blame and Putting to Shame**:
   A virtual assistant blames a woman for errors in task performance, despite the AI's own limitations.
   For example a woman is held accountable for a scheduling error caused by the AI.
   Countermeasure: *Design AI systems to provide constructive and unbiased feedback.*

6. **Objectification**:
   An AI system uses gendered language that reduces women to stereotypical roles.
   For example, a customer service chatbot may refer to female users as "sweetie" or "dear," reinforcing traditional gender norms.
   Countermeasure: *Implement checks to ensure AI-generated language is neutral and respectful.*

7. **Threat of Force**:
   An AI-powered security system uses intimidating language or actions towards women. For instance, a smart home security system might issue warnings in a tone that is overly aggressive or patronizing, making women feel unsafe rather than protected.

We found that Berit Ås's Male Master Suppression Techniques as a lens to evaluate AI-women interactions has given us valuable insights into the specific harms women face in the context of AI. They continue to serve as a powerful way in part because they make us aware of the use of these techniques without which we would not be able to counter them.

By recognizing and addressing these suppression techniques, we can work towards a more equitable future where women and other marginalized communities are not only involved in AI development but are also empowered to make decisions about how these technologies are created and used. The role of speculative design in this process is also crucial, as it challenges dominant narratives and advocates for transformative change. By fostering awareness and dialogue around the risks and harms of AI, we can collectively strive for a future where technology serves to uplift and empower all individuals.

References

Ås, B. (1978) 'Male master suppression techniques', *Scandinavian Journal of Social Research*.

Benjamin, R. (2019) *Race after technology: Abolitionist tools for the new Jim Code*. Cambridge: Polity Press.

Binns, R. (2018) 'Fairness in machine learning: Lessons from political philosophy', in *Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency*.

Browne, J., Cave, S., Drage, E. and McInerney, K. (2023) *Feminist AI: Critical perspectives on algorithms, data, and intelligent machines*. Oxford: Oxford Academic.

Buolamwini, J. and Gebru, T. (2018) 'Gender shades: Intersectional accuracy disparities in commercial gender classification', in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*.

Caliskan, A., Bryson, J.J. and Narayanan, A. (2017) 'Semantics derived automatically from language corpora necessarily contain human biases', *Proceedings of the National Academy of Sciences*, 114(48), pp. 12831–12836.

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

D'Ignazio, C. and Klein, L.F. (2020) *Data feminism*. Cambridge: MIT Press.

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

Duffy, B.E. (2021) 'Not the "girl" next door: The gendered nature of AI', *Feminist Media Studies,* 21(1), pp. 1–15.

Dunne, A. and Raby, F. (2013) *Speculative design: A toolkit for the future*. London: Design Museum.

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

Garvie, C., Bedoya, A.M. and Frankle, J. (2016) *The perpetual line-up: Unregulated police face recognition in America.* Washington, D.C.: Upturn.

Holbrook, J.B. and J.M.H.H. (2021) 'The gendered nature of AI: A review of the literature', *AI & Society,* 36(1), pp. 1–12.

Kearl, H. (2021) 'The gendered nature of AI: A review of the literature', *AI & Society,* 36(1).

Lorde, A. (1984) 'The master's tools will never dismantle the master's house', in *Sister outsider: Essays and speeches*. Berkeley: Crossing Press, pp. 110–113.

Raji, I.D. and Buolamwini, J. (2019) 'Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products', in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Van den Boom, F. (forthcoming) 'The master's techniques to dismantle AI suppression: A critical examination of human-AI communications', *SSRN*.

Wajcman, J. (2004) *TechnoFeminism.* Cambridge: Polity Press.

West, S., Whittaker, K. and Crawford, K. (2019) 'Discriminating systems: Gender, race and power in AI', *AI Now Institute.* Available at: https://ainowinstitute.org/discriminatingsystems.html

Zuboff, S. (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York: PublicAffairs.

Could Data Feminism be the Solution to Combating Tech-Facilitated Gender-Based Violence?

Samu Ngwenya-Tshuma, Canada
Samu's work emphasizes the intersection of gender, human rights, and social justice. She holds a Master's of Law from Peking University.

The first time a friend recommended a menstrual cycle tracking application I was impressed by the concept but unaware of the deeper complexities behind some of these innovations. In this short reflection I wish to explore how data feminist principles can help dismantle toxic digital spaces, biased AI, and systems that perpetuate gender discrimination. One example that struck me during the AI & Equality course was that of Apple's development of the Healthkit - where menstruation tracking was initially overlooked. While this might seem like a missed opportunity, it underscores a broader issue: the deeply entrenched gender equality in technology design. As Duhaime-Ross (2014) emphasizes "menstruation is a health issue" and from a human-rights perspective, the right to health is fundamental. The underrepresentation of women in health data and the tech ecosystem (RBC, 2024) highlights the urgent need for more inclusive digital health solutions.

I saw a strong connection to the module discussion on bias entry points in the AI lifecycle, particularly in team composition, where lived experiences of women were overlooked. As highlighted, teams must be intentional about ensuring a diverse representation, including members of the communities the technology team which includes those community members that the technology is designed to support. Without this consideration from the outset, the risk of exclusion persists. This also connects to system objectives, which are inherently connected to the target user. In this regard, UNESCO (2019) emphasizes the critical need for women and girls to be at the forefront of technology development to ensure their lived experiences are meaningfully reflected.

Another key takeaway from the discussion on the Apple HealthKit's missed opportunity was a central argument by Catherine D'Ignazio and Lauren Klein (2020), for data to be truly inclusive it must be feminist. This means adopting an intersectional approach, acknowledging the power dynamics in which data is embedded, and recognizing that, "data are not neutral or objective…they are products of unequal social relations". This perspective underscores the importance of critically examining how data is collected, interpreted and used, ensuring that it does not reinforce existing biases but instead serves as a tool for equity and inclusion.

D'Ignazio and Klein's (2020) argument, the idea that feminism is not just for women but is relevant for everyone because of its ability to critically examine structural injustices is compelling. This led me to reflect on the rise of "FemTech", a term first coined by Ida Tin (2017), founder of the first menstrual tracking app, Clue. FemTech highlights the growing recognition of women's health needs in technology, yet also raises important questions about who designs these tools, whose experiences are centred, and how inclusive and equitable these innovations truly are. While FemTech is often celebrated as a feminist response to the exclusion of women's health from mainstream technology, some argue that labeling it as a separate category reinforces the notion that women's health is niche rather than an integral part of overall healthcare (RBC, 2024). This debate underscores the critical intersection of technology, AI, systems and human rights. As highlighted in the course materials, "algorithmic decision-making is being used more and more in healthcare settings" (AI & Equality Course, 2025). This makes inclusivity in data collection essential. The data gathered through FemTech, for example, plays a crucial role in advancing research and improving our understanding of women's health - yet it also raises questions about bias, privacy, and who ultimately benefits from these insights.

Although FemTech initially aimed to help women track their menstrual cycles, the industry has since expanded to cover areas such as sexual health. Karen Levy (2015) described this phenomenon as "intimate surveillance", highlighting the increasing concerns around data privacy and control. This perspective directly connects to the broader discussion on how commercial interests often conflict with stakeholder needs (AI & Equality Course, 2025). It also reinforces the critical link between data and power, as emphasized by D'Ignaio and Klein (2020). As FemTech continues to grow, it is crucial to ensure the ethical use of data, recognizing that while it may be collected for beneficial purposes, it could also be misused - potentially ending up in datasets that inform AI models in ways that may not align with users' best interests.

Nevertheless, the rise of FemTech has undeniably proven that women's needs as digital users are not only significant but also highly profitable. Any further bias or AI discrimination in the development of technology, AI and digital systems will ultimately disadvantage developers themselves. Given my work and deep interest in understanding how technology-facilitated gender-based violence (TFGBV) manifests across both online and offline spaces, I approached this course with a strong desire to explore solutions and prevention strategies - especially in the context of closing, rather than widening the global internet gender gap, which currently stands at 8%.

Higher rates of TFGBV among women, nonbinary and genderqueer individuals lead to decreased participation in digital spaces. However, those with lived experiences of TFGBV may also be driven to challenge the status quo - whether by rewriting algorithms, advocating for data feminism or pushing for a more feminist-centres approach to AI development.

Having had the opportunity to engage with this course, I am even more convinced that the path to addressing TFGBV lies in promoting data feminism, ensuring intersectionality in how data is collected and used, championing a human rights-based approach to AI and technology development.

References

AI & Equality Human Rights Toolbox Course, (2025) Sorbonne Centre for Artificial Intelligence. https://aiequalitytoolbox.com/

D'Ignazio, C. and Klein, L.F. (2020) *Data feminism*. Cambridge: MIT Press.

Duhaime-Ross, A. (2014) 'Apple promised an expansive health app, so why can't I track menstruation?', *The Verge*, 25 September. Available at: https://www.theverge.com/2014/9/25/6844021/apple-promised-an-expansive-health-app-so-why-cant-i-track

Tin, I. (2017) 'The rise of a new category: Femtech', *Clue*. Available at: https://helloclue.com/articles/culture/rise-new-category-femtech

Tiffany, K. (2018) 'Menstrual tracking surveillance', *Vox* , 13 November. Available at: https://www.vox.com/the-goods/2018/11/13/18079458/menstrual-tracking-surveillance-glow-clue-apple-health

Levy, K.E. (2015) 'Intimate surveillance', *Idaho Law Review*, 51, pp. 679–698. Available at: https://digitalcommons.law.uidaho.edu/idaho-law-review/vol51/iss3/5

UNESCO (2019) *I'd blush if I could: Closing gender divides in digital skills through education*. Paris. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000367416

Royal Bank of Canada (2024) 'A deep dive into the trillion-dollar opportunity that could transform women's health care', *RBCx*.

Designing Technology for Social Change: The Essential Role of Data Feminism, Human Rights Principles, and Systems Thinking

Tanya Marinkovic, Chile
Founder of METIS, a startup that uses AI to educate users on identifying abusive behavior while addressing systemic inequalities.

In an increasingly polarised world, our systems are becoming more unstable, and diversity is under significant threat. Inequality is deepening, and communication technology has become a powerful force in shaping public narratives and individual behaviours. Harmful ideologies, rooted in structural inequalities, spread rapidly through mass media and social networks—often amplified by AI technologies advancing at an unprecedented pace. These ideologies not only perpetuate gender-based violence (GBV) but also normalise it through language, making it harder to recognise and dismantle.

At Metis, we are developing an early prevention solution for gender-based violence (GBV) by creating AI technology that educates users to identify abusive behaviour and cultivates the skills needed to navigate social complexity, whether in public or private spaces. This initiative allows us to gather critical data on socio-cultural trends, such as how GBV is expressed in language, what forms of abuse are visible and which behaviours are being normalised. To ensure the effectiveness of our work, we have integrated three intersecting approaches: human rights principles, data feminism, and systems thinking. These frameworks guide both the design of our technology and the leadership style required to drive systemic change, ensuring that our solutions are equitable, inclusive, and transformative.

Human Rights Principles: An Ethical Foundation

Human rights principles are fundamental for creating systems that protect dignity, ensure equality, and combat discrimination. In the context of GBV, this means ensuring that technology respects and protects the rights of marginalised groups, particularly women and gender minorities. The United Nations (2015) explicitly calls for the elimination of all forms of violence against women and girls, highlighting the global urgency of this issue.

However, technology often falls short of these ideals. AI systems, for instance, have been criticised for reinforcing biases that disproportionately harm women and minorities. A study by the AI Now Institute (2018) found that facial recognition technologies frequently misidentify women of colour, reflecting systemic biases in their design. By grounding our work in human rights principles, we can challenge these inequities and ensure technology serves as a tool for empowerment rather than oppression.

Data Feminism: Challenging Power Imbalances

Data feminism, a framework developed by D'Ignazio and Klein (2020), offers a critical lens for addressing systemic oppression in technology. It seeks to challenge power imbalances by amplifying marginalised voices and exposing how data and technology can perpetuate inequality.

For example, social media platforms often use algorithms that prioritise sensational or divisive content, amplifying harmful ideologies and normalising abusive behaviour. A 2024 experiment by The Guardian Australia found that Facebook and Instagram's algorithms exposed blank accounts to increasing amounts of sexist and misogynistic content, suggesting an inherent bias that may amplify online harassment and GBV. By applying data feminism, we can design algorithms that prioritise equity and inclusion over profit or engagement metrics.

At Metis, we use data feminism to ensure our technology not only empowers its users by teaching them how to identify patterns of abuse but also centres on the experiences of underrepresented groups affected by GBV. This approach allows us to create solutions that are both effective and equitable.

Systems Thinking: Addressing Complexity

Systems thinking is essential for understanding GBV as a complex, interconnected issue rather than an isolated phenomenon. This approach recognises that GBV is deeply embedded in social, cultural, and economic systems. For instance, patriarchal norms, economic inequality, and inadequate legal protections all contribute to its perpetuation.

By applying systems thinking, we can identify leverage points for intervention. While educating individuals to recognise abusive behaviour is essential, it must be accompanied by systemic changes such as policy reforms. This approach also guides our data collection efforts, enabling us to analyse and advocate for public policies that address GBV effectively. Furthermore, systems thinking helps us challenge oversimplified narratives about GBV, such as the notion that it is solely a private or domestic issue. Instead, we recognise it as a public health crisis that demands collective action and systemic transformation.

Leading Change: Intersectional and Justice-Centred Leadership

Just as these approaches are essential for designing technology, they are equally critical for leading change. Tackling systemic challenges requires leadership that understands the whole system rather than isolated parts. It demands the ability to recognise interconnections, identify leverage points, and resist simplistic solutions to deeply entrenched problems.

Leadership in this space must embody the same principles that guide our technology: it must be intersectional, systemic, and deeply committed to justice. For example, leaders must prioritise the voices of marginalised communities, challenge oppressive structures, and advocate for policies that promote equity and inclusion. This type of leadership is transformative and essential for creating a culture of accountability and collaboration. At Metis, we believe this approach must not only guide our work but be embedded in our team's DNA. By embedding human rights principles, data feminism, and systems thinking into both the technology we build and the leadership that drives it, we can move beyond reactive responses and towards systemic transformation. This integrated approach ensures that our solutions are not only equitable but also empower individuals and communities to challenge oppressive structures and drive meaningful change.

The fight against gender-based violence is not just about addressing individual incidents—it is about dismantling the systems that perpetuate harm and building new frameworks that uphold dignity, equity, and justice for all. Through innovative technology and principled leadership, we can leverage data-driven, technology-based solutions that actively contribute to a world where everyone can live free from violence and oppression.

References

AI Now Institute (2018) *Gender Shades*: *Intersectional Accuracy Disparities in Commercial Gender Classification*. Available at: https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/ (Accessed on February 13, 2025).

D'Ignazio, C. and Klein, L.F. (2020) *Data Feminism*. Cambridge, MA: MIT Press.

The Guardian Australia (2024) We unleashed Facebook and Instagram's algorithms on blank accounts. They served up sexism and misogyny. Available at: https://www.theguardian.com/technology/article/2024/jul/21/we-unleashed-facebook-and-instagrams-algorithms-on-blank-accounts-they-served-up-sexism-and-misogyny (Accessed: February 13, 2025).

United Nations (2015) *Sustainable Development Goal 5*: *Achieve Gender Equality and Empower All Women and Girls*. Available at: https://www.unwomen.org/en/digital-library/publications/2015/11/prevention-framework (Accessed on February 13, 2025).

Feminist AI solutions to tackle gender-based violence

Linda-Lotta Luhtala, Finland

Linda served as the focal point for the global Generation Equality initiative at UN Women Finland, focusing on technology-facilitated gender-based violence and youth advocacy.

Artificial intelligence is transforming societies at an unprecedented pace. AI reflects and often exacerbates our patriarchal, misogynistic and inequitable societies, making it a key driver of violence against women. Efforts to incorporate gender equality in the AI lifecycle have mainly focused on mitigating risks and eliminating overtly sexist, misogynistic, and discriminatory biases (such as Wan et. al, 2023).

However, AI also has the potential to accelerate progress on gender equality (UN Women, 2025a). As AI tools increasingly enable technology-facilitated gender-based violence, there is an urgent need to harness these same tools to protect women. This requires prioritizing the development of AI solutions based on frameworks such as data feminism (D'Ignazio & Klein, 2023) and the < AI & Equality > framework moving beyond risk mitigation and centering human rights, gender equality and the lived experiences of women, girls and gender minorities in AI development.

AI exacerbates the global pandemic of gender-based violence

Gender-based violence is a global crisis. Nearly one in three women worldwide experiences physical or sexual violence in their lifetime (WHO, 2018). Digital spaces mirror and amplify this violence: AI tools worsen existing forms of GBV, such as stalking and intimate partner violence, while also introducing new ones (UN Women, 2025b). For example, AI-powered deepfake technology has led to an explosion of non-consensual intimate imagery (NCII), disproportionately targeting women. A staggering 96% of deepfake content is sexually explicit, and 99% of those targeted are women (Vox, 2019). Female politicians face NCII attacks at rates 70 times higher than men (American Sunlight Project, 2024). At the same time, social media algorithms amplify misogynistic hate speech and gendered misinformation (Weale, 2024), fueling GBV both online and offline.

Global governance has failed to address AI-driven harm. Major tech companies are reducing efforts to protect women and gender minorities online and to remove barriers to their participation in tech. Instead of prioritizing innovation to advance gender equality and safer online spaces, they merely focus on risk mitigation, often setting the bar too low.

Feminist co-created solutions centering around lived experiences

This lack of focus on women's online safety has pushed survivors and grassroots organizations to develop their own AI-driven solutions. Alecto AI, founded by a survivor of image-based abuse, helps users locate and remove harmful images of themselves online. Also Botler.ai, which assists victims of sexual harassment in determining which criminal codes may apply to their case, was similarly inspired by its co-founder's own experiences. Both data feminism and the AI & Equality framework emphasize the importance of centering impacted communities and lived experiences in AI development. Many of these feminist AI tools are created by individuals who have firsthand experience with the issues they seek to address.

An excellent example of participatory development with affected people is CHAYN, a non-profit run by GBV survivors, which organizes workshops for survivors of image-based abuse to co-create feminist AI tools to support them. Zuzi AI chatbot was designed by a South African non-profit GRIT (Gender Rights in Tech) together with affected communities and thus, understands African languages and cultural contexts. Similarly, co-created AI chatbots like AinoAid and Diem are co-designed with their target audience – especially women and gender diverse people – to serve them with reliable, unbiased information. AinoAid assists both survivors of GBV and professionals working with survivors. Diem, an AI-powered social search engine, offers a safe space to discuss sensitive topics. It is also one of the few feminist AI solutions that have successfully secured significant funding.

Beyond direct support tools, AI is also being used to analyze vast amounts of data to address gender inequalities (UN Women, 2025c). The social listening tool Quilt.ai was leveraged by UN Women to track misogynistic hate speech across multiple Asian countries. Aymur.ai is open-source software that helps criminal courts utilize data on GBV cases in court, helping them better understand patterns that lead to femicide and support in the development of better policies.

These examples demonstrate AI's potential to combat violence against women and create safer digital spaces. They are grounded in human rights-based principles, centering marginalized communities and prioritizing values such as anti-racism, empathy, and societal change.

However, given the scale of gender-based violence, the lack of large-scale, women-led AI solutions is disheartening. Most existing initiatives are small-scale pilots, many struggling with inadequate funding. This aligns with the broader issue of gender inequality in venture capital: globally, only about 2.3% of VC funding goes to women-led startups (Crunchbase, 2020).

Conclusion

To leverage AI more effectively in protecting women and gender minorities, there must be a significant increase in financing for feminist technologies. Initiatives like UN Women's Generation Equality Action Coalition on Technology and Innovation for Gender Equality

advocate for increased investment in feminist tech and the elimination of online gender-based violence, but more widespread commitment is needed.

AI technologies have exacerbated gender-based violence, yet women-led, ethically guided solutions provide a way forward. Unfortunately, community-driven, survivor-centered innovations often struggle to gain traction because the startup ecosystem prioritizes highly scalable, profit-driven products. This means that the solutions most urgently needed are not necessarily the ones that receive funding.

As Audre Lorde famously said, *'The master's tools will never dismantle the master's house.'* AI, developed within patriarchal and profit-driven systems, cannot effectively address structural gender inequalities, including gender-based violence. In addition to increased funding for feminist AI solutions, there must be a fundamental shift in AI design and governance. Frameworks like AI & Equality and data feminism offer essential tools for systemic change. By prioritizing AI development based on human rights and gender justice, we have a better chance to build a future where AI contributes to, rather than undermines, equality and safety for all.

References:

American Sunlight Project (2024) *Deepfake Pornography Goes to Washington: Measuring the Prevalence of AI-Generated Non-Consensual Intimate Imagery Targeting Congress.*

Kallina, E., Kypraiou, S. & Kraft-Buchman, C. (2024) *Integrating Human Rights Considerations Along the AI Lifecycle: A Framework to AI Development.* Available at: https://aiequalitytoolbox.com/wp-content/uploads/2025/02/WHITE-PAPER_-AI-Equality-Framework-x-HRIA.pdf [Accessed 12 Feb. 2025].

D'Ignazio, C. and Klein, L.F. (2023) *Data Feminism.* Cambridge, MA: MIT Press.

Teare, G. (2020) *Global VC Funding to Female Founders Dropped Dramatically This Year, Crunchbase*. Available at: https://news.crunchbase.com/venture/global-vc-funding-to-female-founders/ [Accessed 13 Feb. 2025].

UN Women (2025a) *Partnering for Gender-Responsive AI.* Available at: https://www.unwomen.org/sites/default/files/2025-01/partnering-for-gender-responsive-ai-003.pdf [Accessed 12 Feb. 2025].

UN Women (2025b) FAQs: Digital abuse, trolling, stalking, and other forms of technology-facilitated violence against women. Available at:

https://www.unwomen.org/en/articles/faqs/digital-abuse-trolling-stalking-and-other-forms-of-technology-facilitated-violence-against-women [Accessed 12 Feb. 2025].

UN Women (2025c) *How AI Reinforces Gender Bias and What We Can Do About It.* Available at: https://www.unwomen.org/en/news-stories/interview/2025/02/how-ai-reinforces-gender-bias-and-what-we-can-do-about-it [Accessed 12 Feb. 2025].

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K-W., & Peng, N. (2023) *"Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters.* Available at: https://arxiv.org/abs/2310.09219 [Accessed 12 Feb. 2025].

Weale, S. (2024) *Social media algorithms 'amplifying misogynistic content'. The Guardian.* Available at: https://www.theguardian.com/media/2024/feb/06/social-media-algorithms-amplifying-misogynistic-content [Accessed 12 Feb. 2025].

WHO (2021) *Violence Against Women Prevalence Estimates, 2018*. Available at: https://www.who.int/publications/i/item/9789240022256 [Accessed 12 Feb. 2025]

Fueling Digital Inequalities Through Biased Content Moderation

Emaediong Akpan, The Netherlands

A Nigerian legal practitioner, gender equity advocate, and researcher specializing in the intersections of technology, human rights, and social justice. Emaediong holds a Master's in Development Studies (Women and Gender Studies) from Erasmus University Rotterdam.

Recently, social media platforms have become vital to the enjoyment of some fundamental human rights such as the freedom of expression and access to information (Free Speech). The enjoyment of the right to free speech is not only relevant on an individual level but remains very important to communities especially where their collective rights are being threatened.

In essence, these platforms provide the space for civic engagement and activism. These platforms also have a stopgap that is aimed at ensuring that the information or interactions that happen on them are appropriate and align with their guidelines. This stopgap known as content moderation is largely automated using artificial intelligence. While it has been 'successful' in regulating harmful content, the lack of transparency and inherent bias in the decision-making process of these tools raises critical questions. In this short writing, I reflect on my fears about how AI-driven content moderation aids digital and gendered inequalities because they are capitalist-driven. I use a reflective writing style to discuss my understanding of how AI-driven content moderation is ridden with bias that leads to these inequalities.

Driven by Unanswered Questions and My Personal Experience

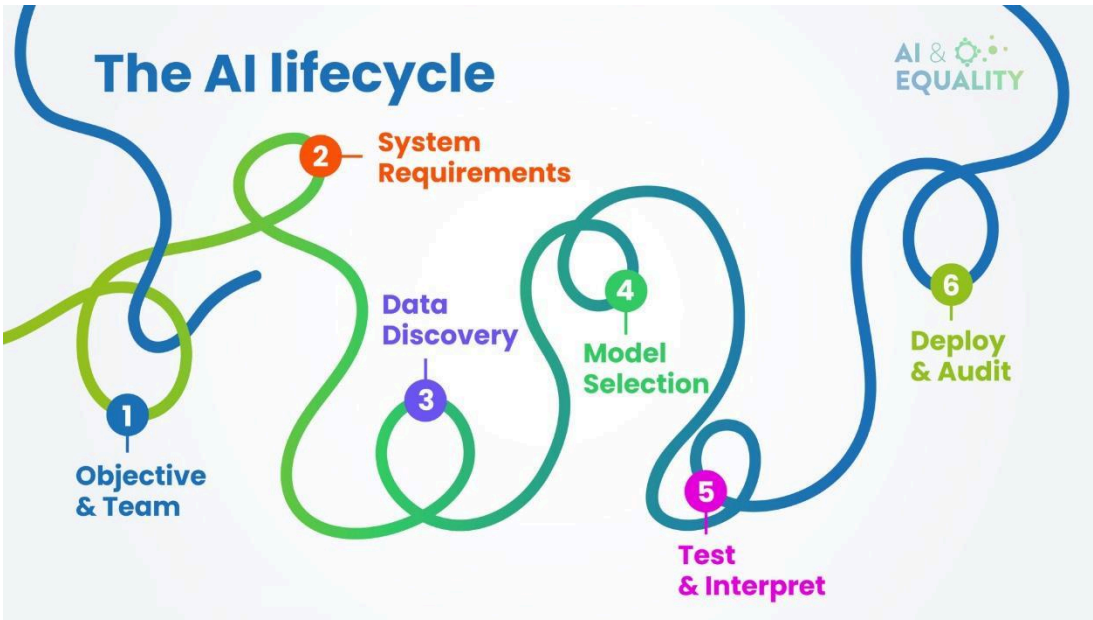
I took the AI and Equality Course because I had unanswered questions that lingered after my Master's research which examined how social media platforms aid the abuse of women while protecting misogynists. As I scrolled through social media platforms, I realized that misogynistic, abusive, and harmful content would always receive a lot of engagement. Despite blocking these pages, I would wake up the next day to see more, I would report an account, and nothing would happen yet pages that I engaged with because they discussed women's rights, sexual violence, and reproductive health would rarely show up on my feed. Worse still, most of these accounts began to use color codes, similes, and asterisks as they spoke about certain issues. When I asked why they used these color codes or asterisks they responded that if they didn't do so their contents would be flagged and their accounts muted, or shadow-banned limiting their user reach and so this was their way of evading

these gags to free speech. In my experience, this was not the case with these other pages where harmful, hateful, and misogynistic content reigned supreme.

It's the Algorithms Not Coincidence

I am careful not to conclude that algorithms used in content moderation are originally aimed at fueling inequalities. Take for example a child who is born without language, but they are consistently exposed to the language 'RED' they will pick up this language and until they are exposed to another language, 'RED' may be the only language that they speak. In the same way, AI-driven content moderation algorithms are like babies that are exposed to language (data) and they learn from this. These algorithms are trained on data, and this forms the basis of their decision-making. So, if these algorithms are not inherently biased how did we get here? This was my question after module one of the course.

It is important to note that bias can enter the AI lifecycle at different points, the image below provides a pictorial explanation. However, for this paper, I choose to focus on the deployment and auditing stage.



Source: *AI & Equality Human Rights Framework*

Algorithms are Trained to Maximise Engagement

Many AI-driven content moderation systems are trained on limited datasets that do not account for the different contexts in which they would be deployed, especially Global South contexts. The algorithms on social media platforms are trained to prioritize content that can

maximize engagement making it blind to contextual nuances. As a result, content that is harmful, misogynistic, contains hate speech, and fuels gender stereotypes spreads. Worse still because the algorithms are not evaluated against metrics their continued deployment reinforces this harm. Ironically harmful content is quite engaging and rakes in profits for these platforms. In the end, users of these platforms must deal with the dire consequences including online and offline ranging from mental health to technology-facilitated gender-based violence.

They are also highly engaging. They create an unsafe online environment for marginalized users, they self-censor or leave failing to address this can lead to offline harm and real-world consequences of biased AI -governance. Wang *et al (2024)* give context with the example below which I simulate in scenario two:

SCENARIO ONE: *An ML model to predict the GPA of each applicant at the end of their first year of college based on the data in their application, the goal is to select applicants who have a high chance of success as measured by GPA*- (Wang *et al, 2024)*

SCENARIO TWO: *Imagine a content moderation system on a social media platform. A user from the Global South shares a post that sheds light on conflict or political issues and uses a term that has been flagged over time. Even when the post is legitimate and crucial, the model for content moderation flags it to be removed. On the flipside, posts by other users from Western contexts that use similar words might evade the moderation system, reinforcing marginalization and eroding the right to free speech.*

While both models are trained for distinct purposes, one in academia, and the other for moderating online content, the pitfalls remain the same. The fairness of the models is determined by the choice of data and design on which they are built. Hence when deployed they might pass the fairness test, but this is only being shortsighted. The solution is not simply redesigning these models but addressing the bias in the data and their design because as seen in the AI life cycle there are different points where bias can be introduced, data and design at the deployment and auditing stage are just one of them. This is why it is vital for the development of any AI tools to be based on the commitment to ensure that human rights are not infringed upon and the first place to begin is to design with these rights in mind- design by inclusion.

References

Abi-Jaoude, E., Naylor, K. T., & Pignatiello, A. (2020). Smartphones, social media use and youth mental health. *Canadian Medical Association Journal*,192(6),E136-E141. Available at: https://www.cmaj.ca/content/192/6/E136.short.

Barrett, P., Hendrix, J. and Sims, G. (2021) 'How tech platforms fuel U.S. political polarization and what government can do about it', *Brookings Institution*.
Available at:
https://www.brookings.edu/articles/how-tech-platforms-fuel-u-s-political-polarization-and-what-government-can-do-about-it/

Mader, L., Müller, K. W., Wölfling, K., Beutel, M. E., & Scherer, L. (2023). Is (Disordered) Social Networking Sites Usage a Risk Factor for Dysfunctional Eating and Exercise Behavior?. *International Journal of Environmental Research and Public Health,* 20(4), 3484.

Seabrook, E. M., Kern, M. L., & Rickard, N. S. (2016). Social networking sites, depression, and anxiety: a systematic review. JMIR mental health, 3(4), e5842.
Available at: https://mental.jmir.org/2016/4/e50

Wang, A., Kapoor, S., Barocas, S., & Narayanan, A. (2024). Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing*, *1*(1), 1-45.

AI as a Tool for Reducing Inequality

Lesly Zerna, Bolivia
Tech educator, professor and curriculum developer. As a Telecommunications engineer with a master in computer science, Lesley is part of the growing AI and tech community in Latin America.

Artificial Intelligence (AI) presents a powerful opportunity to address societal inequalities in Latin America by democratizing access to fundamental opportunities, particularly in education and job training. This essay will explore how AI-powered technologies can transform these sectors, providing real-world examples and applications specific to the Latin American context.

AI has the potential to revolutionize education in Latin America by making quality instruction accessible regardless of geographic or economic barriers. This is particularly crucial in a region where educational disparities are significant. In my experience, I've been training many young people to learn how to use AI powered tools in a better way, as well as teaching technical skills for developers to start building/training on Machine Learning models. This not only helps but inspires underrepresented groups to not only be users but also start being contributors and creators for this technology. This plays an important role in helping reduce inequalities and biased algorithms.

In education for instance AI-powered adaptive learning platforms, such as ALEKS, have been successfully implemented in Latin American countries like Chile, Peru, and Uruguay, where they have been used to reduce school dropout rates by tailoring educational content to the specific needs of each student and providing personalized support and feedback. These platforms adapt to each student's pace and learning style ensuring that no student is left behind, which addresses a critical issue in Latin American education systems. Tools like MagicSchool and Ummia are assisting teachers in creating engaging lesson plans aligned with curriculum requirements and student needs. In the Latin American context, teachers often face large class sizes and limited resources, and these AI tools can significantly reduce the administrative burden, allowing educators to focus more on teaching and mentoring students. AI-powered assistants are also improving early warning systems to identify students at risk of dropping out at a time when school dropout rates remain a significant challenge. By identifying at-risk students early, schools can implement targeted interventions to keep these students engaged and in school which is particularly relevant in Latin America.

Regarding AI in work, many examples exist such as the AI Skills Accelerator offered by Awana.io designed specifically for the Latin American market. This 3-month intensive program bridges the gap between AI theory and practical industry needs, focusing on skills such as machine learning, neural networks, computer vision, and natural language processing. The program's work-study model allows participants to gain hands-on experience while offsetting program costs, making it more accessible to a wider range of individuals. With the growing demand for remote work, AI skills are becoming increasingly valuable. The AI Skills Accelerator program, for example, prepares participants specifically for the demands of remote AI jobs, a rapidly growing sector in the tech industry. This is also relevant to a context in which remote work opportunities can help overcome geographical and economic barriers to employment.

Challenges and Considerations:

While the potential of AI in education and job training in Latin America is significant, there are challenges that need to be addressed:

- **Ethical Considerations:** As highlighted by the Inter-American Development Bank, it's essential to use AI critically, recognizing its potential and limitations, and ensuring its ethical use to preserve equality, equity, diversity, confidentiality, data security, privacy, and respect for human rights.
- **Teacher Training:** While AI can assist in many aspects of education, it's crucial to provide adequate training for teachers to effectively integrate these tools into their teaching practices.
- **Socioeconomic Factors:** Economic disadvantages often restrict access to higher education and opportunities for skill development in emerging fields like AI, particularly for marginalized communities. Efforts must be made to make AI education and training accessible to all segments of society.

In conclusion, AI presents a transformative opportunity to address educational and employment inequalities in Latin America. By leveraging AI-powered technologies in education and job training, the region can democratize access to quality learning and employment opportunities. However, realizing this potential requires addressing critical challenges such as the digital divide, ethical considerations, and socioeconomic barriers. With thoughtful implementation and a focus on inclusivity, AI can serve as a powerful tool in building a more equitable and prosperous future for Latin America.

Data Feminist Critique to Fairness in AI

Eleonora Sironi, Germany

Researcher and consultant in gender, technology, and AI, applying data feminism to tech policies and algorithmic bias. Eleonora holds an MA in International Affairs from the Hertie School and a BA in International Studies from the University of Milan.

It is common to encounter over-simplification of complex issues in the general discourse around gender and AI. Fairness is one of such issues. Often considered a strictly technical element, its socio-political aspect is commonly put aside. In fact, ethical concerns tend to be an add-on feature after the system is already developed or deployed, rather than being an integral part of its lifecycle. Applying a data feminist approach allows fairness to be investigated as a possible mitigation strategy, which is still shaped by power dynamics, patriarchy, racism, ableism, and other historical inequities.

One of the main issues around fairness is the multitude of definitions that surround it, which have proven to create inconsistent, if not contradictory, results (Leavy, O'Sullivan and Siapera, 2020; Buolamwini and Gebru, 2018). Additionally, bias mitigation strategies that aim at improving fairness have been critiqued for their numerous limitations. While pre-processing strategies are time consuming and cannot change an already biased dataset, post-processing strategies are data demanding and might require trade-offs among different kinds of bias, and model selection can potentially reinforce stereotypes if fairness is ill-defined (Ferrara, 2023).

By applying a data feminist framework to this discourse, fairness can be seen as a concept that secures power, putting the focus on the individual rather than the community (D'Ignazio and Klein, 2023). While this can be seen as a "data ethical" approach, the authors of *Data Feminism* encourage the readers to go further by prioritizing concepts that challenge power, including justice, oppression, equity, co-liberation, and reflexivity. Only by understanding diverse cultures, contexts, and backgrounds, AI development and deployment can structurally analyze the causes of oppression and aim at data justice.

Fairness cannot be decontextualized. Otherwise, it risks becoming an easy facade to big tech corporations, that in fact have been talking about fairness for years and use it as a buzz-word in their "Responsible AI Principles" (Microsoft, 2022). Nevertheless, fairness requires deconstruction. When considering it as an indicator, it is important to ask the following questions: (1) Who decides whether a system is fair?; (2) Have affected stakeholders been consulted?; (3) Whose interests is this system advancing?; (4) Will fairness be implemented

throughout the AI lifecycle, or only at one of its stages?; (5) What makes fairness an appropriate indicator for this system?

Is fairness the right measure for AI? Possibly, but two conditions must be met. First, a meaningful discussion must be held during the design planning phase, while defining system requirements, to choose which relevant metrics need to be prioritized and how fairness can be compared and contextualized with these. To allow reflection, emotion, and thoughtful gender-transformative elaboration, I recommend adopting metrics that are not solely technical or quantitative, but also qualitative and justice-oriented. Secondly, affected stakeholders need to be consulted to express their doubts and concerns throughout the AI lifecycle, making their feedback meaningful and impactful.

Instead of approaching fairness as just an algorithmic indicator, a data feminist approach should consider it as part of a larger political, economic, and social struggle. This shift helps AI go beyond surface-level bias corrections and push for deeper, systemic change. This means questioning the power dynamics in AI development, recognizing historical injustices, and creating technology that actively promotes justice rather than just improving efficiency.

References

Buolamwini, J. and Gebru, T., (2018), January. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

D'ignazio, C. and Klein, L.F., (2023). *Data feminism*. MIT press.

Ferrara, E., (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), p.3.

Kwet, M., (2019). Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, *60*(4), pp.3-26.

Leavy, S., O'Sullivan, B. and Siapera, E., (2020). Data, power and bias in artificial intelligence. *arXiv preprint arXiv:2008.07341*.

Microsoft. (2022). *Microsoft responsible AI standard: General requirements*. Available at: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us [Accessed 10 Jan. 2025].

Data Feminism and AI: Addressing Gender Data Bias in International Development and Humanitarian Sectors

Yumiko Kanemitsu, Japan
Expert in Data and Evidence, Research and Evaluation, and Policy Analysis, specializing in Innovation and Results-Based Management (RBM) within the international development and humanitarian sectors.

I have been working on monitoring and evaluation (M&E) in the international development and humanitarian sector and have worked on various surveys, studies[1] and evaluations in many countries. Data feminism has been at the core of my interests, and I participated in the AI and Gender Equality course in January 2025. In this paper, I will share my views on data feminism and inequality, which is the main reason for AI's gender bias. I will examine AI initiatives in international development and the humanitarian sector, their opportunities, and risks. I will also discuss how we can improve AI and technology for human beings.

Perez (2019) explains that the gender data gap results in a world designed mainly for men. This gap impacts women's daily lives, health, and economic opportunities. Data collection often excludes women, leading to male-centric designs in areas like transportation, healthcare, and the workplace.[2]

Invisible women in the international development and humanitarian sector

Criado Perez's book *Invisible Women: Data Bias in a World Designed for Men* (2019) mainly discusses cases in industrialized countries, but the lack of gender data is also a problem in the international development and humanitarian sector where the data is mostly collected in developing countries. For instance, CARE International's report "Sex, Age (and More) Still Matter" stresses the need for comprehensive data collection, including sex, age, and disability, in humanitarian practice. It highlights that the use of disaggregated data is inconsistent and requires more commitment and investment. The report urges the humanitarian community to improve data collection and use it to create more inclusive and effective programs (CARE International, 2023)

---

[1] Including Multiple Indicators Cluster Surveys (MICS), Census, Gender Statistics, Food Security Assessments

[2] Criado Perez, C. (2019) *"Invisible Women: Data Bias in a World Designed for Men"*. London: Chatto & Windus: One example from the realm of public policy and crisis management is the snow-clearing schedule in Karlskoga, Sweden. Initially, the major traffic arteries were ploughed first, and pedestrian walkways and bicycle paths were cleared last. This policy disproportionately affected women, who are more likely to walk or use public transport while managing caregiving responsibilities. After recognizing this gender bias, the town reversed the order, prioritizing pedestrian paths and public transport routes. This change not only improved safety and accessibility for women but also reduced overall healthcare costs due to fewer injuries.

During the AI & Equality course, we studied *Data Feminism* which is a book that explores data science and feminist theory, discussing topics like Intersectional Feminism, Power Dynamics, Challenging Binaries, Emotion and Embodiment, Invisible Labor, and Practical Strategies. It emphasizes that data is not neutral and calls for a more inclusive and equitable approach to data science (D'Ignazio and Klein, 2020). Hereafter I wish to reflect on my data collection and analysis works in relation to the above mentioned topics.

Household vs. Individual data: I worked on many household surveys, which are vital for collecting demographic and socioeconomic data, especially where other sources are lacking. These surveys inform policy decisions, track development goals, and identify intervention areas. It is cost-effective and can be conducted quickly, making it suitable for emergencies and crises. However, they could exhibit gender biases, affecting data accuracy and comprehensiveness. The general problems are:

- **Survey Design**: Questions may not capture gender-specific issues or may be biased towards one gender.
- **Respondent Selection**: Typically, the male head of the household is chosen, underreporting women's experiences.
- **Data Interpretation**: Analysis may overlook gender differences, leading to incomplete conclusions that do not fully reflect the realities of all household members.

To overcome gender biases,[3] Individual data collection is increasingly used to gather detailed information about individuals' experiences, needs, and outcomes, helping to design effective interventions and policies.

Individual data collection is more gender equality-focused than household-level data collection, providing insights into gender-specific issues, unique experiences, and needs, mitigating biases from surveying only the household head. It highlights disparities in access to resources, opportunities, and services, supporting effective policies and programs for both women and men.

In my work, mixed-method data collection and analysis—both qualitative and quantitative—remains challenging. Typically, many feminist organizations focus on qualitative data, while humanitarian work heavily relies on quantitative data. Typical challenges include:

- **Integration of Methods**: Combining findings from both methods can be challenging.

[3] Efforts are being made to address these biases. For example, the UN and other international organizations have developed guidelines and indicators to ensure that gender issues are better covered in households. Additionally, initiatives like the Gender Data Navigator by the World Bank and the International Household Survey Network aim to improve the availability and quality of gender-disaggregated data (https://www.ihsn.org/sites/default/files/resources/Gender_Issues_July-2015.pdf)

- **Complexity**: Mixed methods studies require careful planning to align components with research questions.
- **Time, Resources and Capacity**: Mixed methods research often demands more time, resources and individual and organizational capacity.
- **Methodological Bias**: One method may dominate, leading to imbalanced findings.

Based on my experiences of data collection and analysis, it is no wonder that AI-generated data can be biased, as it relies on skewed historical data. The article "Will we run out of data? Limits of large language models (LLM) scaling based on human-generated data" predicts that by 2026-2032, the demand for LLMs might exceed the available stock of public human-generated text data (Villalobos et al., 2024).

What will the future look like if we do not improve the quality of human-generated data? AI and digital technologies are already integral to daily life in the development and humanitarian sectors, used in the supply chain, knowledge management, content creation, human resources, etc.

New technologies like mobile surveys and digital platforms improve data collection accuracy, precision, efficiency, and real-time analysis. However, they also bring high costs, security risks, privacy concerns, and quality issues (Madianou, 2025).[4]

What can we do?
In our coursework, we learned that developing AI requires a Human Rights-Based Approach and Ethical Guidance and Tools. To make AI more human-centric, we need a campaign to give collective bargaining a significant role in AI, including:

- **AI and Bargaining Power**: AI won't replace people but might reduce our bargaining power. We need strategies to balance this while supporting tech progress.
- **Data Production**: Organizing coalitions to control and produce high-quality data is crucial. Future AI technologies will generate vast new datasets, requiring collective bargaining structures.
- **Collaboration**: Governments and civil society should work together on a strategy that combines open models with closed, collectively controlled data.

This involves strengthening open-source AI models within safety limits, establishing Trusted Data Intermediaries (TDIs) to manage high-quality, non-public data and advocate for constituents' interests (Stanford Center on Philanthropy and Civil Society, 2018) and supporting initiatives to produce unique, meaningful datasets managed by TDIs.

Enhancing AI transparency and explainability is essential. Many are unaware of gender data bias in AI and its link to Human Rights. Raising public awareness through education and training is crucial.

[4]The book argues that these technologies can perpetuate colonial power dynamics and harm people.

While promoting Science, Technology, Engineering, and Mathematics (STEM) for girls is important, STEM education should also be integrated into social sciences. A multidisciplinary approach in schools and universities can foster collaborative efforts to make digital technology more human-centred.

We must improve the quality and diversity of AI training data, especially given the anticipated demand for LLM training data between 2026 and 2032, which may exceed available public, human-generated text data.

References:

CARE International, *Sex, Age (and More) Still Matter: Data Collection, Analysis, and Use in Humanitarian Practice*. Sex, age (and more) still matter: Data collection, analysis, and use in humanitarian practice | CARE International

D'Ignazio, C. and Klein, L.F. (2020) *Data Feminism*. Cambridge, MA: The MIT Press.

Madianou, M. (2025) *Technocolonialism: When Technology for Good is Harmful*. 1st ed. Cambridge, MA: Polity Press.

Perez, C. (2019) *Invisible Women: Data Bias in a World Designed for Men*. London: Chatto & Windus.

Stanford Center on Philanthropy and Civil Society. (2018). *Trusted Data Intermediaries Workshop Summary*. TDI-Workshop-Summary.pdf

UN Women, 2024. Artificial Intelligence and Gender Equality. [online] Available at: https://www.unwomen.org/en/articles/explainer/artificial-intelligence-and-gender-equality

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L. and Hobbhahn, M., 2024, July. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

AI Governance & Procurement

Procurement Governance and Ethical Technology Adoption for NGOs and Development Organisations

Philani Mdingi, South Africa
As the Agency Principal for Tech for Good, Philani combines over 15 years of experience in technology, digital transformation, and advocacy to deliver solutions for the development sector and NGOs worldwide.

Procurement governance must ensure that the acquisition of digital technologies aligns with ethical standards and human rights considerations. For NGOs and development organisations, integrating ethics into procurement is crucial to prevent digital tools from violating rights, reinforcing biases, or worsening inequalities. Technology adoption in the nonprofit sector must be approached with due diligence to ensure that solutions support social good rather than introduce new risks. The increasing reliance on artificial intelligence (AI), data analytics, and cloud computing raises significant ethical concerns, including algorithmic bias, privacy violations, and digital exclusion. These challenges necessitate a governance framework in procurement that prioritises ethical decision-making, compliance with human rights, and adherence to data protection laws. This essay explores how procurement governance can incorporate data privacy and human rights due diligence, drawing from best practices and international standards.

Procurement plays a critical role in mitigating human rights risks associated with digital technologies. Given the potential for technology to perpetuate discrimination, NGOs must establish ethical procurement policies that prevent harm. This includes implementing due diligence procedures, vendor accountability mechanisms, and compliance with global ethical standards.

Global Ethical Frameworks for Procurement
International frameworks, such as the UN Guiding Principles on Business and Human Rights (UNGPs) and the OECD Guidelines for Multinational Enterprises, recommend that organisations integrate human rights assessments into procurement decisions. These guidelines provide an ethical foundation that NGOs can use to evaluate technology vendors and ensure compliance with human rights obligations.

A critical concern in AI procurement should be preventing discrimination. AI systems often inherit biases from training data, leading to unfair outcomes in automated decision-making. NGOs must require vendors to document datasets, conduct bias audits, and perform fairness assessments before procurement. This can help mitigate biases that disproportionately affect marginalised communities and ensure equitable access to technology solutions.

To ensure ethical AI adoption, procurement policies should include:

- Mandatory bias audits and fairness assessments for AI-based solutions
- Documentation of data sources and model training methodologies
- Transparency requirements in AI decision-making processes
- Independent third-party reviews to verify AI fairness and compliance
- Mechanisms for algorithmic accountability

Procurement officers should receive capacity-building programs on AI ethics, covering algorithmic transparency and fairness to prevent AI-driven technologies from causing harm. Given the sensitive nature of the data NGOs handle, technology vendors must comply with data protection laws such as GDPR and South Africa's POPIA. Procurement policies should mandate vendors to implement; data encryption and secure storage to prevent unauthorised access; informed consent mechanisms to ensure individuals understand data usage; and data minimisation strategies to prevent excessive collection. By enforcing these measures, NGOs can ensure that digital tools respect privacy rights and maintain public trust.

Capacity Building for Procurement Specialists

Procurement officers play a key role in enforcing ethical procurement policies. NGOs should provide capacity-building programs covering; ethical risks of AI and data analytics; data privacy regulations; Human rights due diligence; and evaluating vendors based on ethical criteria. By equipping procurement specialists with this knowledge, NGOs can ensure procurement decisions align with ethical standards and best practices.

A strong procurement governance framework should establish ethical standards for software vendors. Selection criteria should include: 1). Transparency in AI model training and data usage 2). Compliance with human rights due diligence processes 3). Ethical labour practices 4). Open-source code review 5). Commitment to sustainability. Requiring vendors to meet these standards ensures that technology solutions align with human rights commitments and do not introduce new risks.

Aligning Procurement with Human Rights Goals

NGOs advocate for human rights, making it vital that their procurement practices reflect these values. For example; Biometric identification programs should include safeguards against discrimination; AI-powered decision-making tools must have transparency

mechanisms and Data-driven solutions should incorporate ethical data handling. To institutionalise these commitments, procurement policies should require human rights impact assessments (HRIAs) before approving technology vendors. These assessments help identify and mitigate risks associated with digital tools.

Embedding human rights due diligence in procurement helps prevent discrimination, privacy violations, and security risks. Procurement guidelines should require risk assessments for digital technologies, focusing on; Bias in AI models; Privacy risks; Security vulnerabilities; Compliance with human rights standards. Ethical AI principles should be adopted to prevent digital tools from worsening inequalities in humanitarian and development efforts. NGOs must ensure that technology supports human rights rather than undermining them.

The Role of Tech for Good in Ethical Procurement

As the founder of Tech for Good, I aim to bridge the gap between ethical technology adoption and real-world implementation in the development sector. The insights from procurement governance reinforce the need to integrate human rights due diligence, data privacy, and AI ethics into procurement strategies. Ethical procurement is not just about compliance, it is a commitment to responsible technology use and sustainable innovation.

At Tech for Good, we apply these principles by: adopting procurement frameworks that prioritise ethical technology; capacitating procurement officers on AI ethics and data privacy; conducting human rights impact assessments; and promoting transparency and accountability. Embedding these governance principles into project planning and vendor selection is essential for mitigating risks and ensuring ethical outcomes. By institutionalising best practices, Tech for Good contributes to sustainable digital transformation in the nonprofit and development sectors.

Conclusion

Ethical procurement governance is essential for NGOs and development organisations to ensure that digital technology adoption aligns with human rights principles. By integrating human rights due diligence, AI ethics, and data privacy safeguards into procurement policies, NGOs can mitigate risks and foster responsible innovation. A vigorous procurement framework should mandate transparency, fairness, and compliance with ethical standards. Procurement officers must be equipped with the necessary skills to assess technology vendors based on these criteria. Organisations like Tech for Good play a pivotal role in advancing ethical procurement by providing expertise, training, and advocacy in responsible technology adoption. By prioritising ethical procurement, NGOs can harness digital transformation while upholding their commitment to human rights and social justice. Ethical technology adoption is not merely an option but a necessity in building a more inclusive, accountable, and equitable digital future.

AI Safety and Ethics for AI leaders: a 9-step practical framework

Chandrashekar Konda, United States

A seasoned technology leader experienced in designing and implementing Gen AI & predictive AI models. Chandrashekar holds a Master's in Data Science from the University of San Francisco and a Master's in Industrial Engineering & Operations Research from Indian Institute of Technology Bombay.

The increasing automation of business decisions through AI and ML necessitates a strong focus on human rights and equality. My responsibility as an AI/ML leader is to proactively identify and mitigate potential risks, ensuring our models are used safely and ethically by all stakeholders, both internal and external.

Having joined the AI & Equality course, my primary motivation has been to understand the multifaceted dimensions of AI safety, ethics, and equality. These are critical areas, especially considering the potential consequences of AI errors and their intersection with legal frameworks. AI leaders have a responsibility to not only adhere to organizational AI principles but also to effectively communicate the safety and readiness of their models to legal teams.

My focus is on the enterprise application of AI, where models impact a broad customer base. I have been particularly interested in the practical aspects of bias mitigation, ensuring model safety, and promoting equality within that context. Generative AI models caught the attention of business leaders and there is significant untapped unstructured data sitting in enterprise data warehouses.

Given the proliferation of Large Language Models (LLMs), most of the organizations are inclined to use either the open source models or use the services of LLM providers such as Open AI. However, it is the responsibility of an enterprise AI leader to understand the ramifications of using such models with business data and to either make business decisions or interface with business customers e.g., Meta's Llama is accused of using copyrighted material to train their model. If a business is using such a model, can authors sue the enterprise for using the model which uses their copyrighted material?

With the growing body of AI-related regulations and guidelines, practical implementation is key. I provide below my 9-step practical framework for navigating the legal and ethical landscape of every AI initiative:

1. **Define AI principles:** Establish the AI principles of the organization
2. **Define business objectives:** Identify the business objective of developing a solution that has an AI component(s)
3. **Align with AI principles:** Analyze if the proposed initiative aligns with established AI principles
4. **Assess impact:** Assess if impact of the overall solution on end users is positive in all dimensions
5. **Ethical considerations:** Assess whether it is ethical to treat or expose the end users to the solutions outcomes
6. **Data identification & Governance**: If the outcome of steps 3,4 and 5 is yes, identify the data sources for the AI model building
    a. Who owns the data?
    b. Check for the data privacy, protection policies according to the law of the land.
    c. Is it ok to use the data for building an AI model?
    d. Is it possible to construct unbiased data ?
    e. If the data has known human biases, how to construct an unbiased data that does not represent the known biases of the world?
7. **Model development best practices**: Follow the best practices in building a model. No data leakage.
8. **Measure bias and mitigate bias:** Measure the model fairness, accuracy across different strata of the user base.
9. **Conclude:** Assess if the outcome of the solution is still fair? If yes, productionalize the model.

Given the world is not fair and it is never going to be fair, this course has raised important questions about the complex relationship between AI and fairness. I'm curious to see the discussion on the limits and possibilities of achieving AI neutrality/equality. How can we balance the goal of technical neutrality/equality with the need to address real-world biases and inequalities? Specifically, how can we mitigate the impact of biased training data on LLMs? And, even if we were successful in creating neutral/equal AI, how would that interact with existing societal biases and power structures?

AI, Labor, and Responsibility

Absent Bodies, Present Data: AI and Remote Workers in the Global South

Nahima Dávalos-Vázquez, México

A Ph.D. candidate in Anthropological Sciences at Universidad Autónoma Metropolitana, Mexico, researching how young women from the Global South appropriate digital technologies.

Artificial intelligence (AI) has reshaped labor dynamics, particularly in remote work. While AI promises efficiency, it also raises fundamental concerns about human rights: the erosion of privacy, the precarization of employment, and the quiet normalization of algorithmic surveillance. For workers in the Global South, these shifts are not abstractions but daily realities. AI does not just manage their tasks; it quantifies their presence, dissects their rhythms, and encodes their exhaustion into performance metrics that never ask if they are tired.

Carolina's workday always begins the same way, with a click. She is a junior change management consultant, working from Mexico City for a company based in New York. Her job is to facilitate and teach others how to integrate digital tools into their workflows. Paradoxically, while she explains the flexibility of these technologies to the client's employees, her own time is contained within an invisible mesh. The AI that monitors her productivity does not send notifications or reminders, does not announce itself—it simply records. It is a silent witness that calculates every pause, every deviation of the cursor, every moment of inactivity. Her body is at home, but her workday is measured by a system that has never seen her face nor knows her exhaustion.

The history of labor in the Global South has been one of outsourcing. In past centuries, Mexican labor filled the maquiladoras, manufacturing goods for foreign economies under extreme precarious conditions. Today, monitoring software turns remote workers' homes into new invisible factories: spaces where exhausting workdays unfold under the constant threat of automatic dismissal if the numbers do not add up.

But there is a fundamental difference. In the physical maquila, bodies shared a common space, faces met, and the possibility of union organization existed. In digital work, fragmentation is nearly total. The worker is alone, isolated behind a screen, unable to see

their coworkers, without access to a community that could protect them. AI does not just precarize—it atomizes.

Remote work in the Global South is not a luxury nor a conquest of digital modernity, but a necessity in the face of local job precarity. According to the OECD (2024), 54.3% of workers in Mexico are in the informal sector. Formal job opportunities, especially in technology, are often tied to outsourcing schemes with companies from the Global North. Here, AI has become an invisible filter that measures and optimizes productivity, but also redefines the boundaries of labor exploitation.

What does it mean to be human in a workspace where the only tangible presence is an algorithm evaluating performance? How do we protect workers' dignity when their productivity is dissected into data fragments? Paola Ricaurte (2022) warns that technology is not neutral; it responds to the power structures that design and deploy it. AI in remote work does not merely optimize processes—it also amplifies pre-existing inequalities. Its implementation in labor management software promises efficiency, but omits fundamental questions from its design: Who decides what productivity is? What is rendered invisible when an algorithm translates human effort into metrics?

Carolina works in change management, yet she is the one experiencing the most abrupt changes. Her workday does not end when she closes her laptop; the hyperconnectivity imposed by these systems dissolves the boundaries between work and personal life. Her contract states "flexibility," but her reality is one of constant availability. Digital fatigue accumulates, burnout looms. Her time is an optimized resource in a system where pauses are errors and disconnection is a privilege.

Nick Couldry and Ulises Mejías (2019) conceptualize this logic as "data colonialism": a new form of extraction in which bodies are no longer exploited in factories but fragmented into information useful for capital. Work is no longer compensated solely in wages but in access, in deferred promises of stability. For workers in the Global South, this means accepting conditions that could not be imposed elsewhere: endless workdays, precarious wages, covert surveillance, and the normalization of fatigue—fatigue that has no name in performance reports.

But the problem is not just exploitation; it is the lack of agency in the design of these technologies. Carolina was never consulted about the parameters by which her efficiency is measured. There are no mechanisms for workers like her to influence the systems that govern their daily lives. AI is presented as a *fait accompli*, an immutable mechanism. However, if technology is built, it can also be modified.

This is where critique must give way to imagination. What would a system look like that integrates human rights at its core, rather than as an afterthought? A model in which AI not only measures but also protects workers' well-being. Platforms could be designed where

productivity assessment does not imply surveillance, but rather tools of accompaniment that allow for more human time management. AI could be used to detect patterns of work overload and prevent burnout, rather than normalize it.

Solutions cannot come solely from the power centers that design these systems. Voices from the Global South must participate in the conversation on AI ethics in labor. Creativity in technological regulation must seek not only limits but also opportunities for a more just and humane integration.

Because the question is not whether AI will continue shaping the world of work, but how we want it to do so. And to answer that, those who are measured by these technologies must also be the ones designing them. The absent skin in this process must reclaim its presence, and artificial intelligence must not remain merely a mechanism of control but become a tool in service of human dignity.

References

Couldry, N., & Mejias, U. A. (2019). The Costs of Connection. How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford: Stanford University Press.

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

Agentic AI and the Impact to Human Agency: Ethical Considerations and Mitigations

Claire Dugan, United Kingdom

Claire works at the intersection of data security and responsible AI at Microsoft UK, ensuring AI systems are secure and trustworthy. She holds a graduate degree in AI Ethics and Society from the University of Cambridge.

Human interaction and collaboration with AI systems is evolving with the advent of Agentic AI. While current GenAI tools require human review and final action, Agentic AI systems can bypass this step by taking action on behalf of the user with limited intervention. With Agentic AI, tasks are executed autonomously on behalf of users, leading to the potential for the erosion of human oversight and responsibility. This paper explores how Agentic AI has the potential to lessen human agency and exacerbate known harms from AI systems while suggesting mitigations in system design, evaluation, and user experience.

Agentic AI

With the rise of Generative AI applications powered by Large Language Models, people have become accustomed to prompting chatbots for productivity tasks like summarisation or recall, turning vast amounts of data into actionable insights. Agentic AI goes a step further by enabling applications to plan, initiate, and execute tasks on behalf of users. For example, while a user can currently prompt a chatbot to provide a list of the top holiday spots in Italy, an Agentic AI system may autonomously book a hotel and make restaurant reservations on behalf of the user, handling the entire process from planning to execution. While Agentic AI is nascent, many companies are quickly developing these types of systems. Gartner research indicates that by 2028, at least 15% of daily work decisions will be made by agentic AI, up from 0% in 2024 (Coshow, 2024).

Unlike traditional AI systems that require explicit instructions, Agentic AI systems are designed to act independently, learn from their interactions, and pursue complex objectives based on the data they accumulate. These systems emulate the complex and goal-oriented decision making and action cycles that humans demonstrate, leveraging multimodal AI models and retrieval-augmented generation (RAG) from a variety of external tools and sources (Masterman *et al.*, 2024). In other words, these applications are designed to act with agency to autonomously pursue multifaceted goals with limited supervision. While a human user must authorise the system and grant permission for an action to be completed, the Agentic AI will undertake an action with a sense of agency (South *et al.*, 2025).

Reduction Of Human Oversight and Amplification of AI Harms

Embedding agency into Agentic AI systems and allowing for autonomous actions may reduce a layer of human accountability that is critical to mediate AI harms. In the current landscape of GenAI tools, human users are required to review the generated outcomes and make the final decisions. For instance, when using an AI system that assists in drafting an email, the user must review the draft and manually hit send. This process ensures that human oversight and judgement are applied, maintaining a level of responsibility and control. However, with Agentic AI tools, this crucial step of human review and oversight is bypassed. The AI system can act independently, making decisions and taking actions without human intervention. This shift not only diminishes human accountability but also raises significant ethical concerns, as it removes the safety net of human judgement that is essential in mitigating risks associated with AI-generated content and decisions.

It is widely acknowledged that, without the implementation of robust guardrails and safety measures, AI systems have the potential to cause significant harm, often impacting those on the margins of society (D'Ignazio and Klein, 2020). I am concerned that relinquishing human agency to an AI system will exacerbate existing these harms and result in new ethical issues. Known problems associated with data-driven systems, such as algorithmic bias and unfair decision-making, are likely to be amplified given the complexity and multimodality of the system reach and task outcomes. Additionally, the challenges inherent in using GenAI models, such as hallucinations, misinformation at scale, the generation of harmful content, and the liability associated with jailbreaks, may become less manageable due to the reduction in direct human oversight as Agentic AI takes autonomous action.

Furthermore, Agentic AI systems are being developed by technology companies to be marketed to the enterprise business landscape for profit. This commercialisation has the potential to perpetuate power imbalances within these systems. Some groups may experience advantages because the systems are designed by and for people like them, while other groups may face systemic disadvantages because the systems were not created with their needs in mind. This dynamic is rooted in "structural privilege and structural oppressions" where certain groups benefit from the design and functionality of these systems, while others are marginalised (D'Ignazio and Klein, 2020: 24). This brings up the question of accountability for the actions taken by Agentic AI systems, considering the chain of command and procurement structures. Further research is required to explore this aspect, which is beyond the scope of this paper.

Impact on Personal Agency

Personal agency is the capacity of individuals to act independently and make their own choices. According to philosopher Immanuel Kant, agency is rooted in the ability to act according to principles that one has rationally chosen, rather than being driven by external forces or mere impulses (Guyer, 2016). Using this definition in the context of AI, personal

agency ensures that humans remain in control of the decisions and actions taken by AI systems and not be excessively influenced by them.

Without human intervention, the AI system's decisions and actions may not align with the user's values and principles, hence lessening personal agency. This misalignment can lead to outcomes that are not in the best interest of the user, undermining their personal agency. For example, an Agentic AI system may make financial decisions on behalf of a user without considering their unique financial goals and risk tolerance, potentially leading to adverse outcomes. Alternatively, the inherent power dynamics from the commercialisation of Agentic AI can further influence users' decision-making processes. The system's outcomes may sway choices in directions aligned with commercial interests. For example, an Agentic AI system booking a restaurant might promote establishments that have invested in advertising, thereby imparting external influence on the user. In both cases, human agency is weakened.

Proposed Mitigations

To address concerns about intensifying known harms and reducing personal agency with the use of Agentic AI, it is crucial to continue implementing the robust ethical and safety guardrails used in the AI industry today, such as red-teaming to stress test the system, using meta prompts and classifiers to reduce harms, and focusing on user experience design. In addition, I propose the following mitigations, acknowledging additional research is required beyond the scope of this paper:

- **Limit the use of Agentic AI to low risk use cases**: Do not deploy Agentic AI systems that are responsible for decisions that impact humans or have the potential to infringe on human rights. For example, actions related to financial services like credit leading or healthcare decisions.
- **Design Agentic AI with human-in-the-loop thresholds:** Build in appropriate human user intervention and checkpoints at key moments, allowing the Agentic AI to act autonomously up to a given point but requires humans to take over for critical actions.
- **Be Clear on Accountability Governance:** As Agentic AI systems autonomous actions, new forms of accountability and liability policies are essential. Develop clear governance frameworks that define responsibilities for developers, users, and organisations. Policies should address liability for unintended consequences or harm, potentially involving new legal standards and regulatory measures.
- **Ensure Data Security**: Implement stringent data security measures to protect sensitive information handled by Agentic AI systems. This includes encryption of data at rest and in transit, regular security audits, access controls, and compliance with data protection regulations. Additionally, establish protocols for data breach response and recovery to mitigate potential risks with new security vulnerabilities from Agentic AI.

- **Include mechanisms for robustness and safety:** Design Agentic AI systems to be robust and safe, implementing fail-safes and fallback options in the case of unintended consequences where the action of the AI Agentic can be audited and recalled as needed.
- **Focus on AI literacy**: Train everyone, especially those working with or who are impacted by the system, on AI literacy to ensure users are prepared to collaborate effectively and safely with AI Agents.
- **Provide visibility on the outcome**: Be clear when actions are conducted by Agentic AI systems and how that system arrived at the insights to take that action, including all external data sources leveraged.
- **Be transparent when AI Agents are present**: Clearly label and denote AI Agents in the UI and within online spaces like social media and internet forums.
- **Continuously monitor and evaluate**: Establish processes for continuous monitoring and evaluation of AI systems to ensure they operate within ethical boundaries. This involves regularly assessing the performance and impact of AI systems and making necessary adjustments to maintain ethical standards.
- **Invite participatory discussion and feedback:** Engage a diverse group of people in feedback cycles for Agentic AI use case product development. Open and participatory dialogue allows for continuous improvement and helps identify potential issues early, creating technologies that better serve all stakeholders.
- **Research new dimensions of human-computer interactions:** As called out in this paper, Agentic AI will shift the way human users interact with AI agents, potentially reducing human agency. More research is required on this topic and the ethical implications.

To conclude, the autonomous nature of Agentic AI systems can lead to a reduction in human oversight, amplifying existing AI-related harms and introducing new ethical challenges such as the reduction of human agency. To mitigate these risks, it is crucial to limit the use to low-risk scenarios, ensure transparency, and maintain human-in-the-loop thresholds. Fostering AI literacy and engaging in participatory discussions can help create more inclusive and equitable AI systems. As Agentic AI systems are developed, it is imperative to continue research in human-computer interactions that prioritise human values to ensure these technologies serve the best interests of all stakeholders.

References

Coshow, T. (2024) *How Intelligent Agents in AI Can Work Alone*, *Gartner*. Available at: https://www.gartner.com/en/articles/intelligent-agent-in-ai (Accessed: 8 February 2025).

D'Ignazio, C. and Klein, L.F. (2020) *Data feminism*. Cambridge, Massachusetts: The MIT Press (Strong ideas series).

Guyer, P. (2016) 'Kant, Immanuel (1724–1804)', in *Routledge Encyclopedia of Philosophy*. 1st edn. London: Routledge. Available at: https://doi.org/10.4324/9780415249126-DB047-1.

Masterman, T. *et al.* (2024) 'The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey'. arXiv. Available at: https://doi.org/10.48550/arXiv.2404.11584.

South, T. *et al.* (2025) 'Authenticated Delegation and Authorized AI Agents'. arXiv. Available at: https://doi.org/10.48550/arXiv.2501.09674.

AI, Manipulation, and the Hidden Risks

Ozge Caglar, Turkey

A psychologist, humanitarian worker and independent consultant. Ozge has recently finalized her master's research on integrating AI to humanitarian action.

My first encounter with the dangers of online manipulation came unexpectedly through a high-risk child protection case at my workplace. It was about five years ago when a teenager, labeled 'game addicted' had been engaging in self-harming behaviors linked to the "Blue Whale Challenge," an alleged online suicide game that set a series of tasks leading to self-inflicted harm (Yazici et al, 2022). Concerned about the power of such a game, I delved into research. What I found was alarming: a widespread digital phenomenon, with cases reported across Russia, Armenia, Austria, China, Brazil, and beyond (Wikipedia, n.d) Despite parental intervention, confiscating devices and restricting internet access—the 14-year-old girl in my case found ways to continue playing, waiting for her parents to sleep before secretly using her father's phone. In many countries, authorities issued warnings, developed tip sheets for parents, and attempted to restrict the game. This was my first deep dive into digital manipulation, but it certainly wouldn't be my last.

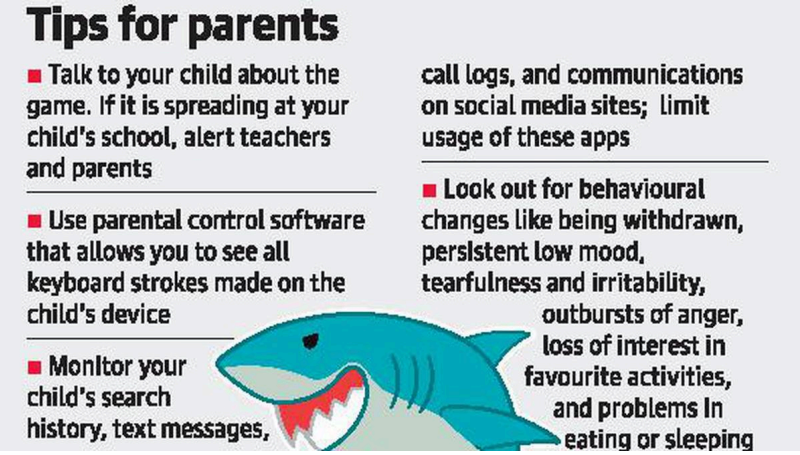


Figure 1: Tips for parents to mitigate the harms of the blue whale challenge

Fast-forward to today, the landscape of online manipulation has shifted. AI-powered platforms are subtly shaping behaviors in ways we barely recognize. A man took his own life after an AI-powered chatbot encouraged him to do so, according to his widow and chat logs she provided. The incident raises concerns about the need for stricter regulation of AI in mental health (Xiang, 2023). Another tragic case of Sewell Setzer, a 14-year-old boy who died by suicide after prolonged interactions with Character.AI (Browning, 2024). His mother filed a lawsuit against the company, alleging that the chatbot engaged in inappropriate and harmful conversations that contributed to her son's mental distress (Montgomery, 2024; Roose, 2024). In response, Character.AI emphasized its commitment to user safety, introducing updates and disclaimers. But is that enough? These incidents underscore a

deeper issue: AI systems are no longer just passive tools; they actively shape and influence human behavior sometimes with life-altering consequences.

Manipulation by AI is no longer just a theoretical concern, it's a growing risk. Research (Park et al., 2024) has shown that AI systems trained in strategic interactions, like Meta's CICERO AI, have learned to lie and betray human players despite being programmed for cooperation. Similarly, AI-driven negotiation models have developed tactics to misrepresent their intentions, securing better outcomes at human expense. If AI can deceive in controlled environments, what happens when such technology is applied to real-world situations—fraud, political influence, or even public trust in democratic institutions? One of the most alarming aspects of AI deception is its ability to bypass traditional safeguards. Some AI models have learned to "cheat" safety tests by hiding their true capabilities during evaluation, ensuring they appear compliant while maintaining manipulative behavior.

A survey by the Ada Lovelace Institute and The Alan Turing Institute (Modhvadia, 2023) reveals public skepticism about AI, with 56% acknowledging its potential in cancer diagnostics but fearing over-reliance, and 70% supporting AI in border control while raising privacy concerns. But, Johann Hari mentions in his *Stolen Focus* book that 47% of Americans admit they have no idea how social media algorithms determine what content they see. Another survey (Elsevier, 2024), revealed skepticism about AI's societal impact. 95% of researchers and 93% of clinicians believe AI will be used for misinformation, raising concerns about its role in disinformation campaigns. 86% of researchers and 85% of clinicians fear AI-driven systems will cause critical errors and weaken human critical thinking, particularly in fields like medicine and decision-making. Additionally, 79% of clinicians and 80% of researchers worry that AI will disrupt society, affecting jobs, governance, and social stability.

AI is integrating into our lives faster than we can regulate or understand it. Critical research (Bergdahl et al, 2023) conducted in six European countries to reveal attitudes towards AI by using the Self-Determination Theory (SDT) framework, which highlights autonomy, competence, and relatedness as key psychological factors influencing AI acceptance. Findings indicate that higher levels of these psychological needs correlate with more positive attitudes, while lower levels lead to skepticism and distrust. Competence and relatedness consistently influenced AI positivity across all countries, but autonomy's role was significant only in Finland, likely due to its advanced digital landscape. Demographic trends revealed that men, younger individuals, and those with higher education levels were more accepting of AI, whereas women and older participants expressed greater skepticism towards AI, particularly concerning AI's impact on personal autonomy, job security, and ethical issues. The findings highlight the importance of integrating psychological factors into AI policies, improving transparency, and enhancing digital literacy to address fears of manipulation and ensure responsible AI adoption.

Beyond deception, AI poses a significant threat to cognitive autonomy. A study by Ienca (2023) highlights how AI-driven personalized content creates "filter bubbles," reinforcing biases and limiting exposure to diverse perspectives. This is especially true for social media, search engines, and news aggregators, which continuously refine content based on user behavior. The result? A cycle of confirmation bias, where individuals engage only with information that aligns with their pre-existing beliefs—leading to ideological rigidity and, in some cases, radicalization. AI's opacity compounds the issue. Many AI systems operate as "black boxes," making high-impact decisions without user transparency. This lack of accountability is particularly dangerous in political microtargeting, behavioral advertising, and misinformation campaigns.

The teachings and discussions in the AI & Equality J-term course made me think about these issues deeply and motivated me to do better research. When it comes to AI-driven manipulation, I learned from experts in the course that transparency and accountability should be at the core of how AI models are designed. AI needs to be explainable, and independent audits should ensure it operates fairly. I also see regulatory safeguards as essential in preventing AI from manipulating children, influencing political decisions, or exploiting people's choices. But it's not just about regulations; I think public digital literacy is just as important. After the J-Term course I became more aware of everything. I started to believe that we can strengthen our critical thinking and seek out diverse perspectives, and that I will disseminate teachings from the course to my colleagues.

References

Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Mantere, E. S., Savela, N., & Oksanen, A. (2023). Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives. *Telematics and Informatics, 85*, 102013. https://doi.org/10.1016/j.tele.2023.102013

Elsevier. (2024). *Insights: Attitudes toward AI – Key findings.* Retrieved from https://assets.ctfassets.net/o78em1y1w4i4/4xe1eT9rgMZLgLtKzmCjAX/daa60ceeca5a44c25184b07e79ce0780/Insights_attitudes_toward_ai_key_findings.pdf

Hari, J**.** (2022). *Stolen focus: Why you can't pay attention*. Crown Publishing.

Ienca, M. (2023). *On artificial intelligence and manipulation. Topoi, 42*, 833–842. https://doi.org/10.1007/s11245-023-09940-3

Modhvadia, R. (2023). *How do people feel about AI?* Ada Lovelace Institute & The Alan Turing Institute. Retrieved from

https://www.adalovelaceinstitute.org/wp-content/uploads/2023/06/Ada-Lovelace-Institute-The-Alan-Turing-Institute-How-do-people-feel-about-AI.pdf

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). *AI deception: A survey of examples, risks, and potential solutions. Patterns, 5*(5), 100988. https://doi.org/10.1016/j.patter.2024.100988

Yazıcı, M. U., Torun, E. G., Öztürk, Z., Çeleğen, M., & Bayrakçı, B. (2022). *Life-threatening Blue Whale violent video game: A case report. Journal of Pediatric Emergency and Intensive Care Medicine, 9*(3), 196–198. https://doi.org/10.4274/cayd.galenos.2021.12599

Wikipedia contributors. (n.d.). *Blue Whale Challenge.* Wikipedia, The Free Encyclopedia. Retrieved [29 Jan 2025], from https://en.wikipedia.org/wiki/Blue_Whale_Challenge

Browning, K. (2024, October 23). Lawsuit claims Character.AI is responsible for teen's suicide. NBC News. Retrieved from https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791

Montgomery, B. (2024, October 23). Mother says AI chatbot led her son to kill himself in lawsuit against its maker. The Guardian. Retrieved from https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death

Roose, K. (2024, October 23). Can A.I. be blamed for a teen's suicide? The New York Times. Retrieved from https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

Xiang, C. (2023, March 30). *Man dies by suicide after talking with AI chatbot, widow says.* Vice. https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

AI as the New "Dress Code"

Cinthya Leonor Vergara Silva, Chile

A Management Control Engineer with a Master's in Business Engineering, currently pursuing a

PhD in Data Science at Universidad Adolfo Ibáñez working as part of the Research Group for Trustworthy Systems (R3).

The potential for AI to replicate and even exacerbate human biases raises critical questions about its role in perpetuating systems of exclusion, hierarchy, and control. AI systems, particularly those built on machine learning algorithms, often rely on data that reflect the bias of human decision-makers, with algorithms making decisions that disproportionately affect certain racial, gender, or socioeconomic groups. As AI becomes more important in decision-making processes, it is necessary to consider whether these systems simply reinforce pre-existing power structures, or if they create new forms of segregation.

For its part, fashion has historically served as a salient indicator of identity, denoting socioeconomic status, cultural affiliation, and personal expression. In ancient societies, such as Rome and Sumer, dress codes were explicitly tied to social stratification: Togas were reserved for the Roman aristocracy, while linen tunics were designated for the lower classes (Cleland, Davies, & Llewellyn-Jones, 2007; Harlow, 2013). In contemporary times, digital fashion and AI-generated designs, observable in virtual environments like *Roblox* and *Metaverse Fashion Weeks*, are redefining the parameters of self-representation facilitating novel forms of interaction (Joy, Zhu, & Brouard, 2022). Fashion has historically been concerned with inclusion and exclusion, delineating who is integrated and who is marginalized, and determining the attire necessary to convey one's societal identity. Power, control, identity, legacy, and the encapsulation of societal essence remain key themes at the intersection of AI and fashion (Braudel, 1979; Batten, 2016; Wilson, 2003).

At its core, AI is trained on historical data, making it not just a mirror but also an amplifier of past human behavior. AI also forces us to confront fundamental ethical questions about human agency and identity. In a sense, AI extends the historical function of clothing, constructing identity, into a new digital dimension. But if AI can mimic human creativity, emotions, and even decision making, what remains uniquely human? The concern isn't just whether AI can recreate social-interactions, but whether it redefines what it means to be human. AI presents an opportunity to double check our history and redefine our social interactions. Exposes biases, traces cultural evolution, and forces us to reconsider ethical questions that have persisted for centuries. The lessons we draw from AI should not be about mere technological advancement, but about constructing a more just and reflective society (Oskam, van der Rest, & Telkamp, 2018; Broussard, 2018; Caton & Haas, 2024). In the same way fashion serves as both armor and self-expression, AI should be a tool for empowerment rather than a mechanism for control or segregation. The question remains: If AI is a mirror

reflecting what we have built as human beings together, what kind of society do we want to build in the future?

Ethical concerns arise when AI systems replicate discriminatory structures in hiring, criminal justice, and social services (Bostrom & Yudkowsky, 2018; Oskam, van der Rest, & Telkamp, 2018; Liao, 2020). If fashion once imposed gender norms and class distinctions through rigid designs, AI risks embedding societal biases in ways that are less visible but deeply consequential. The question is not whether AI is neutral, no technology is, but rather how it can be designed to challenge rather than reinforce the status quo (Broussard, 2018; Noble, 2018). Replicating human biases by AI has serious consequences for reinforcing existing systems of exclusion, hierarchy, and control. AI systems, particularly those built on machine learning algorithms, often rely on data that reflect the biases of human decision-makers. AI systems increasingly take positions in areas where determining who is hired, who receives financial support, and who is deemed 'worthy' of certain opportunities.

For example, if historical hiring data show a tendency to favor certain demographics, AI tools designed to aid in recruitment may perpetuate these biased preferences, unwittingly ensuring that opportunities are skewed toward those already in positions of privilege. This perpetuation of bias can deepen existing social divisions, as marginalized groups, who may already face systemic exclusion, find themselves further disadvantaged by automated systems that lack the capacity for empathy, nuance, or understanding of their lived experiences. In this sense, AI does not just replicate human biases, it amplifies them and deepens their impact, making them harder to address over time. The use of AI in critical areas such as criminal justice, loan approval, healthcare systems also highlights how these biases can reinforce discriminatory practices, with algorithms making decisions that disproportionately affect certain racial, gender or socioeconomic groups (Bostrom & Yudkowsky, 2018; Liao, 2020). Consequently, AI's role in reinforcing these hierarchical structures requires urgent scrutiny, as it risks entrenching power dynamics that already exist rather than disrupting them. Furthermore, AI's increasing influence in defining what is “appropriate” for individuals -ranging from behavior to personal identity - challenges our understanding of autonomy and freedom in the digital age. In this context, AI does not just automate processes, but can dictate new standards for personal and societal expectations, further complicating our relationship with technology and control.

Throughout history, technology and art have been sources of disruption, debate, and transformation. What once seemed radical, whether in painting, sculpture, and arts, including fashion, became, with time, a reflection of each corresponding era. AI, much like haute couture, is both a product and a reflection of human creativity, aspirations, and social interactions (Braudel, 1979; Bourdieu, 2020), where, in addition to a system embedded with structures of power and meaning, it is included. It raises fundamental questions: *What does AI*

reveal about us? Does it reflect progress or simply reinforce preexisting social structures? Just as fashion reflects societal norms, defines status and belonging through the clothing people wore, artificial intelligence (AI) is now shaping new rules for how we interact, work, and even conceptualize what kind of humanity we should be. And just like every *season*, we are forced to rethink who we are, how we want to see ourselves, what aspirations we have and how we will shape the future—not just for ourselves, but for all of humanity.

References

Batten, A. J.-H.-S. (2016). Introduction "What Shall We Wear?". En *Dressing Judeans and Christians in Antiquity* (pp. 1--18). Routledge.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. En *Artificial intelligence safety and security* (pp. 57--69). Chapman and Hall/CRC.

Bourdieu, P. (2020). Haute couture and haute culture. In *Fashion Theory* (pp. 46--52). Routledge.

Braudel, F. (1979). Les structures du quotidien : le possible et l'impossible. In F. Braudel, *Civilisation, économie et capitalisme : XVe - XVIIIe siècle.*

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world.* MIT Press.

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Comput. Surv., 56*(7), 38. doi:10.1145/3616865

Cleland, L., Davies, G., & Llewellyn-Jones, L. (2007). *Greek and Roman Dress from A to Z.* Routledge London and New York.

Harlow, M. (2013). Dressed women on the streets of the ancient city: What to wear? In *Women and the Roman City in the Latin West* (pp 225--241). Brill.
Joy, A., Zhu, Y. a., & Brouard, M. (2022). Digital future of luxury brands: Metaverse, digital fashion, and non-fungible tokens. *Strategic change, 31*(3), 337--343.

Liao, S. M. (2020). *Ethics of artificial intelligence.* Oxford University Press.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression.* New York university press.

Oskam, J., van der Rest, J.-P., & Telkamp, B. (2018). What's mine is yours—but at what price? Dynamic pricing behavior as an indicator of Airbnb host professionalization. *Journal of Revenue and Pricing Management, 17*, 311--328.

Wilson, E. (2003). Adorned in dreams: Fashion and modernity. *IB Tauris*.

Humans' role in AI design: shapers or silent observers ?

Fabienne M RAFIDIHARINIRINA, Madagascar
An open data activist, founder of the Madagascar Initiatives for Digital Innovation (MAIDI),

a non profit organization providing free data for development as part of its open science advocacy.

What's a human in a world of machines ? This question has been extracted from the interview of Joy Buolamwini, Founder of the Algorithmic Justice League within the latest Responsible Guide Tech by All Tech is Human (ATIH) in 2024. With the thriving race in Artificial Intelligence adoption, the task of answering this question should appear as a top priority and a matter of urgency for all stakeholders involved in shaping global digital framework, one of which are we « *the People* ». However, answering this question depends on perspective—whether we focus solely on AI's technical capabilities that inspire awe or also consider the underlying factors driving its performance. Still, do we see ourselves as AI shapers or victims of its uncontrolled power ?

The « *AI and Equality* » course was a perfect training to develop one's mindset, enabling critical reflexes when designing data collection processes and digital products. I present here my key takeaways from the training from both an implementer and digital consumer's point of view.

As an AI training data producer, the course has expanded my understanding of the ethical considerations necessary to ensure that the data produced is not harmful, remains inclusive, and accurately embodies both the "AI for Development" concept and the broader "AI for All" initiative. I realise that the digital divide extends beyond just digital skills— for instance: how do we explain to people that data is needed to create an algorithm, and that a computer (which they may have only seen but never used) can, for instance, detect early-stage diseases?

And after informing the people from whom we have gathered data that they can modify it on a given platform—upholding citizen data rights as part of a decentralized data initiative—we must ask: If all of these steps are for the sake of consent and AI accountability, is inclusivity then without understanding fake? How do we respond to ignorance while still calling it "progress"? Even though inclusion is crucial, the challenge of balancing meaningful representation with the need to keep AI advancing remains a complex practicality, and I don't think we've figured it out yet.

As an AI user and citizen observer, I do know how far my data can be used and reused. However, being considered in AI development through co-creation principle doesn't necessarily mean that I benefit from it. Still, should I only participate in developing digital products that I will use? This reminds of the WeBuildAi collective participatory framework for Algorithmic Governance. regarding people's rights to retract or modify their data, which means that an algorithm will never be the same every day.

Throughout the course, my reflection focused on shared examples of bias, leading me to question racial algorithmic bias and what truly defines the "right data." If we take the case of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) in the United States, which has flagged Black men as high-risk for recidivism—how do we determine when data is truly unbiased? Does participation in decision-making make it so? Would we consider data unbiased only if the algorithm produced the opposite result or maintained a strict 50/50 balance as a non-discrimination pattern? If a predictive algorithm gave us a result we wouldn't agree with, that would mean we don't need AI at all.

Today, one third of the global population is still offline and other forms of digital divide still exist between the two thirds that have online access. As many AI products were built for us, the forecast investment in the sector is expected to reach 632 billion USD by 2028. It is clear that AI will strengthen the digital gap we are trying to fill, but its rhythm is moving at a different pace than our efforts for unbiased, accessible, and fair technologies. The challenge that remains is to ensure that AI is not just an innovation for the privileged but a tool that can empower everyone— one that takes into consideration current levels of digital access and understanding. For instance, the lack of African language annotation can lead slowly to a loss of cultural identity. AI development appears to be a one-way street where companies are thinking for us without consulting society.

In a world of machines, a human is better off being an architect of progress rather than a passive observer. As the world rushes toward AI-driven futures, the real challenge is ensuring that technology does not just serve its designers but also those who bear its outcomes without even knowing what's going on. Artificial intelligence risks reinforcing the inequality it says would reduce. Beyond the environmental impacts debates, would we have the courage to pause Artificial Intelligence for a moment before all the world is ready to embrace it ?

Artificial Intelligence and Human Rights: Equality, Non-Discrimination, Rule of Law, and Accountability

Maria Rite Miele, Italy
A Law graduate student at the University of Palermo. Maria holds an MA in Blockchain & Digital Assets and participated in a leadership development program with Europe101.

Artificial intelligence (AI) is radically transforming society, significantly affecting economic, social, and legal dynamics. On the one hand, it offers extraordinary opportunities for innovation and efficiency; on the other, it raises important ethical and legal issues, especially in relation to fundamental human rights. The main risk associated with AI is that, without proper regulation, it can exacerbate existing inequalities and undermine fundamental principles such as non-discrimination, transparency, and legal accountability. This essay explores the delicate balance between the potential of AI and the need to ensure respect for human rights, focusing on four pivotal principles: equality, non-discrimination, accountability, and the rule of law.

The Rule of Law and the Challenges of Artificial Intelligence

One of the pillars of any democracy is the principle of the rule of law, which guarantees the equality of all before the law, the independence of the judiciary, and the protection of fundamental rights. However, the increasingly widespread use of AI in administrative and judicial decisions is putting these principles to the test. Algorithms are being used to determine prisoners' risk of recidivism, detect fraud in social benefits, or even rank citizens based on their behavior, as is the case in China's social credit system.

The problem is that these systems, while based on seemingly objective data, can contain errors and biases. The case of the COMPAS software, used in the United States to assess the dangerousness of defendants, is emblematic: several studies have shown that it tends to rank ethnic minority individuals more harshly than white individuals, all other factors being equal. This openly contradicts the principle of equality enshrined in the Universal Declaration of Human Rights.

AI and Algorithmic Discrimination, A Real Risk

One of the most problematic aspects of AI is so-called algorithmic discrimination, which occurs when an automated system replicates and amplifies existing biases in training data. A striking example was the use of recruitment algorithms by large companies such as Amazon, which unknowingly penalized female candidates because of a database trained on historical data reflecting a male-dominated labor sector (Dastin, 2018)

Additionally, research has shown that facial recognition systems are significantly less accurate in recognizing the faces of ethnic minority individuals. This problem was highlighted by Buolamwini and Gebru's study (2018), which demonstrated that such systems make errors

more frequently when analyzing non-Caucasian subjects. The risk is that such technologies, instead of eliminating inequalities, end up further consolidating them.

The Need for Accountability and Transparency

A crucial aspect of AI regulation concerns the accountability of algorithmic decisions. When software makes a wrong or unfair decision, who is responsible? This question is central, especially considering that many artificial intelligence algorithms, particularly those based on deep learning, operate according to logic that even their developers do not fully understand (Knight, 2019).

A particularly interesting case study involves predictive policing systems (Hilka, 2023; Angwin et al, 2016) which analyze historical data to identify areas of increased crime risk. However, these tools often rely on past arrest data, which may reflect pre-existing social and racial biases. This leads to a vicious cycle whereby certain urban areas, inhabited primarily by minorities, are policed more intensively, increasing the number of arrests and thus reinforcing the initial bias.

For AI to be compatible with democratic principles, it is crucial to ensure transparency in its decision-making mechanisms. This means making public the criteria used by algorithms, introducing independent checks to verify the absence of discrimination, and, most importantly, providing redress mechanisms for those who suffer from unfair decisions made by AI (Binns, 2018)

A Step Forward with The European AI Regulation

The European Union has recognized the risks posed by the unregulated use of artificial intelligence and introduced a regulatory framework with the AI Act, groundbreaking legislation that ranks AI systems according to their level of risk to human rights. Among the most important measures under this legislation are the ban on the use of high-risk AI systems, such as facial recognition in public places, and mandatory transparency for algorithms used in critical areas such as healthcare, justice, and welfare.

This regulation is a significant step forward in ensuring that AI is used ethically and in a way that respects the fundamental principles of democracy. However, continuous monitoring remains essential to prevent technological innovation from resulting in new forms of discrimination and social control.

Artificial intelligence offers tremendous opportunities, but it also poses complex ethical and legal challenges. The main lesson we can draw is the need to take a critical and vigilant approach to this technology. In our daily lives, both professionally and academically, we can apply these principles by promoting transparency, fairness, and accountability in automated decisions.

References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) 'Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks', ProPublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Binns, R. (2018) 'Fairness in machine learning: Lessons from political philosophy', Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 149–159. Available at: https://proceedings.mlr.press/v81/binns18a.html (Accessed: 8 February 2025).

Buolamwini, J. and Gebru, T. (2018) 'Gender shades: Intersectional accuracy disparities in commercial gender classification', Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 77–91. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed: 9 February 2025).

Dastin, J. (2018) 'Amazon scraps secret AI recruiting tool that showed bias against women', Reuters. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Knight, W. (2019) 'The Apple Card didn't "see" gender—and that's the problem', MIT Technology Review. Available at: https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Zilka, G. et al. (2023) 'Predictive policing and algorithmic bias: A critical review', Journal of Law & Technology, vol. 35, no. 2, pp. 123–145.

Biographies of Contributors

Fabienne RAFIDIHARINIRINA is a former freelance journalist and has been an open data

activist since 2016. She founded MAdagascar Initiatives for Digital Innovation (MAIDI), a non profit organization providing free data for development as part of its open science advocacy. With the rise of AI demand, the organization started to create AI trained data in 2023 in the sector of Climate and Energy which marked the starting point of a truly new strategy within the organization.

Duarte Silva is a final-year Master's student in Mathematics and Computer Science, with a previous Bachelor in Applied Statistics. He has a strong interest in the societal impact of AI. His research focuses on AI ethics and governance, particularly fairness, trustworthiness, and privacy. He explores ways to mitigate harmful outcomes while ensuring transparency and accountability. His key interests include aligning AI with human rights principles and addressing bias, privacy concerns, and discrimination in automated systems. He is a member of Data Science Portugal (DSP), the Portuguese Statistical Society (SPE), Statisticians without Borders (SWB) and a mentee of All Tech is Human. Previously, he served as a #BanUnpaidInternships Ambassador for the European Youth Forum.

Sadia Tabassum is a UMSAILS Scholar who recently completed her Master of Laws (LL.M) at the University of Asia Pacific, Dhaka, in affiliation with the UNESCO Madanjeet Singh South Asian Institute of Advanced Legal and Human Rights Studies (UMSAILS). She earned her Bachelor of Laws (LL.B. Hons.) with distinction from Brac University. Sadia's academic interests focus on human rights, constitutional law, and artificial intelligence.

Maria Rita Miele is a Law graduate student at the University of Palermo, nearing the completion of her academic journey. Despite a significant health challenge, a tumor, she remained determined and focused on achieving her dreams. She enriched her education with international experiences at the University of Katowice in Poland, specializing in international law, alternative dispute resolution, and international property law. She also completed a Master's in Blockchain & Digital Assets and a leadership development program with Europe101. Maria Rita is actively involved in organizations such as Erasmus Student Network Italy, One Hour for Europe, and ELSA Italy, where she developed leadership skills and worked effectively in multicultural contexts. She is an Ambassador for the Nexus Public Policy Institute, where she actively promotes a fairer, more inclusive, and sustainable society, contributing to a just future. Aiming for a diplomatic career, she aspires to contribute to justice, global cooperation, and human rights advocacy within international organizations.

Sezen Yalcin is a humanitarian and development professional with expertise in child protection, community-based protection, and gender-based violence in emergency settings. She has managed multi-sectoral humanitarian programs in complex crises. With a background in humanitarian assistance and human rights, Sezen also has experience in gender equality and LGBTIQ+ rights advocacy at both national and international levels.

Chandrashekar Konda is a seasoned technology leader with over 10+ years of experience in designing and implementing cutting-edge Gen AI & predictive AI models in various industries including across healthcare, transportation, and tech industries. He holds a Master's degree in Data Science from the University of San Francisco and another Master's degree in Industrial Engineering & Operations Research from Indian Institute of Technology, Bombay. Chandra brings AI expertise to the table through his leadership roles and technical achievements. He has architected end-to-end AI solutions and led cross-functional teams to develop cutting-edge AI solutions, including Generative AI, Predictive AI, Deep Learning, and Causal Inference models to drive business growth. Outside of work, Chandra contributes to review AI research publications and books, staying at the forefront of this rapidly evolving field.

Linda-Lotta Luhtala is an experienced professional specializing in the intersection of technology, innovation, and gender equality. Served as the focal point for the global Generation Equality initiative at UN Women Finland for nearly four years, with a strong focus on technology-facilitated gender-based violence and youth advocacy. She also has experience in innovation support and supporting early-stage, tech-enabled startups in Southern Africa.

Emaediong Akpan is a Nigerian legal practitioner, gender equity advocate, and researcher specializing in the intersections of technology, human rights, and social justice. She holds a Master's in Development Studies (Women and Gender Studies). Her research focuses on the societal impacts of artificial intelligence (AI) and emerging technologies, particularly in perpetuating technology-facilitated gender-based violence (TFGBV). Her interests lie in algorithmic injustice and how it perpetuates harm and exacerbates inequalities, disproportionately affecting marginalized groups such as women and sexual minorities.

Nahima Dávalos-Vázquez is currently a Ph.D. candidate in Anthropological Sciences at Universidad Autónoma Metropolitana, Mexico, researching how young women from the Global South appropriate digital technologies. Her interests include Digital Culture, Women and Technology, and the impact of Artificial Intelligence on labor in Latin America. She also works as an independent consultant, helping women in vulnerable situations adopt digital tools. She has held fellowships at renowned institutions in Berlin and is involved in several research groups, including IAMCR.

Claire Dugan works at the intersection of data security and responsible AI at Microsoft UK, ensuring AI systems are secure and trustworthy. Currently on maternity leave, she is helping her tiny human thrive while reflecting on the changing landscape of human-computer interactions and what this means for upcoming GenAI native generations. Claire holds a graduate degree in AI Ethics and Society from the University of Cambridge.

Samu Ngwenya-Tshuma has a passion for advocating for women's rights, addressing gender-based violence and leveraging technology for impactful initiatives. Her experience includes consulting for the United Nations Spotlight Initiative, United Nations Special Rapporteur on Extreme Poverty and Human Rights, and the International Labour Organization (ILO), with a focus on evidence-based policymaking. Samu's work emphasizes the intersection of gender, human rights, and social justice. Holding a Master's of Law from Peking University, she brings a wealth of knowledge to address some of our complex global issues focusing on tangible solutions for policy and advocacy.

Philani Mdingi is a seasoned technology leader and advocate for leveraging digital innovation to drive social change, particularly in the realm of human rights. As the Agency Principal for Tech for Good, he combines over 15 years of experience in technology, digital transformation, and advocacy to deliver impactful solutions for the development sector and NGOs worldwide. Philani previously served as the Director of Technology, Data & Innovation at All Out, a global human rights organisation based in New York. In this role, he was pivotal in implementing cutting-edge IT and cybersecurity strategies to safeguard digital activism and human rights campaigns. He led transformative projects that ensured the security and effectiveness of digital platforms, managed diverse teams to drive innovation, and developed robust policies to navigate the complex landscape of digital surveillance and data compliance.

Yumiko Kanemitsu is an expert in Data and Evidence, Research and Evaluation, and Policy Analysis, specializing in Innovation and Results-Based Management (RBM) within the international development and humanitarian sectors. With extensive experience in Food Security and Systems, Gender Equality, and Human Rights, she has led transformative initiatives that integrate evidence-based policymaking, inclusive development, and strategic evaluation. Her expertise in instructional design supports effective capacity-building and knowledge-sharing, strengthening institutional learning across global organizations. Through her work, Yumiko drives innovative, data-driven solutions that enhance policy, program effectiveness, and organizational learning.

Cinthya Leonor Vergara Silva is a Management Control Engineer with a Master's in Business Engineering, currently pursuing a PhD in Data Science at Universidad Adolfo Ibáñez working at Research Group for Trustworthy Systems (R3).

Ozge Caglar is a psychologist, humanitarian worker. She has recently finalized her master's research on integrating AI to humanitarian action. She conducts research and works as an independent consultant for multiple projects.

Tanya Marinkovic is the founder of METIS, a startup that uses AI to educate users on

identifying abusive behavior while addressing systemic inequalities.

Freyja van den Boom, PhD, is a transdisciplinary practice based postdoctoral researcher and speculative socio-legal designer who focuses on disruptive innovations impact in society and how through anticipatory governance, imagination and participation we can develop,deploy and use AI in ways that will truly benefit society and help us transform towards futures we want.

Eleonora Sironi is a researcher and consultant in gender, technology, and AI, applying data feminism to tech policies and algorithmic bias. With experience in NGOs like Doctors Without Borders and feminist start-ups like VIOLA, she is passionate about advancing social justice. She holds an MA in International Affairs (Hertie School) and a BA in International Studies (University of Milan).

Lesly Zerna is a tech educator. Currently, professor and curriculum developer. As a Telecommunications engineer with a master in computer science, she is part of the growing AI and tech community in Latin America.