PART 3: GLOBAL MAJORITY FACING AI

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

Elise Racine

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

Introduction

Artificial intelligence (AI) has emerged as a transformative force across sectors, offering immense potential to tackle complex global challenges (Vinuesa et al., 2020). But AI's use also raises pressing ethical concerns, including the possibility for algorithmic systems to amplify biases, reproduce injustices, and exacerbate global inequities (Ashar, Ginena, Cipollone, Barreto, & Cramer, 2024; Davis, Williams, & Yang, 2021; Igarapé Institute, 2024; Racine, 2024). These concerns are especially acute in the Global South, where "wicked" problems marked by resource constraints, infrastructure limitations, and unique socio-cultural considerations are more prevalent.

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

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

11.1. Challenges in Ensuring AI Benefits the Global Majority

The dominant discourse in Western technological spaces is one of hype, where the promise of AI to address global challenges and improve lives worldwide is emphasized (Crawford, 2021; Dežman, 2024). But in reality, these benefits are not equally distributed (Benjamin, 2019; Igarapé Institute, 2024; Mohamed, Png, & Isaac, 2020). Several key challenges prevent AI from serving the Global Majority. One of the most substantial challenges is the lack of Global Majority involvement throughout the AI lifecycle—even when directly impacted. This extends to the AI workforce itself (Okolo, 2023). Critical AI functions like data labeling and content moderation are routinely outsourced to the Global Majority, often subjecting workers to traumatic conditions and low pay (Igarapé Institute, 2024; Okolo, 2023; Perrigo, 2022, 2023).

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

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

Marginalized and minoritized groups, like sexual and gender minorities, are especially vulnerable to these epistemic impositions. For example, commercial facial-analysis technologies only return binary labels, entirely excluding non-binary/genderqueer individuals (Scheuerman et al., 2019). However, these Western labels (e.g., man/woman, homosexual/heterosexual) may be unsuitable in other cultural settings and repress invaluable diversity and complexity (Young & Meyer, 2005; Racine, 2023, 2024). This includes local self-determined identities that operate outside these binary categorizations, such as the *bacha bereesh* of Afghanistan, *hijra* of India, and *māhū* and *fa'afafine* of the Pacific Islands. Not only are these identities and rich histories erased when Western frameworks/norms are imposed, but such acts of epistemic violence can perpetuate significant, long-lasting harm.

This epistemic injustice is compounded by the concentration of AI power in the hands of a few tech giants. As of July 2024, 14 of the 15 largest AI companies by market cap were US-based, with the

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

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

11.2. Reparative Algorithmic Impact Assessments

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

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

to (a) consider how power flows through these systems and (b) place these developments within broader patterns of oppression, privilege, marginalization, and disadvantage (Johnson, 2021; Kalluri, 2020; Racine, 2024).

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

The key components of R-AIAs are:

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

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

11.3. Incorporating Decolonial, Intersectional Principles

To foster inclusive and equitable AI for the Global Majority, adopting decolonial, Intersectional principles in AI design, development, deployment, research, and governance is essential. Intersectionality, as introduced by Crenshaw (1991), recognizes that individuals experience overlapping forms of discrimination/oppression based on aspects of their identity, such as race/ethnicity, gender, class, sexual orientation, disability, and religion. A reparative praxis builds on this foundation by not only addressing these intersections but actively working to repair associated harms (Davis et al., 2021; Racine, 2024). By weaving Intersectionality into every step of assessment process and centering marginalized voices, R-AIAs aim to not only dismantle inequitable structures, but drive material benefits and systemic change for those most affected by algorithmic injustice.

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

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

1. Epistemological diversity: Actively incorporating diverse knowledge systems into AI development.
2. Data sovereignty: Respecting the rights of communities to control their data.

3. Technological self-determination: Empowering communities to develop and deploy AI that aligns with their values and needs.
4. Cultural preservation: Ensuring AI respects and promotes cultural diversity.
5. Reciprocity: Establishing mutually beneficial relationships between AI developers and communities.

11.4. From Principles to Practice: Implementing Decolonial R-AIAs

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

11.4.1. Socio-Historical Research

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

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

11.4.2. Participant Engagement and Impact/Harm Co-Construction

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

- *Reparative:* Utilize socio-historical research to inform participant recruitment. Implement mechanisms to navigate divergent values, iterate based on new knowledge, and alleviate burden for participants (e.g., offer compensation, mental health resources, flexible participation options). Make certain accessibility needs are met and participant feedback directly shapes chatbot functionality
- *Harmful:* Define and assess impacts based on superficial consultations with highprestige experts while neglecting input from affected communities (PAI, 2024).

11.4.3. Sovereign and Reparative Data Practices

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

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

11.4.4. Ongoing Monitoring and Adaptation

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

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

11.4.5. Redress

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

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

Concluding Remarks

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

References

Ada Lovelace Institute. (2021). Algorithmic impact assessment: A case study in healthcare. Ada Lovelace Institute. Retrieved from

https://www.adalovelaceinstitute.org/wp-content/uploads/2022/02/Algorithmicimpact-assessment-a-case-study-in-healthcare.pdf

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

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

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed, 17*(2), 389-409.

https://doi.org/10.2966/scrip.170220.389

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

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

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

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

In *Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '22)

(17 pages). ACM, New York, NY. https://doi.org/10.1145/3551624.3555290

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

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

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

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

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

Davis, J. L., Williams, A., & Yang, M. W. (2021). Algorithmic reparation. *Big Data &*

Society, 8(2), 20539517211044808. https://doi.org/10.1177/20539517211044808 Igarapé Institute. (2024). *Responsible Artificial Intelligence efforts in the Global South*. Igarapé Institute.

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

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

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

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

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp.

735-746). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445935

Miller, K. (2022, March 21). The movement to decolonize AI: Centering dignity over dependency. Stanford HAI. Retrieved from

https://hai.stanford.edu/news/movement-decolonize-ai-centering-dignity-overhttps://hai.stanford.edu/news/movement-decolonize-ai-centering-dignity-over-dependencydependency

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*(4),

659-684. https://doi.org/10.1007/s13347-020-00405-8

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

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

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

Perrigo, B. (2022, February 14). Inside Facebook's African sweatshop. *TIME*.

https://time.com/6147458/facebook-africa-content-moderation-employeehttps://time.com/6147458/facebook-africa-content-moderation-employee-treatment/treatment/

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

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

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

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

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

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

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

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

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

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

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

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

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

12. AI ETHICS FOR THE GLOBAL MAJORITY: LESSONS FROM DECOLONIAL FEMINIST BIOETHICS

ALICE RANGEL TEIXEIRA

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

INTRODUCTION

Artificial Intelligence (AI) has been celebrated for potentially empowering humans and stimulating human flourishing, from applications that save human time with automation in areas overloaded such as the judiciary system, healthcare and business, to the prevention, diagnostic and treatment of diseases, the reduction of poverty, inequalities in healthcare and education and gender-based discrimination (Baclic et al. 2020; Floridi et al. 2018; García-Micó & Laukyte 2023; Goralski & Tan 2023; Topol 2019). However, its applications are often reported to be biased and discriminatory, untrustworthy, harming individuals and marginalized groups, and reinforcing inequalities (López Belloso 2022; Mhlambi & Tiribelli 2023; Mohamed et al. 2020; Morondo Taramundi 2022; Ricaurte 2022). These ethical concerns and the recent developments on AI capabilities have stimulated the debate of AI ethics and the publication of several frameworks and guidelines from different sectors such as business, institutional and governmental (Floridi & Cowls 2019).

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

12.1. THE PRINCIPLE-BASED APPROACH TO AI ETHICS

Analyzing the global landscape of AI ethics guidelines, Jobin et al. (2019) identified over 11 principles across 84 documents, highlighting that the dominant approach to AI ethics is principle-based. The work also demonstrates how unclear these principles are, with each containing an abundance of codes. The principle of "justice & fairness", for example, includes 16 codes that present ill-defined or

broad terms such as reversibility, challenge, inclusion, and equity. As the authors observe, this diversity indicates divergences in how AI ethical challenges are addressed. Furthermore, because most guidelines come from the Global North, it raises concerns on how well-equipped these strategies are to deal with the global scenario of AI without neglecting the particularities of knowledge, needs and interests of underrepresented regions. Mhlambi and Tribelli (2023) observe that despite attending to several principles, guidelines tend to prioritize autonomy, perpetuating historical and abusive practices of racial and gender control and oppression.

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

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

12.2. CRITIQUES OF THE PRINCIPLE-BASED APPROACH

Because AI ethical frameworks are predominantly from Global North regions and neglect power asymmetries and systemic injustice, they allow the deployment of technologies that reinforce coloniality and the matrix of domination (Collins, 2000), contributing to the economic, social and epistemic oppression of marginalized social groups. This can be observed in the uneven effects of AI that disproportionately exploit human labor and natural resources of the Global South regions, while benefits disproportionately benefit regions of the Global North (Couldry & Mejias, 2019; Ricaurte, 2023; Van Dijck, 2014). Scholars focused on the effects of coloniality and neo-colonialism through data or computation colonialism, point to problems in the Western philosophical traditions that serve as foundations to the principle-based approach. Arguing that these traditions have served the interest of those in power as a tool for continuous oppression of coloniality. It is therefore necessary to consider a diversity of epistemologies instead of assuming a core shared value system (Mhlambi &

Tiribelli, 2023; Mohamed et al., 2020; Ricaurte, 2022; Valente & Grohmann, 2024). Feminist scholars approaching the subject of ethical AI from gender analysis have also pointed to the same problems. As they argue, because feminism has been historically attuning to power inequities, it presents itself as a framework that can be applied to AI, allowing the continuous examination of its power asymmetries as a method to prevent exacerbating oppression (Ciston, 2019; D'Ignazio & Klein, 2020; Hancox-Li & Kumar, 2021; Katell et al., 2020).

There has been less work on proposing moral and epistemological theories that could sustain these models, which can be problematic. By leaving untouched the philosophical assumptions behind a proposed framework, there is a risk of repeating the problems on how to interpret the concepts that underpin the main concerns of the field, contributing to the current scenario of confusion and ethical-washing. If the assumed interpretation is left undiscussed, it also risks favoring a specific standpoint while neglecting others. Finally, it closes the possibility of learning from similar fields by looking to the theories developed or strengthened by them. This can be observed, for instance, by the lack of cross-work between feminist and decolonial AI scholars, or the absence of feminist, racial and decolonial theoretical foundations, with a few exceptions such as the work of Ricaurte (2022) and Birhane (2021). These works greatly contribute to the cohesive critique of AI principle-based approaches that takes into account different systems of oppression, however they are less focused on discussing ethical theory. The complexity and potential impact of AI technology can be better understood through a careful examination of its risks and possibilities that can in turn guide the regulation and governance of the technology with a clear social-political direction (Floridi, 2018). Noteworthy, these fields have been an essential part of feminist bioethics' long standing criticism of the principle-based approach in bioethics and crucial contributors in feminist ethics (Rogers et al., 2022) that can serve as lessons to inform a decolonial feminist AI ethics.

12.3. FROM BIOETHICS TO AI ETHICS: LESSONS FROM FEMINIST BIOETHICS

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

12.3.1. FEMINIST EPISTEMOLOGY AND METHODOLOGY

Feminist bioethics critiques the methods and conclusions of scientific research by highlighting the relationship between knowledge and power. It challenges biased assumptions, such as those found in eugenics and the pathologization of women's bodies, or racial differences in pain threshold, arguing that abstract reasoning, detached from social context, is impossible (Ganguli-Mitra, 2022; Hutchison, 2022). Scholars like Patricia Hill Collins and Maria Lugones, building on Standpoint Theory and intersectionality, propose alternative models that prioritize those in the margins that are often oppressed by these bodies of knowledge (Collins, 2000; Hutchison, 2022; Lugones et al., 1983;

Stoetzler & Yuval-Davis, 2002). Others, influenced by postmodern feminist thought, claims the impossibility that true objectivity knowledge can ever be achieved (Anderson, 2024), they focus instead on the current situation that people actually face (Hutchison, 2022), prioritizing empirical research over normative judgment, but with a feminist methodology that takes into account the power structures that delineate any research, including the relationship between researcher and subjects (Scully, n.d.).

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

12.3.2. JUSTICE

Feminist critiques of justice in bioethics highlight the influence of Rawls's Theory of Justice, Utilitarianism, and Distributism. They criticize Rawls for prioritizing abstract principles over practical realities (Fourie, 2022), neglecting social injustices like gender and race (Jaggar, 2009)). Utilitarianism, focused on maximizing benefits, similarly overlooks what equality and well-being might mean (Scully et al., 2010), while distributism often emphasizes resource distribution without addressing non-quantifiable injustices, such as epistemic injustice—the devaluation of marginalized knowledge systems (Fricker, 2007). Current AI ethics, with its focus on resource distribution, leading to the under-analysis of structural injustice and how it relates to AI technology (Lin & Chen, 2022)). Instead, Iris Young's model of shared responsibility (Young, 2011), which addresses systemic oppression collectively, offers a more inclusive approach to justice in AI that does neglect systemic injustice and places responsibility in the collective (Lin & Chen, 2022), diverging from current discussions of AI's responsibility that are limited to the liability model.

12.3.3. NEO-LIBERAL AUTONOMY VS RELATIONAL AUTONOMY

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

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

12.3.4. UNIVERSALISM AND PARTICULARISM

Starting with a critique of the abstraction and generalization of the principle-based approach, feminist bioethics, intersecting gender with social identities such as race, ethnicity, and sexuality, has had to incorporate non-Western perspectives and accommodate these differences through its history of activism and self-reflection. This critique of universalism, combined with the adoption of a relational view of the self promotes a focus on the local and particular, seeking to extract context-specific experience to make explicit the social process in which gender and other differences are transformed into inequities (Marway & Widdows, 2015). AI ethics needs to draw from this same framework, concerning itself more with the particularities of moral life in its local applications instead of applying universalist models that invisibilize the differences in power.

12.3.5. CRISIS PROBLEMS AND MUNDANE PROBLEMS

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

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

12.4. CONCLUSION

This paper has demonstrated that mainstream AI ethics frameworks, dominated by Global North perspectives, follow a principle-based approach influenced by mainstream bioethics. A review of feminist bioethics provided insight on how principle-based ethics neglect power asymmetries and perpetuate systems of oppression. The alternative theories and methods developed in decolonial feminist bioethics offer significant advances beyond the principlism. Their emphasis on power relations, relational autonomy, shared responsibility, empirical evidence and local contexts creates opportunities for rethinking the ethical AI, offering distinctive tools for addressing structural injustice. By integrating insights from decolonial feminist bioethics, it is possible to build an AI ethics that challenges existing systems of oppression, centering decision-making on the needs and interests of the global majority.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ricaurte, P. (2023). Epistemologias de dados, colonialidade do poder e resistência. *Dispositiva*, *12*(22), 6–26. https://doi.org/10.5752/P.2237-9967.2023v12n22p6-26

Rogers, W. A., Scully, J. L., Carter, S. M., Entwistle, V. A., & Mills, C. (2022). Introduction. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Scully, J. L. (n.d.). Feminist Empirical Bioethics. In *Empirical Bioethics Theoretical and Practical Perspectives* (pp. 195–221). https://doi.org/10.1017/9781139939829.013

Scully, J. L., Baldwin-Ragaven, L. E., & Fitzpatrick, P. (2010). *Feminist Bioethics: At the Center, on the Margins*. Johns Hopkins University Press.

Stoetzler, M., & Yuval-Davis, N. (2002). Standpoint theory, situated knowledge and the situated imagination. *Feminist Theory*, *3*(3), 315–333. https://doi.org/10.1177/146470002762492024

Stoljar, N., & Mackenzie, C. (2022). Relational autonomy in feminist bioethics. In *The Routledge Handbook of Feminist Bioethics*. Routledge.

Tang, L., Li, J., & Fantus, S. (2023). Medical artificial intelligence ethics: A systematic review of empirical studies. *DIGITAL HEALTH*, *9*, 20552076231186064. https://doi.org/10.1177/20552076231186064

Tong, R. P. (2019). *Feminist Approaches to Bioethics: Theoretical Reflections and Practical Applications*. Routledge.

Valente, J. C. L., & Grohmann, R. (2024). Critical data studies with Latin America: Theorizing beyond data colonialism. *Big Data & Society*, *11*(1), 20539517241227875. https://doi.org/10.1177/20539517241227875

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

Young, I. M. (2011). Two Structure as the Subject of Justice. In I. M. Young & M. Nussbaum (Eds.), *Responsibility for Justice* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.003.0002

13. EXPLOITATION ALL THE WAY DOWN: CALLING OUT THE ROOT CAUSE OF BAD ONLINE EXPERIENCES FOR USERS OF THE "MAJORITY WORLD."

ZEERAK TALAT AND HELLINA HAILU NIGATU

Abstract. Global Majority users are exposed to multitudes of harm when interacting with online platforms. This essay illuminates how exploitation in the advances of Artificial Intelligence is tied to historical exploitation and how the use of blanket terminology overshadows the layers of exploitation and harm "Global Majority" populations face. It first discusses the multitude of harm content moderators from the Global Majority face, arguing against the current trend of protection through exploitation, then it illustrates the nuances and differences within the Global Majority, and finally, it outlines actionable items to move away from such harm.

INTRODUCTION

Global Majority users are disproportionately affected by the more extreme harms caused due to harmful content online. For instance, failures in moderation on Facebook have resulted in physical harm and escalation of violence in countries like Myanmar and Ethiopia (Akinwotu, 2021) the spread of misinformation on WhatsApp led to violent attacks on minorities in India (Samuels, E. 2020); and YouTube users from countries that do not have English as their primary language are at 60% higher rate of being exposed to content they will "regret" watching (McCrosky et. al. 2021). Such lackluster moderation and failure of automatic detection for the majority of the world's languages emboldens malicious content creators to post policy-violating videos (Nigatu et. al, 2024).

Platforms use a combination of automated systems and human moderators to moderate content (Roberts 2019). Generally, automated content moderation involves using trained machine learning models to determine if a post should be sanctioned due to breaches of policy, e.g., on hate speech and toxicity. However, not all users are protected equally (Dias Oliva, 2020). The field of natural language processing (NLP) has paid little attention to non-European languages, which has lead to a lack of data and technological resources to train robust automated detection systems. Moreover, platforms focus their efforts disproportionately on Western countries. For instance, in 2020 while 90% of its users live outside of the United States (US) and Canada, Meta (then Facebook) spent 87% of its time moderating posts in the US (Tworek, 2021). Such disparity is also reflected in moderation personnel: YouTube reports that 89.2% of its human moderators operate in English (Google, 2023), neglecting that 67% of videos are posted exclusively in languages other than English and 5% in multiple languages including English (Van Kessel et al, 2019).

The harm that speakers of the majority of the world's languages face in relation to content moderation extends beyond exposure to harmful content as users of online platforms. Big Tech companies hire content moderators from the Global Majority, which appears like an increased effort to protect users from those communities. However, these moderators often operate under deplorable working conditions and without fair compensation for conducting deeply traumatizing work (Perrigo, 2022). Such workers, who are often employed from African, South American, South East Asian, and South Asian countries, also provide labeled data for guardrails of Large Language Models like ChatGPT

(Perrigo, 2023), models which do not work well in languages spoken by the Global Majority (Ojo et al., 2023), or are entirely unavailable.

Understanding and implementing effective policy to protect users of Global Majority must begin by uncovering what lies beneath blanket terminology that serves to obscure nuances; starting with the term Global Majority. While the term has been adopted as a reclaiming of power by appealing to the number of people grouped under it, it is still a blanket term covering several geographies, hundreds of cultures, and thousands of languages whose common predicament is exploitation by the powers on the other side – a concern that remains unresolved by the adoption of the term. Prior work has demonstrated the cultural nuances that result in the under-moderation or over-moderation of online users from the "Global Majority" or "Global South" (Shahid et al, 2023). Hence, to effectively impact practical policies, we must start by examining these nuances and uncovering what is underneath the blanket terminologies.

In this essay, we first dive deeper into multitudes of harm faced by content moderators from the Majority world, reflecting on how the common denominator is exploitation. Then, we examine the current alternatives in online moderation which pose a false dichotomy for moderation to be effective, for which surveillance is an inevitable consequence. We call out the root problem that presents these alternatives as the only options. Next, we detail the social, political, and economic structures within the "Global Majority" to illustrate the nuances in different communities that would render blanket policies ineffective. Finally, we put forth a call to action to ensure the effective protection of "Global Majority" users on online platforms. We argue that what ties the experience of Global Majority people is the continued exploitation and disregard for well-being by Big Tech and states outside of the Global Majority, which bears similarities to exploitation by colonial bodies during the period of European colonization.

13.1. Discussion

13.1.1. The Cycle of Harm In Moderation and Inclusion

In 2021, Meta (then Facebook) faced scrutiny after a whistleblower, Frances Haugen, leaked internal documents detailing the harms the platform was fostering, in some cases not taking action to rectify the situation even after becoming aware of it (Horwitz, 2021). One trend in the moderation landscape has been to hire moderators in Global Majority countries, sometimes through third-party companies. However, the working conditions of the moderators are usually dire (Perrigo, 2022). While cases brought directly against companies like Microsoft and Meta have resulted in settlement payments and some policy changes for moderators hired directly by the companies (Newton, 2020), moderators hired by third-party companies risk mass layoffs and threats against forming unions (Perrigo, 2022). This double standard is a parallel to other exploitative work performed in "Global Majority" countries (e.g. the externalization of "Global Minority" pollution and trash to the "Global Majority" (Liboiron, 2021)), where workers are treated differently for the same work when it is performed in "Global Minority" countries. The exploitation does not stop there. Perhaps ironically, such moderators are hired to moderate OpenAI models like ChatGPT, which do not work for the African languages that they speak (Ojo et al, 2023). In fact, ChatGPT was not available in countries like Ethiopia until November 2023 (Shega, 2023). In this way, the labor of the "Global Majority" is extractive, and the conditions under which moderators work are for the benefit of the privileged few who can operate the internet in languages like English and Spanish.

Communities from the "Global Majority" are exposed to harm (1) while using the platforms, due to weak platform policy enforcement and limited performance of technologies used in the moderation pipeline; (2) while moderating harmful content by virtue of exposure to traumatic content; (3) through poor working conditions and exploited labor; and (4) through technologies that exploit their labor but leave out their whole communities from whatever benefit the technology might provide. At the center of this cycle of harm is the exploitation and neglect of the wide swath of communities. The current systems that sustain the digital landscape are an extension of the history of colonization and exploitation that have ravaged the "Global Majority" (Kwet, 2019). Even when these communities are included in Artificial Intelligence research, they are treated as "bottom billion petri dishes"(Sambasivan et al, 2021, p.320)–their diversity and the weak policies protecting them make them an attractive test-bed for evaluating model robustness with little-to-no consequence or cost.

13.1.2. False Dichotomies of Harm: Either you are surveilled or you are left in the trenches.

Communities that have largely been excluded from policy and technological advances in the moderation space are exposed to harmful content daily. These unmoderated harmful content could be due to (1) policies that exist but are not enforced properly for these communities, or (2) policies that do not exist since the design of policies takes place under contexts that do not account for the diverse realities of "Global Majority." When policies do exist and are under-enforced, malicious actors exploit the under-enforcement to propagate policy-violating content. As such, communities who have already been exploited by global structures are exploited again in our failure to effectively moderate online spaces.

When policies that reflect the diverse cultural context in the "Global Majority" simply do not exist, entire communities and cultures are left in a vacuum. Indeed, some companies seek to enforce a single standard upon all users, disregarding cultures, customs, and traditions. For instance, Facebook's one-size-fits-all approach resulted in the removal of a post of village kids swimming in a pond for violating the platform's policy against child nudity; although in the context of the poster, it is a common activity for children to swim naked in their local ponds to avoid "being scolded by their parents" (Shahid & Vashistha, 2023, p. 5).

With the rapid advances of Large Language Models and the "low-resource language" NLP community trying to increase the representation of these languages, harmful, toxic, and culturally nonrepresentative content on online spaces risks trickling down to model development and deployment. Generative models are trained using data from YouTube, Twitter, and general web scraping (Cole, 2024). However, training models for the majority of the world's languages present a particular risk as effective content moderation technologies and practices are not deployed for such languages. Thus, risks of harm are compounded by a lack of appropriate moderation, thereby compounding the risks of harm that have been documented for English (Talat et al. 2022).

Platforms that benefit from their users should adhere to their end of the bargain and provide a "positive experience for everyone on [their] platforms no matter where they [the users] are in the world" (Google, 2023, p. 8). Effective content moderation infrastructures, both human and automated, are required for safely building language technologies and content moderation technologies. However, many language technologies have risks of dual-use (Kaffee et al., 2023), including the risk of surveillance (Solaiman et

al. 2023). It is therefore particularly important to consider how technologies are deployed and used, in addition to how data is gathered for the technologies themselves.

Here we would like to pause and reflect on what exactly effective moderation is, especially in the current context of the moderation pipeline. If the premise of moderation was not capitalistic and exploitative, could we have safer online experiences that put the power in users and not in companies that are out for profit?

13.1.3. What Lies Under Blanket Terminologies?

The degree and type of harm communities from the Global Majority face are shaped by the social, political, and economic realities of each community. Take two YouTube users studied by Nigatu & Raji, (2024) who studied the experiences of Ethiopian women on YouTube: a migrant domestic worker and a software engineer in the United States. Both users are Ethiopians, women, and of the Global Majority; yet have completely different realities. Migrant domestic workers cross borders to countries like Qatar and Lebanon en masse, either legally or via human traffickers. Once there, most of these women are subject to inhumane treatment, and sexual harassment and are often left without access to legal or medical services (Diab et al., 2023). Nigatu & Raji, (2024) show how these migrant domestic workers are exposed to harm through exposure to graphic and sexual videos while seeking medical help on online platforms. On the other hand, the Ethiopian Software Engineer living in the US is exposed to the same policy-violating content as the migrant workers when they search in their language. That is, a shift of location does not indicate a shift in types of policy-violating content. Change in policy enforcement might, for instance, remove policy-violating posts that expose both sets of users to harm. However, removal would not satisfy the need for information from the migrant worker, in this case, medical advice.

Political responses of different countries towards platform policies, or failures of platform policies also vary drastically. Countries like Ethiopia, Somalia, and Sudan ban online platforms when policies do not align with their values or when policies do not protect citizens from violent content. However, this has little impact on the actual problem as users resort to VPN services to access the platforms. Additionally, representatives for these platforms are most often subject to regulatory scrutiny in Global Minority countries, even when the harms are primarily impacting people in the Global Majority. It is clear the platforms respond to the callouts by powerful governments; Europe has constantly been praised for the GDPR and its requirements against online harm to its citizens.

While the term “Global Majority” is an evolution from prior binaries based on social and economic status or geographic location (Khan et al., 2022), it is still a binary. The realities–and needs–of Indigenous and Aboriginal communities who continue to suffer the consequences of colonization and occupied land are different from those of African and Asian countries that faced the brunt of exploitation colonialism. Within the Global Majority several layers of class, ethnicity, and power result in the exploitation and harm of some communities over others. There is no single “AI from the Global Majority” because the “Global Majority” is many.

Call to Action: Throughout this essay, we have discussed the degree and depth of harm and exploitation that Global Majority users face. However, Global Majority users are not idly waiting for the mercy of the powers that be; to the extent that they can, they devise ways to protect themselves from

harm[128]. We can augment their efforts by designing interventions that support them and relying on methods like participatory design as we build AI tools. Additionally, members of the Global Majority face layers of barriers to entering academic and policy spaces at a Global scale (Septiandri et al., 2023). Those who do make it, ourselves included, have degrees of privilege not afforded to the many who are organizing on the ground. Hence, we are responsible for engaging with community organizations–to the degree they are interested–to connect the academic and policy space with community organizing.

Conclusion

The manifold of communities that the "Global Majority" encompasses makes it challenging to enforce one-size-fits-all policies. The harms members of these communities face vary across the diverse social, economic, and political axes each community has. Most of the current policies for protecting users in the digital age have been designed, tried, and tested in the "Global Minority" context. Our response to the fact that we have ignored the majority of the world's population in policy making and implementation should not be to blindly extend these policies to the communities we ignored. In moving from neglect to blind inclusion, we risk the exploitation of community members at several levels of the pipeline. Instead, we should focus our efforts on augmenting community efforts and building interventions that center community needs.

References

Akinwotu, E. (2021). Facebook's role in Myanmar and Ethiopia under new scrutiny. The Guardian. Retrieved from https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny

Cole, S. (2024). Leaked Documents Show Nvidia Scraping 'A Human Lifetime' of Videos Per Day to Train AI. 404 Media. https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project

Diab, J. L., Yimer, B., Birhanu, T., Kitoko, A., Gidey, A., & Ankrah, F. (2023). The gender dimensions of sexual violence against migrant domestic workers in post-2019 Lebanon. Front. Sociol., 36741584. Retrieved from https://pubmed.ncbi.nlm.nih.gov/36741584

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2021). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. Sexuality & Culture, 25(2), 700–732. doi: 10.1007/s12119-020-09790-w

Google. 2023. U Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report. Technical Report. https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf

Horwitz, J. The Facebook Files. (2021, October 01). The Wall Street Journal. Retrieved from https://www.wsj.com/articles/the-facebook-files-11631713039

[128] Instagram users create Fake-Instagram or "Finsta" accounts to share more intimate content with a close group of friends. A YouTube user in Nigatu & Raji (2024) study created multiple accounts for different aspects (religious, educational, and general) becuase she did not "want to be hit with disturbing content when [I] was watching a religious sermon or looking at a lecture."

Kaffee, L.-A., Arora, A., Talat, Z., & Augenstein, I. (2023). Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing. ACL Anthology, 13977–13998. doi: 10.18653/v1/2023.findings-emnlp.932

Khan, T., Abimbola, S., Kyobutungi, C., & Pai, M. (2022). How we classify countries and people—and why it matters. BMJ Global Health, 7(6). doi: 10.1136/bmjgh-2022-009704

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. Race & Class. doi: 10.1177/0306396818823172

McCrosky, J., Geurkink, B., Zawacki,K., Jay,A., Afoko, C., Gahntz, M., and Bennet, O..(2021) YouTube Regrets.https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

Newton, C. (2020). Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. Verge. Retrieved from https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

Nigatu, H. & Raji, I.D. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. | Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. (2024, June 05). https://doi/10.1145/3630106.3658546

Ojo, J., Ogueji, K., Stenetorp, P., & Adelani, D. I. (2023). How good are Large Language Models on African Languages? arXiv, 2311.07978. Retrieved from https://arxiv.org/abs/2311.07978v2

Perrigo, B. (2022). Facebook Faces New Lawsuit Alleging Human Trafficking and Union-Busting in Kenya. Time. Retrieved from https://time.com/6147458/facebook-africa-content-moderation-employee-treatment

Perrigo, B. (2023). Universities Are Wondering How to Adapt New Artificial Intelligence Tool ChatGPT. Time. Retrieved from https://time.com/6247678/openai-chatgpt-kenya-workers

Roberts, S.T. Behind the Screen. (2024, August 12). Retrieved from https://yalebooks.yale.edu/book/9780300261479/behind-the-screen

Sambasivan, N. & Arnesen, E. & Hutchinson, B. & Doshi, T. & Prabhakaran, V. Re-imagining Algorithmic Fairness in India and Beyond | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. (2021, March 01). https://doi/10.1145/3442188.3445896

Samuels, E. (2020). How misinformation on WhatsApp led to a mob killing in India. Washington Post. Retrieved from https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india

Septiandri, A.A., Constantinides, M., Tahaei, M., Quercia, D., 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT? In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 160–171. https://doi.org/10.1145/3593013.3593985

Shahid, F. & Vashistha, A. (2023) Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? | Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. (2024, July 01). Retrieved from https://doi/10.1145/3544548.3581538

Shega Team. (2023). ChatGPT Now Available in Ethiopia. https://shega.co/post/chatgpt-now-available-in-ethiopia

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., Daumé III, H., Dodge, J., Duan, I., Evans, E., Friedrich, F., Ghosh, A., Gohar, U., Hooker, S., Jernite, Y., Kalluri, R., Lusoli, A., Leidinger, A., Lin, M., Lin,

X., Luccioni, S., Mickel, J., Mitchell, M., Newman, J., Ovalle, A., Png, M.T, Singh, S., Strait, A., Struppek, L., Subramonian, A. (2023). Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv, 2306.05949. Retrieved from https://arxiv.org/abs/2306.05949v4

Talat, Z., Neveol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ...Van Der Wal, O. (2022). You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. ACL Anthology, 26–41. doi: 10.18653/v1/2022.bigscience-1.3 and Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv, 2403.00742. Retrieved from https://arxiv.org/abs/2403.00742v1

Tworek, H. (2021). Facebook's America-centrism Is Now Plain for All to See. Centre for International Governance Innovation. Retrieved from https://www.cigionline.org/articles/facebooks-america-centrism-is-now-plain-for-all-to-see

Van Kessel, P., Toor, S. and Smith, A. (2019). 1. Popular YouTube channels produced a vast amount of content, much of it in languages other than English. Pew Research Center. Retrieved from https://www.pewresearch.org/internet/2019/07/25/popular-youtube-channels-produced-a-vast-amount-of-content-much-of-it-in-languages-other-than-english/#:~:text=Meanwhile%2C%2067%25%20posted%20videos%20exclusively,the%20first%20week%20of%202019

14. Countering False Information: Policy Responses for the Global Majority in the Age of AI

Isha Suri and Shiva Kanwar

Abstract. False information including misinformation and disinformation is being recognized as a severe global risk anticipated over the coming years. Access to generative artificial intelligence (AI) has dramatically increased the capacity for creating and disseminating falsified information. This is further compounded by algorithmic promotion of divisive content and creation of filter bubbles, leading to a precarious environment. We analyse the role of AI in exacerbating the false information crisis, evaluate regulatory responses to false information across various jurisdictions, and propose strategic policy recommendations for the Global Majority to effectively counter the threats of misinformation and disinformation in the age of AI.
Keywords. Misinformation; Disinformation; False Information; Artificial Intelligence; Algorithms; Algorithmic Bias; Recommender Systems; Intermediary Liability; Platform Governance; Content Moderation

Introduction

The proliferation of false information poses a significant threat to societal cohesion and democratic integrity in the contemporary digital landscape. As open access to advanced technologies, particularly artificial intelligence (AI), becomes increasingly prevalent, the capacity for generating and disseminating falsified information has increased dramatically. Sophisticated AI models have democratized creation of synthetic content, including realistic images and videos, voice cloning and counterfeit websites, blurring lines between authentic and fabricated narratives. This phenomenon is further compounded by an erosion of trust in information sources and institutions, leading to a precarious environment where societal cohesion and legitimacy of electoral processes and governance are jeopardized.

In response, governments worldwide are implementing evolving regulatory frameworks to curb dissemination of false information online. These measures often grapple with the delicate balance between safeguarding free speech and mitigating risks associated with falsified information. Particularly in global majority nations, where the intersection of digital authoritarianism and false information may exacerbate political repression, the need for tailored policy responses becomes imperative. This essay seeks to analyse the role of AI in exacerbating the false information crisis, evaluate regulatory responses to falsified information across various jurisdictions, and propose strategic policy recommendations enabling the Global Majority to counter pervasive threats of false information in the AI age.

World Economic Forum's Global Risks Report 2024 recognises false information as the most severe global risk anticipated over the next two years.[129] As such, different jurisdictions have been grappling with this menace and its ability to undermine democratic ideals for the past decade, albeit with limited success. Today, a handful of dominant technology firms are largely responsible for how users traverse

[129] World Economic Forum. (2024, January). The Global Risks Report 2024. weforum.org. Retrieved from https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf; Ezrachi, A., & Stucke, M. E. (2022). How Big-Tech Barons Smash Innovation—and How to Strike Back. Harper Collins.

the internet. Acting as gatekeepers, these firms control access to digital markets,[130] including internet-based communication services. Social media companies operate in multi-sided markets, as intermediaries for distinct user groups. Predominantly, on one side they interact with users accessing the platform for generating content (the 'free side' of the market), and on the other, they sell placement for digital advertisers.[131] Therefore, advertising-led business models dependent on massive data collection, profiling, and personalisation are a major source of revenue for social media platforms.[132]Research suggests that toxic and fabricated content is likely to be more engaging, with one study reporting that disinformation was likely to spread six times faster than the truth.[133] And recent research demonstrates that AI-generated disinformation may be more convincing than human-generated disinformation.[134] Multiple studies have demonstrated that social media is designed to reward and amplify divisive content, hate speech and disinformation.[135] With algorithms designed to maximise user engagement, content likely to trigger user attention is amplified including extreme content and content that contributes to formation of filter bubbles.[136] For instance, an internal study by Facebook revealed that its News Feed algorithms exploit the human brain's attraction to divisiveness and if left unchecked it would feed users "more and more divisive content to gain user attention and time over platform".[137] Similarly, employees at Google sought to improve issues pertaining to filter bubbles and enhance diversity of content by modifying YouTube's recommendation algorithm. However, it reduced viewer retention (which would eventually reduce advertising income) because of which the change was suspended.[138] Therefore, owing to their integrated structures, and profit maximising incentives, platforms continue to employ algorithms that recommend divisive content.

14.1. Regulatory Responses

Regulatory responses to tackle false information can vary, For instance, some nations have proposed or enacted laws specifically targeting false information such as the European Union's Digital Services Act (DSA), while others ground their proposed amendments or legal frameworks on existing legislation,

---

[130] Khan, L. M. (2019). The Separation of Platforms and Commerce. *Columbia Law Review*, *119*(4). Retrieved from https://columbialawreview.org/content/the-separation-of-platforms-and-commerce

[131] Stasi, M. L. (2023). Social media markets: A pro-competitive approach to free speech challenges. [Doctoral Thesis, Tilburg University].

[132] Article19. (2023, January). *Taming Big Tech: A pro-competitive solution to protect free expression*. Retrieved from https://www.article19.org/wp-content/uploads/2023/02/Taming-big-tech-UPDATE-Jan2023-P05.pdf

[133] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. Retrieved from https://www.science.org/doi/10.1126/science.aap9559

[134]Williams, R. (2023, June 28). Humans may be more likely to believe disinformation generated by AI. MIT Technology Review. Retrieved October 21, 2024, from https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/

[135] O'Carroll, T., Elsayed-Ali, S. (2024, August 20). Musk is the symptom, Big Tech is the crisis. Retrieved from https://www.linkedin.com/pulse/musk-symptom-big-tech-crisis-sherif-elsayed-ali-wppge/?trackingId=ZoQKBSHBQ4GP6PxMlTygHQ%3D%3D

[136] Ezrachi, A., & Stucke, M. E. (2022). How Big-Tech Barons Smash Innovation—and How to Strike Back. Harper Collins.

[137] Ibid.

[138] Ibid

including penal codes, civil law, electoral law, or cybersecurity law.[139] These regulatory frameworks either aim to hold perpetrators accountable as purveyors of false information or transfer the obligation to internet communication corporations to oversee or eliminate specific types of content.[140] The paper discusses DSA since it is the only regulation on recommender systems, and is therefore helpful in addressing various nuances involved in regulating algorithmic recommender systems. Germany's Network Enforcement Act is also discussed in an effort to contrast its intermediary liability framework with India, as outlined in the case study.

Ascribing criminal liability, particularly in cases where false information is defined broadly, carries significant risks of censorship[141]. For instance, Malaysia passed the Malaysia Anti-Fake News Act which criminalised the publication and dissemination of false news, punishable by up to six years in jail and a fine of $128,000. The law which was repealed in December 2019, made online service providers responsible for third-party content on their platforms.[142] Sri Lanka also amended its penal code in 2019 to prohibit fake news and hate speech that is "harmful to harmony between nations and national security" and enabled prosecution for spreading false statements or hate speech.[143] This law has been criticised for its potential to stifle free speech, usher in censorship, and facilitate mass surveillance.[144]

Electoral regulations have also been used to combat false information. In the run-up to the 2018 general elections, Brazil introduced several draft bills criminalizing electoral misinformation with penalties ranging from fines to imprisonment for crimes ranging from spreading fake news stories on social media to publishing inaccurate press accounts.[145] In 2019, Brazil amended its electoral code to define the crime of "slanderous denunciation for electoral purpose", with a penalty of two to eight years of imprisonment.[146]

More recently, instances of AI deepfakes being used to manipulate political narratives and public opinion have been reported in countries including Moldova, Slovakia, and Bangladesh.[147] And, deepfakes pose a grave threat to democratic processes with consequences such as voter confusion

---

[139] UNESCO, International Telecommunication Union, & Broadband Commission for Sustainable Development. (2020). Balancing Act: Countering Digital Disinformation while Respecting Freedom of Expression: Broadband Commission Research Report on 'Freedom of Expression and Addressing Disinformation on the Internet'. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379015

[140] Ibid

[141] Ibid

[142] Poynter. (n.d.). A guide to anti-misinformation actions around the world. Retrieved August 25, 2024, from https://www.poynter.org/ifcn/anti-misinformation-actions/#malaysia

[143] Poynter. (n.d.). A guide to anti-misinformation actions around the world. Retrieved August 25, 2024, from https://www.poynter.org/ifcn/anti-misinformation-actions/#sl

[144] Schiffrin, A, Cunliffe-Jones, P. (2022). Online Misinformation: Policy Lessons from the Global South. In H. Wasserman, D. Madrid-Morales (Ed.). *Disinformation in the Global South*. (pp. 161-178). USA: Wiley-Blackwell.

[145] Poynter. (n.d.). A guide to anti-misinformation actions around the world. Retrieved August 25, 2024, from https://www.poynter.org/ifcn/anti-misinformation-actions/#brazil

[146] UNESCO, International Telecommunication Union, & Broadband Commission for Sustainable Development. (2020). Balancing Act: Countering Digital Disinformation while Respecting Freedom of Expression: Broadband Commission Research Report on 'Freedom of Expression and Addressing Disinformation on the Internet'. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379015

[147] Swenson, A., & Chan, K. (2024, March 14). Election disinformation takes a big leap with AI being used to deceive worldwide. Retrieved from https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd

and manipulation.[148] While measures such as labelling AI-generated content are being developed to combat deepfakes, they are ineffective in preventing the spread of false information.[149]

Legislative proposals have also sought to tackle this issue through intermediary liability for online platforms regarding false information or hate speech. Germany's Network Enforcement Act, 2017 mandates that for-profit social media platforms with over two million registered users are required to act against hate speech and offences outlined in the German criminal code. Such entities are required to implement transparent procedures for reporting content and managing complaints, and remove/block "manifestly unlawful" content within 24 hours and "unlawful" content within 7 days.[150] Countries have also established specialized task forces to monitor and investigate false information campaigns. In 2018, Indonesia established the National Cyber and Encryption Agency intending to assist intelligence agencies and law enforcement to combat online misinformation and hoaxes in anticipation of nationwide regional elections, although the specific authorities granted to this agency remain ambiguous.[151]

The European Union's DSA introduces due diligence and transparency obligations regarding algorithmic decision-making by online platforms. It applies to all "intermediaries"[152] providing services in the EU and deems platforms and search engines with over 45 million monthly users in the EU as Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs).[153] Failure to comply with any obligation under the DSA can result in a fine of up to 6 per cent of the annual worldwide turnover in the preceding financial year.[154] The DSA mandates enhanced transparency in recommender systems and advertising by requiring intermediary service providers to disclose their content moderation tools and algorithmic decision-making processes in their terms and conditions.[155]

Apart from regulatory responses, fact-checking, especially on social networks, is also being used to counter false information. While published fact-checks provide people with an authoritative source of information, they often receive fewer shares on social media than the mis/disinformation they aim to debunk.[156]

---

[148] Ibid
[149] Ibid
[150] UNESCO, International Telecommunication Union, & Broadband Commission for Sustainable Development. (2020). Balancing Act: Countering Digital Disinformation while Respecting Freedom of Expression: Broadband Commission Research Report on 'Freedom of Expression and Addressing Disinformation on the Internet'. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379015
[151] Poynter. (n.d.). A guide to anti-misinformation actions around the world. Retrieved August 25, 2024, fromhttps://www.poynter.org/ifcn/anti-misinformation-actions/#indonesia
[152] Intermediary here includes social media platforms, search engines, online marketplaces, and internet service providers. See Baker, G. (2024, April 4). The EU Digital Services Act: A Win for Transparency. Retrieved from https://freedomhouse.org/article/eu-digital-services-act-win-transparency
[153] European Commission. (2023, April 25). Press Release - Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines. Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413
[154] Article 52 Digital Services Act
[155] Article 14(1) Digital Services Act
[156] UNESCO, International Telecommunication Union, & Broadband Commission for Sustainable Development. (2020). Balancing Act: Countering Digital Disinformation while Respecting Freedom of Expression: Broadband Commission Research Report on 'Freedom of Expression and Addressing Disinformation on the Internet'. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379015

Meta has the only large-scale international "third-party verification" programme among the dominant technology companies. Launched after the 2016 US presidential elections, the programme collaborates with independent fact-checking organizations, to assess accuracy of information on Facebook, Instagram and WhatsApp.[157] Fact-checkers are compensated by Meta. However, there is lack of transparency regarding payments made to the third-party fact-checking collaborators and ambiguity around the initiative's effectiveness in curbing the spread of false information.[158] An increasing reliance on a system where more content is flagged initially by Meta's AI tools raises concerns about potential algorithmic errors, and concerns about Meta developing AI tools based on data acquired through partnerships from this programme.[159]

Fact-checking initiatives also face added challenges in the Global Majority with low digital literacy, lack of connectivity, and rural-urban and gender divides affecting efficacy. Multilingual societies also result in misinformation in regional languages being 'ignored'.[160] Emerging research suggests that falsified information manifests differently across the globe, necessitating a nuanced and contextual approach to addressing the problem. For instance, during the COVID-19 pandemic, while India accounted for 16 per cent of global misinformation, the nature of content differed from that in the West. In the West, anti-vaccine related fake information gained traction, however, in India the myths ranged from using home remedies for treatment of COVID-19, thereby requiring distinct tactics from regulators and advocacy groups.[161]

14.2. India: Case Study

India relies predominantly on the Information Technology Act, 2000 (IT Act) and the recently amended Information Technology Rules to curb false information related harms in the country. Provisions such as Sections 69-A of the IT Act enable State Authorities to send content takedown orders to intermediaries whenever they find it "necessary or expedient" for national security, integrity, friendly relations with foreign states, and prevention of offences related to these grounds.[162] Any intermediary failing to comply with such an order is liable to pay a fine and/or face imprisonment for up to seven years.[163] While corresponding blocking rules provide a framework for implementation of the law, experience suggests that the rules cause an excessive restriction on freedom of speech and expression. For instance, research has highlighted that content creators are rarely notified or afforded a

[157] Meta. (n.d.). Meta's Third-Party Fact-Checking Program. Retrieved from https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking
[158] UNESCO, International Telecommunication Union, & Broadband Commission for Sustainable Development. (2020). Balancing Act: Countering Digital Disinformation while Respecting Freedom of Expression: Broadband Commission Research Report on 'Freedom of Expression and Addressing Disinformation on the Internet'. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000379015
[159] Ibid
[160] Ugwa, J., & Jain, M. (2023, December 18). Big tech 'failing' to curb fake news in global South. Retrieved from https://www.scidev.net/global/scidev-net-investigates/big-tech-failing-to-curb-fake-news-in-global-south/
[161] Baybars Orsek. (2023, June 5), How India can show the way in combatting fake news in the Global South. Retrieved from https://indianexpress.com/article/opinion/columns/india-show-way-combatting-fake-news-global-south-8646961/.
[162] Section 69A(1), Information Technology Act, 2000.
[163] Section 69A(3), Information Technology Act, 2000.

hearing.[164] Furthermore, the blocking rules also require confidentiality, effectively preventing content creators from viewing or challenging orders issued under them.[165] Recently X also challenged the blocking rules in the Karnataka High Court arguing, inter alia, that blocking orders did not contain reasons recorded in writing and were not communicated to users, consequently preventing users from effectively challenging them. X also claimed that directions of the Ministry of Electronics and Information Technology (MeitY) to block entire accounts rather than specific tweets were disproportionate and excessive. However, a Single Judge Bench of the Court rejected X's arguments and dismissed their challenge and imposed a penalty of Rs. 50 lakh (US$59,655).[166] An appeal against the order is currently pending before a Division Bench of the High Court.[167]

Furthermore, empirical evidence from India confirms that online blocking solely at the discretion of the executive has far-reaching effects on freedom of expression.[168] If an intermediary is legally obligated to respond to an overwhelming number of content takedown requests under the fear of losing its legal immunity, they are likely to over-comply to avoid sanction. This has the potential to chill online expression.[169]

Indian law has no regulations dealing with algorithms used by intermediaries and their potential harms, thereby limiting its ability to effectively counter AI-fuelled false information. Although reports suggest that the proposed Digital India Act will have provisions pertaining to algorithmic accountability, there remains ambiguity around the legislation and its timelines for implementation.[170]

---

[164] Gupta, A. (2015, March 27). But what about Section 69A? *The Indian Express*. Retrieved from https://indianexpress.com; Sakar, T., & Grover, G. (2020, February 15). How India is using its Information Technology Act to arbitrarily take down online content. *Scroll.in*. Retrieved from https://scroll.in/article/953146/how-india-is-using-its-information-technology-act-to-arbitrarily-take-down-online-content

[165] Mukhopadhyay, D. (2019, December 16). Delhi HC issues notice to the government for blocking satirical Dowry Calculator website. Retrieved from https://internetfreedom.in/delhi-hc-issues-notice-to-the-government-for-blocking-satirical-dowry-calculator-website/

[166] *X Corp* v. *Union of India*, W.P. No. 13710 of 2022, Karnataka High Court, Order dated 30-06-2023

[167] Karnataka High Court stays order imposing Rs 50 lakh fine on X Corp. (2023, August 10). *Business Standard*. Retrieved from https://www.business-standard.com

[168] Sehgal, D., & Grover, G. (2023, April). *Online Censorship: Perspectives From Content Creators and Comparative Law on Section 69A of the Information Technology Act*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4404965

[169] See Dara, R. (2011). Intermediary Liability in India: Chilling Effects on Free Expression on the Internet. *Centre for Internet & Society*. Retrieved from https://cis-india.org/internet-governance/intermediary-liability-in-india.pdf

[170] Ministry of Electronics and Information Technology. (2023). Proposed Digital India Act, 2023. Retrieved from https://www.meity.gov.in/writereaddata/files/DIA_Presentation%2009.03.2023%20Final.pdf

Table 1. A Brief Outlook on Regulatory Responses to False Information

Country	Instrument/Response	Status	Criminal Sanctions	Intermediary Liability	Transparency and Accountability Provisions
Argentina	Bill on creating a Commission for the Verification of Fake News 2018	Not passed	✓	✓	
Bangladesh	Digital Security Act 2018	In force	✓		
Brazil	Draft Bills 2018	Not passed	✓		
Brazil	Amendment to electoral code 2019	Passed by Congress	✓		
Egypt	Regulating the Press and Media 2018	In force	✓		
Egypt	Anti-Cybercrime Law 2018	In force	✓	✓	
Egypt	Penal Code	In force	✓		
European Union (EU)	Digital Services Act 2022	In force	✓	✓	✓
Germany	Network Enforcement Act 2017	In force	✓		✓
India	Information Technology Act, 2000 Information Technology Rules, 2009	In force	✓	✓	
Kenya	Computer Misuse and Cybercrimes Act 2018	In force	✓		
Malaysia	Anti-Fake News Act 2018	Repealed	✓	✓	
Nigeria	Anti-social Media Bill; Originally titled Protection from Internet Falsehood and Manipulation Bill 2019	Not passed	✓	✓	
Singapore	Protection from Online Falsehoods and Manipulation Act, 2019	In force	✓	✓	✓
Sri Lanka	Amended Penal Code 2019	In force	✓		

Source: Compiled by Authors

14.3. Conclusion and Way Forward

Countering the menace of fake information poses significant challenges for policymakers worldwide. And the advent of generative AI tools has further exacerbated the problem. Amongst other things, it requires treading a fine balance between restricting harmful speech without violating fundamental rights to free speech and expression. However, it will require efforts from all stakeholders including platforms, policymakers, and regulators to effectively address these threats. Based on our research, we recommend:

Consider unbundling platforms: The internet today is characterised by large social media platforms controlling the flow of information and communication between users. These dominant platforms rely on advertising as a source of revenue, with most of them being the largest providers of online advertising services, thereby creating a conflict of interest that requires intervention by regulators.[171] To

[171] Stasi, M. L. (2023). Social media markets: A pro-competitive approach to free speech challenges. [Doctoral Thesis, Tilburg University].

counter this problem, regulators should consider unbundling content curation services[172] (excluding content moderation services) from content hosting services within a platform.[173] Given that large social media platforms are global in nature, decisions taken in one jurisdiction are likely to have a spillover effect in another. For example, in Germany, as part of a remedy to respond to competition concerns from third-party sellers, Amazon agreed to amend its terms of business for sellers on Amazon's online marketplaces across Europe, North America, and Asia.[174]Such cross-country benefits could be further leveraged by promoting international cooperation between antitrust authorities across jurisdictions. This could also create opportunities to shift revenue models away from advertising and disincentivise promoting user engagement with divisive content.

Adopting a co-regulatory approach: Regulations alone will struggle to eliminate false information from digital platforms; a comprehensive strategy involving efficient regulatory interventions, along with self-regulation by platforms is required. A co-regulatory response involving government and platforms working together is likely to achieve better results. Through such a collaboration, regulators could gain access to information on algorithmic recommender systems, and make better decisions on how to shape their design to achieve desired outcomes. A co-regulatory response tailored to each jurisdiction's requirements is also likely to make enforcement easier.[175]

Develop inclusive AI-assisted tools for content moderation: Automated hate speech detection systems that have shown success in English and European languages struggled in countries such as Myanmar, India, and Ethiopia, due to lack of cultural contextualisation.[176] Miscreants evade keyword-based machine detection through astute combinations of words, misspellings, satire, changing syntax and coded language.[177] Platforms must develop AI tools using diverse high-quality data sets, and employ local teams proficient in the local language and cultural context.

Global majority countries must consider domestic realities before succumbing to the Brussels effect: The DSA introduces due diligence and transparency obligations regarding algorithmic decision-making by online platforms that complement other EU AI regulatory efforts such as the AI Act.[178]

---

[172] This paper the term "content curation" has been defined as the measures taken by social media platforms that affect the availability, visibility and accessibility of content, such as ranking, promotion, demotion. These measures are performed by fully or partially automated systems based on algorithms. Content curation differs from content moderation, which usually indicates the activities undertaken by social media platforms to detect, identify and address illegal content or content incompatible with their terms and conditions, such as demotion and removal.

[173] Stasi, M.L. (2021). Unbundling hosting and content curation on social media platforms: between opportunities and challenges. UCLA Journal of Law and Technology, 28 (2).

[174] Amazon in deal with German watchdog to overhaul marketplace terms. (2019, July 17). CNBC. Retrieved October 21, 2024, from https://www.cnbc.com/2019/07/17/amazon-in-deal-with-german-watchdog-to-overhaul-marketplace-terms.html

[175] Ibid.

[176] Udupa, S., Maronikolakis, A., Schütze, H., & Wisiorek, A. (2022, June). *thical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence*. The Shorenstein Center on Media, Politics and Public Policy. Retrieved from https://shorensteincenter.org/wp-content/uploads/2022/06/Ethical-Scaling.pdf

[177] Ibid.

[178] Chander, A. (2023). When the Digital Services Act Goes Global. *Berkeley Technology Law Journal*, 38(3), 1067–1088. https://doi.org/10.15779/Z38RX93F48

Adoption of such regulatory provisions without accounting for contextual realities, especially by authoritarian regimes and fragile democracies can leave nations vulnerable to potential misuse.[179] It could also incentivize platforms to adopt uniform content moderation policies that align with European standards, which, while promoting global consistency, may inadvertently suppress local norms and practices in global majority countries, resulting in over-censorship.[180] Furthermore, emulating complex regulations such as the DSA may pose challenges for developing countries, which often lack administrative and judicial capacities required for effective implementation, thereby increasing the risk of inconsistent application and exploitation by powerful entities. While comparative regulatory analysis is helpful, countries should tailor these regulations through studies grounded in their jurisdictions and also enhance regulatory capacity, where required.

Promote transparency in content recommender systems: Most regulatory and legislative responses focus on content moderation from the lens of eliminating potentially harmful user-generated content without addressing how individual pieces of content achieve high impact through recommender systems. Prioritizing transparency in recommender systems is essential to tackle harms that arise from algorithms promoting divisive content. It enhances comprehension of algorithmic decisions, fosters trust, and alleviates bias and privacy concerns while ensuring compliance with ethical AI standards.

[179] Ibid
[180] Ibid

15. Addressing the Challenges of AI Content Detection in the Global South

Richard Ngamita

Abstract. The rapid adoption of artificial intelligence (AI) in content creation has raised significant content moderation challenges, particularly in the Global South, where 'cheapfakes'—manipulated media created with basic tools—pose a serious threat. Existing detection systems, primarily designed for deepfakes, are inadequate for cheapfakes, which exploit low-tech environments to spread misinformation. To address this, initiatives must focus on developing AI models trained on local data, enhancing research and development, and implementing inclusive content moderation policies. These efforts protect civic participation and democracy in the Global South.

Introduction

The widespread use of artificial intelligence (AI) in content creation has posed significant challenges for content moderation, particularly in the Global South. While much attention has been given to detecting deepfakes, there is growing concern about the more common threat of 'cheapfakes'—AI-manipulated media created using basic editing tools. These cheapfakes can have serious consequences in regions with limited technological infrastructure, where misinformation or disinformation can easily incite violence and political instability. Current detection mechanisms, primarily designed for deepfakes, are insufficient for identifying cheapfakes, which include manipulated audio and video created with minimal resources. These types of content can be easily spread across social media platforms, making them difficult to detect and regulate (Paris & Donovan, 2019).

In 2023, Chinese smartphone brands such as iTel, Infinix, Huawei, and Tecno captured a 48% market share in Africa (Statista, 2023). While these devices have made digital technology more accessible, they often produce low-quality video content. This presents a challenge for automated detection systems, which may mistakenly flag these videos as fake, not due to manipulation, but simply because of their inherently poor quality. This issue highlights the limitations of current detection technologies, which are often ill-equipped to consider the context in which content is created and consumed, especially in the Global South.

Adding to this complexity is the significant geopolitical influence of China, which plays a significant role in shaping Africa's technological landscape. China's strategic economic and political engagement in Africa has facilitated the widespread adoption of its smartphone brands. While these devices are affordable and provide much-needed access to technology, they have raised concerns about surveillance and propaganda. Chinese technology companies, often influenced by state directives, may embed software that enables data tracking and collection. This duality—affordable access alongside the potential for digital surveillance—complicates the benefits of these smartphones, particularly in terms of privacy and control over information flow.

The issue is further compounded by the infrastructural challenges faced by countries in the Global South. Limited access to high-speed internet and reliance on low-end smartphones result in a higher prevalence of low-quality content. Videos created under these conditions are often flagged as suspicious by AI detection tools—not because they have been tampered with, but because poor video quality is mistakenly linked to inauthenticity. This not only leads to false identifications but also

undermines the credibility of legitimate content from these regions. Thus, the interplay of technological limitations, geopolitical influences, and infrastructure challenges creates a precarious digital environment, where access to technology can both empower and marginalize.

While the Global South faces challenges related to infrastructure and cheapfakes, the Global North contends with the more sophisticated threat of deepfakes. Politically motivated deepfakes have increasingly been used to manipulate public opinion. For example, a recent instance in the U.S. involved a fake voice message falsely claiming to be from President Joe Biden, which was sent to voters in New Hampshire during the primary election to discourage voting. Although the nature of manipulated media varies between the Global North and South, the dangers remain significant in both contexts—cheapfakes in the South, given their ease of creation, and deepfakes in the North, due to their technical complexity (Lewandowsky et al., 2012).

For example, a poorly edited video showing a political figure endorsing a controversial policy could spread quickly, especially in places with limited access to reliable news sources. A cheapfake featured Donald Trump endorsing Umkhonto we Sizwe (MK) and encouraging South Africans to vote for the party. Another involved an AI-generated video of Joe Biden falsely claiming that if the ANC won the election, the USA would impose sanctions on South Africa. Additionally, a manipulated image of Julius Malema of the Economic Freedom Fighters (EFF) appeared to show him crying after a perceived political defeat.

15.1. Results

Platforms like Meta, YouTube, and TikTok have introduced content moderation guidelines to address manipulated media, but these measures are largely focused on deepfakes. For example, Meta's manipulated media policy applies primarily to deepfakes, while YouTube's misinformation policy targets content that poses a risk of egregious harm (Meta, 2023; YouTube, 2023). TikTok prohibits AI-generated realistic scenes of fake people unless labelled by the creator. However, these policies inadequately address the proliferation of cheapfakes, which pose a more immediate threat to civic participation in the Global South.

Another challenge for detecting and moderating AI-generated content in the Global South is the region’s linguistic and cultural diversity. Many AI detection tools are trained on datasets that primarily consist of content in English or other widely spoken languages. This limits the effectiveness of these tools in detecting manipulated content in languages underrepresented in training data, leading to gaps in detection capabilities across different regions.

Moreover, the Global South's socio-political context presents additional content moderation challenges. Over 70% of the world's population lives under authoritarian regimes, primarily in low- and middle-income countries (Freedom House, 2023). In these environments, the disclosure of AI-generated content or the identity of the content creator could lead to severe repercussions, including imprisonment or worse.

15.2. Recommendations

To effectively address these challenges, several initiatives can be proposed to enhance AI research and development focused on content detection in the Global South to address these challenges. One key approach is developing AI models trained on local data. This would involve collecting and

annotating large datasets of content from the Global South, including texts, images, videos, and audio in local languages and dialects. By training AI models on this data, detection tools would be better equipped to recognize the nuances of manipulated content in these regions.

Collaborative efforts between local governments, academic institutions, and international organizations are essential to support research and development in this area. Funding should be directed towards building the necessary infrastructure for data collection and analysis, as well as for training local researchers and developers. This would not only improve the detection of AI-generated content but also empower local communities to participate in the global conversation on AI ethics and regulation.

One such initiative is Thraets, a company that is actively involved in combating the spread of AI-generated misinformation and disinformation, particularly in Africa. Through initiatives like the 'Safeguarding African Elections' project, Thraets is working to develop open-source AI tracking tools and knowledge hubs that focus on monitoring AI-generated content related to elections. This is particularly significant in regions where the proliferation of cheapfakes—manipulated media created with basic tools—poses a threat to civic participation and democratic processes (Thraets, 2024). Thraets also trains journalists and civil society organizations to detect and counter AI-generated disinformation. This capacity-building effort is especially crucial in regions where resources and expertise are often limited, and where the impact of misinformation can be particularly destabilizing. Thraets' efforts represent a significant step forward in the fight against AI-generated disinformation in the Global South.

An important initiative can be the development of clearer and more inclusive content moderation policies by social media platforms. These policies should explicitly address the issue of cheapfakes and outline specific measures for detecting and mitigating their spread. Platforms should also invest in tools that allow users to report suspected manipulated content and provide clear guidelines on how this content will be reviewed and acted upon.

It's important to prioritize raising awareness about the dangers of manipulated media in the Global South. Educational campaigns should aim to improve digital literacy and critical thinking skills among the population to reduce the impact of misinformation. These campaigns should be conducted in local languages and customized to the specific cultural contexts of different regions.

Conclusion

To effectively combat the issue of cheapfakes and ensure digital inclusivity, it is particularly essential to develop AI models trained on local data, support research and development initiatives, and implement clearer and more inclusive content moderation policies. We can better protect the citizens of the Global South from the harmful effects of manipulated media and ensure that they can participate fully in the digital age by taking these steps.

References

Freedom House. (2023). *Freedom in the World 2023*. Retrieved from https://freedomhouse.org/report/freedom-world/2023/global-2023

Paris, B., & Donovan, J. (2019). *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. Data & Society.* Retrieved from Data & Society — Deepfakes and Cheap Fakes (datasociety.net)

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106-131. Retrieved from https://doi.org/10.1177/1529100612451018

Meta. (2023). *Community Standards on Manipulated Media*. Retrieved from https://www.facebook.com/communitystandards/manipulated_media

Statista. (2023). *Smartphone market share in Africa in 2023*. Retrieved from https://www.statista.com/statistics/1171017/smartphone-market-share-by-vendor-in-africa/

Thraets. (2024). *Thraets secures grant to protect African elections from AI-generated mis/disinformation*. Retrieved from Thraets.

YouTube. (2023). *Misinformation policy*. Retrieved from https://support.google.com/youtube/answer/10834785?hl=en

16. Bridging the gap between the North and South in the governance of dual-use artificial intelligence technologies

Guangyu Qiao-Franco and Mahmoud Javadi

Abstract. This article examines the complex challenges of regulating dual-use artificial intelligence (AI) technologies within international arms control frameworks, amid a growing divide between the Global North and Global South. The intangible nature and dual-use potential of AI make traditional monitoring, verification, and classification methods ineffective. Developed nations are integrating civilian AI research into defense applications and imposing strict access controls to maintain military advantages, which exacerbates geopolitical tensions and stifles global innovation. In contrast, many Global South countries, unable to match these technological advancements, advocate for outright bans on autonomous weapons systems to mitigate their disadvantages. This dynamic undermines global cooperation and increases the risk of interstate conflict. The article advocates for a paradigm shift toward inclusive AI governance that addresses the needs and aspirations of both developed and developing nations. By fostering international dialogue, capacity building, and equitable access to AI technologies, it proposes establishing a transparent, multilateral framework for responsible AI use to bridge the North-South divide, reduce tensions, and promote global security and prosperity.
Keywords. Artificial Intelligence, Dual-Use Technologies, North-South Divide, AI Governance

Introduction

Artificial intelligence (AI), a key driver of economic growth, holds significant implications for international peace and security. In the early 2010s, concerns about the autonomous use of force enabled by AI prompted intergovernmental negotiations on arms control under the United Nations Convention on Certain Conventional Weapons (CCW). These discussions have revealed a growing divide between the Global North and the Global South regarding the military use of AI and regulatory approaches. Over a decade later, this gap appears to be widening rather than closing.

The dual-use nature of AI, which allows for both civilian and military applications, further complicates the path to a comprehensive arms control agreement. Developed countries are increasingly integrating civilian research and development (R&D) into defence, raising concerns about the military use of dual-use technologies by adversaries. This has led to stricter access controls, such as the United States tightening semiconductor export restrictions to China, supported by Japan and the Netherlands (Allen, et al., 2023). These restrictions permeate the civilian domain and raise security concerns.

In response, many Global South countries, unable to develop AI weapons, have opted for an outright ban on the use of autonomous systems to offset their technological disadvantage (Bode & Qiao-Franco, 2024). Meanwhile, emerging economies have taken a more rigid stance in military AI governance due to fears that broader control measures might be imposed under the guise of national security. For instance, China's unexpected abstention on a UN General Assembly resolution concerning lethal autonomous weapons systems (LAWS) in 2023 contradicted its earlier support for a legal ban on LAWS at the UNCCW. This has contributed to growing distrust and tension between states, undermining efforts to build confidence and coordinate on AI governance, while increasing the likelihood of extreme responses that could trigger interstate conflict.

To prevent this negative trajectory, a stepwise paradigm shift is needed in arms control regarding dual-use AI. Measures must account for the needs of both the Global South and Global North. International dialogue and partnerships should be fostered to promote capacity building, knowledge transfer, and inclusivity. These initiatives would help create an incentive structure encouraging responsible AI use and broader engagement. Ultimately, whether AI is used for peaceful or military purposes is determined by social factors. A new arms control paradigm should address the current insecurity dynamic, reduce the push for rival states to accelerate civil-military technology transfers, and pave the way for a 'global AI order' (Kissinger & Allison, 2023).

This article outlines the inherent challenges of controlling dual-use technologies and emphasises the different economic conditions and aspirations of the Global North and Global South. It concludes by proposing measures for achieving a harmonised approach to AI security regulation, aiming to build an inclusive arms control regime for AI safety.

16.1. Intricacies and Challenges of Arms Controls for Dual-Use AI

AI is an intangible technology, unlike other tangible and recognisable technologies, making traditional restrictive measures less, if not entirely, applicable and effective. Three main reasons justify this challenge. First, the intangible nature of AI software enables effortless cross-border transfer, circumventing monitoring by enforcement agencies (Brockmann, 2022). Unlike physical goods, AI algorithms can be transmitted digitally across borders with little to no physical trace, making it difficult for authorities to track and regulate their movement effectively.

Secondly, the verification of AI capabilities is complex due to the extensive lines of code involved, rendering it challenging for enforcement agencies to assess (Kaur et al., 2023). Unlike conventional technologies where physical characteristics can be examined, AI systems often consist of intricate algorithms with millions of lines of code, making it daunting to verify their functionalities, especially when those functionalities could have both benign and harmful applications.

In addition to the monitoring and verification challenges, AI is increasingly provided as a service rather than a standalone product, complicating export controls and oversight of its use across multiple countries (Klein & Patrick, 2024). With the rise of cloud computing and Software-as-a-Service (SaaS) models, AI capabilities can be accessed remotely, blurring the lines of jurisdiction and making it challenging for regulators to enforce compliance with arms control and usage restrictions (Cespedes & van der Kooij, 2023).

The dual-use nature of AI introduces another layer of hurdles in classification and regulation. Unlike other revolutionary technologies, whose progress relies heavily on government investments, AI technologies are propelled forward by private actors, ranging from technologists and entrepreneurs to corporations (Perifanis & Kitsios, 2023). Restrictive measures will likely pose significant risks to global commerce and can provoke dissent among private sectors reliant on overseas markets.

States, primarily from the Global North, have developed national frameworks and transnational regimes—such as the Wassenaar Arrangement and the Australia Group—to maintain control lists for dual-use items. However, the composition of these lists remains subjective and politically driven, largely due to the absence of international consensus on the definitions and scope of dual-use technologies (Benson & Putnam, 2023). In the absence of established criteria for controlling dual-use

AI within existing transnational regimes, these Global North states have increasingly asserted their authority by imposing restrictions on access to AI technologies, their components, and applications. Managing AI items on these lists is particularly challenging given the widespread use of general-purpose AI software. Excessive access controls designed to limit the export or use of AI technologies with dual-use potential risk stifling innovation, hindering economic growth, and unnecessarily escalating geopolitical tensions.

16.2. The Widening Gap between the Global North and Global South

While some emerging economies, such as China, India, and Turkey, are becoming leading technology innovators, most of the Global South, particularly the poorer regions, can only adopt AI technologies previously developed in the Global North. The significant technological gap has led to differing views on issues such as the adequacy of the existing legal framework to regulate autonomous weapons, the permissible forms of AI use in armed conflict, and measures to ensure human control, as discussed during the UNCCW negotiations over the regulation of military AI (Bode et al., 2023). While several Global South countries, in collaboration with a few small developed states and civil society, have succeeded in securing a mandate to negotiate a new legally binding instrument on lethal autonomous weapons at the UN General Assembly (UNGA, 2024), this instrument is unlikely to be endorsed by developed nations, which seek to modernise their armed forces to maintain a military advantage.

Instead of focusing their efforts on UN negotiations, several AI safety initiatives have emerged in the Global North, including those within the G7, OECD, NATO, and the EU. Other inclusive multilateral frameworks, led by countries such as the Netherlands, South Korea, the UK, and the US—such as REAIM (Government of the Netherlands, 2023), the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (U.S. Department of State, 2023), and the Bletchley Declarations (UK Prime Minister's Office, 2023)—have not been well received in the Global South. In contrast, Global South forums such as BRICS, ASEAN, and the African Union have primarily concentrated on AI's developmental potential, particularly its impact on the digital economy, with security concerns often receiving less attention (See e.g., Jin, 2024; ASEAN Secretariat, 2024; African Union, 2024). Consequently, the North-South divide in the global governance of dual-use AI technologies appears to be widening.

Programmes aimed at transferring technology to bridge capability gaps in various sectors have proven difficult to sustain in the field of AI, largely due to concerns in developed countries about the potential for malicious use. The Global North, particularly the United States, prioritises maintaining its qualitative edge in AI, often monopolising technology and securitising access to prevent its diffusion from civilian to military applications. Common measures include export controls, foreign investment reviews, and the suspension of R&D partnerships (Moller-Nielsen, 2024). Notable examples include US de-risking policies (The White House, 2023a), NATO's Defence Innovation Accelerator for the North Atlantic (DIANA) (NATO, 2024), and the Action Plan on Synergies between Civil, Defence, and Space Industries (European Economic and Social Committee, 2021).

The need for access control conflicts with the desire for rapid advancements in the Global South, especially among emerging economies subject to these stringent measures. China, for instance, has persistently sought to acquire and develop AI technologies, using them to advance various domestic and international agendas. The 2024 remarks delivered by Chinese Prime Minister Li Qiang at the

World Economic Forum highlight these diverging perspectives, criticising the restrictions on technology access and innovation while advocating for more open technological cooperation (WEF, 2024).

The restrictions on access to AI technologies, even in civilian domains, have exacerbated geopolitical tensions, diminishing the sense of security among states and making meaningful progress in cooperative military AI governance increasingly unlikely. The US's "chip war" with China offers a pertinent example. On 7 October 2022, the Biden administration issued new regulations (U.S. Department of Commerce, 2022) limiting US exports of advanced AI chips and Chinese acquisitions of companies capable of producing chips smaller than 14 nm. This was followed by an Executive Order in August 2023, establishing mechanisms to limit outbound investment in sectors such as semiconductors, quantum information, and AI in China and other designated countries of concern (The White House, 2023b). In response, the US undertook extensive efforts to dissuade countries in the Middle East and Africa from maintaining ties with Chinese technology companies.

China's reaction was swift: it imposed licensing requirements on the export of rare-earth metals, such as gallium and germanium, and their derivatives, which are essential for semiconductor manufacturing (Shivakumar et al., 2024). Additionally, following the restrictions from Washington and its allies, China has refocused its military-civil fusion-driven semiconductor investment policies to enhance state autonomy (Waldie, 2022). These policies have supported less competitive enterprises, facilitated the substitution of outdated foreign chips with domestically produced alternatives in critical military equipment, and allowed military-focused research to continue without fear of foreign embargoes. In a likely response to Western restrictions, China, the world's second-largest military spender, allocated an estimated €270 billion to its military in 2022, accounting for 13 per cent of global military spending. This represents a significant 63 per cent increase since 2013 and a 4.2 per cent rise from 2021 (Tian et al., 2023).

Although national measures like those adopted by Washington and Beijing—while not exclusive to these countries (Sterling, 2023)—aim to control access to dual-use AI, they risk reinforcing protectionism and isolationism, worsening global geopolitical dynamics rather than effectively managing dual-use AI regulation.

In addition to triggering a securitisation spiral that reduces both the Global North's and Global South's sense of security, this imbalanced regulatory approach may lead to a race to the bottom in AI safety standards. States may be incentivised to adopt more lax safety regulations to attract investment in AI industries, while competitive pressures could prompt AI producers to release products prematurely, sacrificing thorough testing and risk management.

16.3. Towards a Paradigm Shift for Dual-Use AI Governance

Governing dual-use technologies, particularly AI, necessitates a paradigm shift that reconsiders the multifaceted benefits and threats these technologies pose to nations across both the Global North and Global South. This shift involves identifying shared interests and common challenges to foster international collaboration and build consensus. Scholarly analyses and policy proposals (Kissinger & Allison, 2023; Reppy, 2006) emphasise the urgency of this approach, a sentiment echoed by the adoption of the United Nations' Global Digital Compact in September 2024 (Reiland, 2024).

AI's pervasive impact on various dimensions of human life—economic, social, and political—makes its governance especially critical. Implementing AI export controls and arms control mechanisms is vital to prevent the malicious proliferation of AI technologies that could compromise global security. However, when states exploit and weaponise AI against one another, it undermines efforts to establish a global AI governance framework essential for maximising benefits while minimising risks.

For developing nations, AI offers unprecedented opportunities for economic growth and social advancement. To realise these benefits, it is imperative that the Global South is actively included in global AI governance discussions. Inclusive policies are crucial to prevent the widening of the technological divide and to ensure that AI contributes to poverty eradication and sustainable development in less-developed regions.

To this end, Track Two and Track 1.5 diplomacy—facilitated by epistemic communities such as technologists, scientists, and industry leaders—provide promising avenues for initial engagement (Qiao-Franco, 2022). These non-governmental channels can foster mutual understanding, build trust, and promote informed discussions on managing dual-use AI technologies. This is particularly important in contexts where influential nations, such as the United States and China, may perceive each other antagonistically. By facilitating nuanced debates and identifying common ground, these communities can develop pragmatic solutions that balance national security concerns with the imperatives of innovation and competitiveness.

These efforts can lay the groundwork for an inclusive and transparent dual-use AI control framework within a multilateral setting, open to all states and viewpoints. Incorporating measures to bridge the North-South gap—such as technology transfer agreements, capacity-building initiatives, and equitable access to AI advancements—can promote understanding and trust between developed and developing nations.

The proposed framework should aim to fulfil the security needs of developed nations by preventing malicious AI use while simultaneously addressing the goals of developing nations for economic and social development. This includes supporting poverty eradication through AI-driven solutions in sectors like healthcare, education, and agriculture. By fostering a global AI order free from weaponisation and politicisation, AI can serve as a tool for global good rather than a source of conflict.

Ultimately, mitigating geopolitical tensions and enhancing global stability reduces the impetus to convert civilian technologies into military applications. By actively bridging the North-South gap and cultivating an inclusive international environment, the international community can harness AI's transformative power to promote global prosperity and security for all nations.

References

African Union. (2024). Continental artificial intelligence strategy. Retrieved October 21, 2024, from African Union website: https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy

Allen, G. C., Benson, E., & Putnam, M. (2023, April 10). Japan and the Netherlands announce plans for new export controls on semiconductor equipment. Retrieved October 21, 2024, from Center for Strategic and International Studies website: https://www.csis.org/analysis/japan-and-netherlands-announce-plans-new-export-controls-semiconductor-equipment

ASEAN Secretariat. (2024). ASEAN Guide on AI Governance and Ethics. Retrieved October 21, 2024, from ASEAN Secretariat website: https://asean.org/book/asean-guide-on-ai-governance-and-ethics/

Benson, E., & Putnam, M. (2023, April 11). Export controls and intangible goods. Retrieved October 21, 2024, from Center for Strategic and International Studies website: https://www.csis.org/analysis/export-controls-and-intangible-goods

Bode, I., Huelss, H., Nadibaidze, A., Qiao-Franco, G., & Watts, T. F. A. (2024). Algorithmic warfare: Taking stock of a research programme. *Global Society*, *38*(1), 1–23. https://doi.org/10.1080/13600826.2023.2263473

Bode, I., & Qiao-Franco, G. (2024). The geopolitics of AI in warfare: Contested conceptions of human control. In R. Paul, E. Carmel, & J. Cobbe (Eds.), *Handbook on Public Policy and Artificial Intelligence* (pp. 281–294). Cheltenham: Edward Elgar Publishing. https://doi.org/10.4337/9781803922171.00030

Brockmann, K. (2022). Applying export controls to AI: Current coverage and potential future controls. In T. Reinhold & N. Schörnig (Eds.), *Armament, Arms Control and Artificial Intelligence* (pp. 193–209). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11043-6_14

Cespedes, F. V., & van der Kooij, J. (2023, April 18). The rebirth of software as a service. *Harvard Business Review*. Retrieved from https://hbr.org/2023/04/the-rebirth-of-software-as-a-service

European Economic and Social Committee. (2021, March 28). Action Plan on synergies between civil, defence and space industries. Retrieved October 21, 2024, from European Economic and Social Committee website: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/action-plan-synergies-between-civil-defence-and-space-industries

Government of the Netherlands. (2023, February 13). REAIM 2023. Retrieved October 21, 2024, from Government of the Netherlands website: https://www.government.nl/ministries/ministry-of-foreign-affairs/activiteiten/reaim

Jin, Z. (2024, October 18). Tapping AI's potential. Retrieved October 21, 2024, from China Daily website: https://www.chinadailyhk.com/hk/article/595673#Tapping-AI%E2%80%99s-potential-2024-10-18

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Kissinger, H. A., & Allison, G. (2023, October 13). The path to AI arms control. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/united-states/henry-kissinger-path-artificial-intelligence-arms-control

Klein, E., & Patrick, S. (2024, March 21). Envisioning a global regime complex to govern artificial intelligence. Retrieved October 21, 2024, from Carnegie Endowment for International Peace website: https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en

Moller-Nielsen, T. (2024, January 25). EU reveals new economic security plan to resist “fierce” Chinese tech competition. Retrieved October 21, 2024, from Euractiv website: https://www.euractiv.com/section/economy-jobs/news/eu-reveals-new-economic-security-plan-to-resist-fierce-chinese-tech-competition/

NATO. (2024, July 5). Defence Innovation Accelerator for the North Atlantic (DIANA). Retrieved from NATO website: https://www.nato.int/cps/en/natohq/topics_216199.htm

Perifanis, N.-A., & Kitsios, F. (2023). Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review. *Information*, *14*(2), 85. https://doi.org/10.3390/info14020085

Qiao-Franco, G. (2022, May 25). Can track II dialogues be the new “ping-pong” diplomacy to thaw the sino-us relationship on military ai? - Autonorms. Retrieved October 21, 2024, from AutoNorms website: https://www.autonorms.eu/can-track-ii-dialogues-be-the-new-ping-pong-diplomacy-to-thaw-the-sino-us-relationship-on-military-ai/

Reiland, P. (2024, October 1). United nations: Global digital compact adopted by UN member states. Retrieved October 21, 2024, from Friedrich Naumann Foundation website: https://www.freiheit.org/human-rights-hub-geneva/global-digital-compact-adopted-un-member-states

Reppy, J. (2006). Managing dual-use technology in an age of uncertainty. *The Forum*, *4*(1), 000010220215408884 1116. https://doi.org/10.2202/1540-8884.1116

Shivakumar, S., Wessner, C., & Howell, T. (2024, February 21). Balancing the ledger: Export controls on u. S. Chip technology to China. Retrieved October 21, 2024, from Center for Strategic and International Studies website: https://www.csis.org/analysis/balancing-ledger-export-controls-us-chip-technology-china

Sterling, T. (2023, June 30). Dutch curb chip equipment exports, drawing Chinese ire. *Reuters*. Retrieved from https://www.reuters.com/technology/amid-us-pressure-dutch-announce-new-chip-equipment-export-rules-2023-06-30/

The White House. (2023a, April 27). Remarks by national security advisor Jake Sullivan on renewing American economic leadership at the Brookings institution. Retrieved October 21, 2024, from The White House website: https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/04/27/remarks-by-national-security-advisor-jake-sullivan-on-renewing-american-economic-leadership-at-the-brookings-institution/

The White House. (2023b, August 9). Executive order on addressing united states investments in certain national security technologies and products in countries of concern. Retrieved October 21, 2024, from The White House website: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/08/09/executive-order-on-addressing-united-states-investments-in-certain-national-security-technologies-and-products-in-countries-of-concern/

Tian, N., opes da Silva, D., Liang, X., Scarazzato, L., Béraud-Sudreau, L., & Assis, A. (2023). *Trends in world military expenditure, 2022*. Solna: SIPRI. Retrieved from SIPRI website: https://www.sipri.org/publications/2023/sipri-fact-sheets/trends-world-military-expenditure-2022

UK Prime Minister's Office. (2023, November 2). The Bletchley declaration by countries attending the ai safety summit. Retrieved October 21, 2024, from UK Prime Minister's Office website: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

UNGA. (2024). *Lethal autonomous weapons systems: Report of the secretary-general*. UNGA. Retrieved from https://docs-library.unoda.org/General_Assembly_First_Committee_-Seventy-Ninth_session_(2024)/A-79-88-LAWS.pdf

U.S. Department of Commerce. (2022, October 7). Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC). Retrieved from U.S. Department of Commerce website: https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3158-2022-10-07-bis-press-release-advanced-computing-and-semiconductor-manufacturing-controls-final/file

U.S. Department of State. (2023, November 9). Political declaration on responsible military use of artificial intelligence and autonomy. Retrieved October 21, 2024, from U.S. Department of State website: https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/

Waldie, B. (2022, April 1). How military-civil fusion steps up china's semiconductor industry. Retrieved October 21, 2024, from DigiChina website: https://digichina.stanford.edu/work/how-military-civil-fusion-helps-chinas-semiconductor-industry-step-up/

WEF. (2024, January 17). Davos 2024: Special Address by H.E. Li Qiang, Premier of the State Council of the People's Republic of China. Retrieved from World Economic Forum website: https://www.weforum.org/agenda/2024/01/li-qiang-china-special-address-davos-2024/