Media and Communication
2026 • Volume 14 • Article 10969
https://doi.org/10.17645/mac.10969

cogitatio

ARTICLE

Open Access Journal

Geographies of Hope: Rethinking Deepfake Harms and Gender AI Safety in the Global South

Weijie Huang [1], Payal Arora [1], and Marta Zarzycka [2]

[1] Department of Media and Cultural Studies, Utrecht University, The Netherlands
[2] Google, USA

Correspondence: Weijie Huang (w.huang2@uu.nl)

Submitted: 23 July 2025 **Accepted:** 20 November 2025 **Published:** 28 January 2026

Issue: This article is part of the issue "Digital Geographies of Hope: The Transformative Power of Media" edited by Cornelia Brantner (Karlstad University), Kaarina Nikunen (Tampere University), and Georgia Aitaki (Karlstad University), fully open access at https://doi.org/10.17645/mac.i501

Abstract

This article builds on the "geographies of hope" (Hazlewood et al., 2023) to better understand and address the gendered challenges posed by AI technologies in the Global South. AI-powered surveillance and technology-facilitated gender-based violence have reshaped digital geographies, leading to the rise of non-consensual synthetic intimate images—often called "deepfakes" or "deepfake pornography"—that disproportionately target women, LGBTQI+ communities, and racialized groups. These harms reveal the urgent need for inclusive AI safety and AI regulation frameworks that reflect the diversity of material and cultural geographies across the Global South. Through a cross-regional analysis of emerging AI safety policies in Asia, Africa, and Latin America, this article critiques the limitations of top-down, risk-based governance models and introduces a cross-cultural Gen(der) AI Safety framework rooted in decolonial and feminist praxis. Using critical discourse analysis, it identifies three systemic challenges—exclusionary legal-technical architectures, overreliance on individual responsibility, and entrenched power asymmetries. In response, the article proposes "geographies of hope" that emphasize localized, community-driven, and pleasure-positive interventions to counter digital harms. By centering intersectional and decolonial approaches, it calls for an AI safety agenda that affirms gender agency, collective joy, and justice.

Keywords

AI governance; decolonial AI; deepfakes; digital rights; feminist AI; gender AI safety; Global South; technology-facilitated violence

1. Introduction

In an era marked by digital technologies, datafication, algorithmic surveillance, and broader digital inequalities, new digital geographies of both harm and hope are emerging. AI is now a powerful socio-technical force reshaping social structures, digitally mediated governance, and daily life. The rise of AI-enabled tools is transforming societies worldwide—from restructuring gig economies in India (Bansal et al., 2024), to predictive welfare models in Denmark (Amnesty International, 2024), to diagnostic systems in Mexico's public health sector (Bandhakavi, 2024). However, AI is not neutral; it is embedded in spatial, cultural, and geopolitical contexts, shaped by global power dynamics, colonial legacies, and the socio-technical landscapes of its locations. Its widespread adoption can exacerbate existing social inequalities (UN, 2024b). In response, a countercurrent is emerging, grounded in justice-oriented interventions. For example, the UN Women AI School ("AI for gender equality," 2025) and Pivotal Ventures' gender-focused tech investments (Confino, 2024) reimagine AI from the margins by advancing gender-inclusive AI capacity building and equity-oriented design models.

AI safety refers to ensuring that AI technology is designed, developed, and deployed with sustainable reliability to mitigate harm, emphasizing technical safety and social impacts (Leslie, 2019). As AI becomes widespread, Fearnley et al. (2025) argue that the definition of AI safety must extend beyond physical harm to include the psychological harm caused by AI. For instance, technology-facilitated gender-based violence (TFGBV) is a broad term covering behaviors that use digital technology to promote gender-based harm (UN Women, 2022). AI can not only amplify the threat of gender-based violence, but also create new forms of violence, which have profound psychological and social consequences for victims (Dunn, 2021). Among them, non-consensual synthetic intimate images (NCSII), as a new type of TFGBV, severely invade the privacy and dignity of victims, particularly across platforms and regions with weak legal protections (Sheikh & Rogers, 2024).

Building upon data feminism (D'Ignazio & Klein, 2020), which emphasizes how data practices intersect with gender and power, this article frames AI safety as a practice rooted in relational care, accountability, and justice, extending beyond technical risk-management approaches. We further adopt De Sousa Santos' (2015) "pluriversal" epistemologies, emphasizing that Indigenous, feminist, and relational knowledge systems should guide AI governance, moving beyond the assumptions of Western liberal frameworks that prioritize individual autonomy, rationality, and universal norms. This study adopts a framework of harm, care, and hope: Harm reveals structural injustices in AI gender safety, while care and hope offer relational and political paths for responding, rebuilding, and imagining alternative futures (Hazlewood et al., 2023; Held, 2005; Tronto, 1993). From this decolonial perspective, the "Global South" is understood not as a homogeneous region, but as a political category. Following Spivak's (1988) notion of "strategic essentialism," we employ it as a solidaristic category, enabling countries with shared colonial legacies, weak institutions, and limited infrastructure to form strategic solidarities. This approach provides a cross-cultural comparative lens to understanding and addressing gender AI safety, fostering collaborative, justice-oriented practices that are sensitive to historically underrepresented contexts and communities.

Although some studies have explored TFGBV safety and governance in Kenya (Amatika-Omondi, 2022) and Pakistan (Batool et al., 2024), dominant AI safety frameworks remain shaped by Western ethical traditions. These frameworks often center the individual, rationality, and universal rules or outcomes, sidelining collective responsibility, relational obligations, ethical cosmologies, or community-led norms—central to many traditions

in the Global South, such as Ubuntu in Africa, Dharma in South Asia, or Confucian relational ethics in East Asia (Goffi, 2021; White et al., 2024). Therefore, global governance models often overlook the majority of digital users in Africa, Asia, and South America (Arora, 2024a).

Current AI governance often adopts a paternalistic stance, assuming Western values as universally applicable. This unidimensional normative output suppresses cultural diversity and inadequately addresses differences within Global South norms and values, including clashes between genders, states, and citizens, as well as enforcement capacities, thereby poorly integrating the needs of the Global South as technology users and governance stakeholders (Goffi, 2021). More critically, the existing framework generally lacks the collection of "gender data" (disaggregated data on women and girls, non-binary and other gender-diverse individuals), which undermines evidence-based policymaking and reinforces social inequality and systemic bias (Arora & Huang, 2025). Therefore, conceptualizing AI safety as a technical and regulatory issue risks obscuring its deep social, political, and geographically sensitive dimensions. To introduce hope as an analytical lens, this study draws on Hazlewood et al. (2023) to situate digital geographies as spaces where critical imagination informs transformative practice. We advocate repositioning digital geography as a site of hope, resistance, and rebuilding through transformative policy, challenging gender biases and injustices while recentering pleasure-positive approaches to digital intimacy.

This article examines emerging forms of TFGBV in the Global South, particularly NCSII in "deepfake" and "shallowfake" involuntary pornography. By analysing insufficient legal protections within cross-cultural and patriarchal systems, we highlight the urgency of addressing global gender AI safety and call for a contextualized, cross-cultural examination. Current AI safety discourse around deepfake technology predominantly frames harm as risk aversion, rather than considering gendered agency and justice (Fabuyi et al., 2024). The feminist perspective remains marginalized in the AI safety framework, with insufficient attention to the root causes of gender inequality (Arora & Huang, 2025).

Methodologically, this study adopts a critical discourse analysis approach, examining policy documents, feminist advocacy texts, and media coverage through a feminist and decolonial lens (Catalano & Waugh, 2020). This approach allows us to critically examine existing power structures (harm), explore ethical practices grounded in relationality, accountability, and justice (care), and identify actions and discourses that actively envision more equitable digital futures (hope).

To move beyond the dominant discourse on *harm* and to foster a vision of digital geographies guided by *care* and *hope*, this article proposes a Gen(der) AI Safety ABCDE Framework: Alliances, Beta-testing, Collective rights, Design justice, and Empowerment (see Table 1). This framework operationalizes a feminist and decolonial approach, and provides practical steps for policymakers, technologists, and other stakeholders in the Global South to develop gender-sensitive AI safety frameworks that address local contexts and protect vulnerable communities, contributing to a more equitable digital future for all.

Table 1. The ABCDE Framework for Gen(der) AI safety.

Framework element	Focus	Power shift	Core principle	Problem-solving
A: Alliances cross-cultural and cross-sectoral	Cross-sector, cross-cultural AI safety governance	From Western platform monopoly to local–global feminist coalitions	Power: Shifting control to those historically excluded	Centering affected communities in shaping rules and responses
B: Beta-testing approach to evidence-led policymaking	Participatory AI governance and empirical co-evaluation mechanisms	From elite lab models to lived-reality experimentation	Responsiveness: Iterative, community-driven policymaking	Co-developing tools and policy from the ground up with users
C: Collective rights and responsibilities	Local data sovereignty and group data justice	From individualist privacy to collective data governance	Justice: Centering group-based redress and recognition	Redressing harms by protecting group identities and community data
D: Design justice of affordances and constraints	Inclusive, transparent, explainable AI systems	From opaque engineering to co-created, accountable design	Inclusion: Designing AI systems that serve all, especially the marginalized	Embedding feminist values into AI architectures and interfaces
E: Empowerment and enablement-oriented	Survivor-centered, agency-enhancing tech policies	From harm reduction to flourishing and autonomy	Self-determination: Enabling individuals and communities to govern their own digital identities and futures	Creating binding obligations for platforms and states to protect users

2. Literature Review

Mainstream research on AI safety remains rooted in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) paradigms, often relying on abstract risk frameworks detached from the lived realities of users in the Global South (Arora, 2024a). This research takes a different approach. Grounded in feminist, interdisciplinary, and postcolonial studies, it understands technology as both a site of harm and hope. It centers the experiences of women and gender-diverse groups in the Global South, redefining gender AI safety as a pursuit of justice, accountability, and relational care. In these regions, patriarchy, gendered institutions, data extraction, and platform power are intertwined, creating unique digital vulnerabilities. By situating digital harms within their geopolitical and cultural contexts, this section reveals gaps and structural inequalities in current AI governance frameworks, the socio-technical logics of NCSII—commonly referred to as "deepfake pornography," though this term should not normalize or trivialize the harms involved the feminist political economy of synthetic media.

2.1. Mapping TFGBV: The Gendered Landscape of Digital Harm in the Global South

TFGBV is defined as follows:

> Any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological,

> social, political, or economic harm, or other infringements of rights and freedoms. (UN Women, 2022, p. 4)

TFGBV encompasses a broad range of digital harm, including network harassment, image-based sexual abuse, doxxing, defamation, stalking and monitoring, threats, and hate speech, that disproportionately target women and marginalized groups (Dunn, 2021). Definitions of TFGBV across legal and platform frameworks remain fragmented, reflecting divergent understandings of harm, responsibility, and accountability.

In this context, women and gender-diverse groups in the Global South face unique vulnerabilities shaped by intersecting inequalities. A global survey found 85% of women experiencing or witnessing online violence, with rates reaching 98% in the Middle East and 90% in Africa (The Economist Intelligence Unit, 2021). Furthermore, intersectional reports show that some groups are disproportionately affected by TFGBV. For instance, the UNESCO project The Chilling highlights that Black, Indigenous, and Jewish women journalists report higher exposure to online violence than white women journalists (Chowdhury & Lakshmi, 2023). UN reports also indicate that women in rural or low-connectivity regions remain disproportionately vulnerable to TFGBV due to digital literacy gaps and limited access to protection resources (UN, 2024a).

These individual-level harms are not isolated incidents—they are rooted in gendered power relations embedded in digital technologies and institutional structures (Dunn, 2020). TFGBV operates beyond the interpersonal level, revealing structural and institutional dynamics through the affordances of digital platforms, including virality, anonymity, and permanence. As such, it increases the circulation of abuse and diminishes perpetrators' accountability (Henry et al., 2020; Powell et al., 2021). Women in many low- and middle-income countries are increasingly misappropriated to generate "deepfake" content, resulting in harassment, blackmail, reputational damage, and social exclusion (Sheikh & Rogers, 2024). In countries like Ghana, Namibia, and Senegal, "deepfake pornography" has been weaponized against female politicians and journalists to undermine their credibility (Miliza et al., 2025). Structural dynamics of TFGBV in the Global South, where harms intersect with deep-rooted patriarchy, honor-based cultures, low digital literacy, and limited protections, produce systemic formations of vulnerability (Bansal et al., 2024). In such contexts, accountability often shifts from offenders to victims, leading to silencing and exclusion from public and digital life. For instance, Moroccan activist Ibtissame "Betty" Lachgar faced online threats, harassment, and a criminal sentence after her feminist social media post went viral ("Moroccan court upholds," 2025), and an 18-year-old woman on Facebook was killed after a manipulated photo of her next to her boyfriend went viral (Hussain, 2023).

2.2. "Deepfake Pornography" as Socio-Technical Gendered Harm

With the advancement of AI, NCSII has become increasingly widespread online, marking a disturbing evolution of TFGBV disproportionately targeting women (Umbach et al., 2024). Unlike early forms of image-based sexual abuse, which circulated authentic non-consensual intimate images, NCSII does not require pre-existing explicit material, and uses publicly available images to generate synthetic sexual content (Thomasen & Dunn, 2021). Of the 95,820 identified deepfake videos globally in 2023, pornographic content constituted 98% of the total, with 99% of the victims being women (Home Security Heroes, 2023, as cited in Birrer & Just, 2024).

NCSII builds on existing structural gendered inequalities, exacerbated by AI tools and social media algorithms that facilitate its rapid production and dissemination (Li et al., 2024; Viola & Voto, 2023). Generative AI has significantly lowered both technical and financial barriers to producing such content, while platform recommendation algorithms, driven by engagement metrics, inadvertently amplify its viral circulation (Kalpokas & Kalpokiene, 2022, pp. 65–71).

Across regions, AI governance has adopted distinct regulatory frameworks in response to this phenomenon. While the EU model integrates NCSII into broader digital governance, emphasizing transparency and data privacy, and the US focuses on criminalizing specific harms and safeguarding individual rights within a free speech context (Fabuyi et al., 2024), both fail to address the structural, gendered nature of digital harm (Birrer & Just, 2024). In the Global South, these harms are further exacerbated by deep-rooted socio-cultural norms, legal gaps, and limited legal aid and digital literacy (Sheikh & Rogers, 2024).

Feminist and decolonial theories reveal how AI exacerbates gender inequalities, challenging claims of technological "neutrality" and arguing how these systems are structurally embedded in and reinforce existing power dynamics. From a data feminism perspective (D'Ignazio & Klein, 2020), patriarchal structures within AI systems lead to the non-consensual extraction and commodification of marginalized groups' identities and experiences, thereby reproducing and deepening intersecting gendered, racial, and economic inequalities. Benjamin (2023), in her analysis of the "New Jim Crow," further reveals that technology not only passively replicates but actively amplifies structural racial and gender hierarchies and fosters new mechanisms of social control. Arora and Natale (2025) advocate for "Situated AI" practices, to reinvigorate engagements with the global that can account for the local, the cultural, and the particular in the context of generative media. Based on these theoretical insights, NCSII can be understood as more than an individual harm, as they are situated at the intersection of gendered technological infrastructures and global power structures.

From this perspective, our decolonial feminist approach operates on three interrelated levels: epistemologically, by centering experiences from the Global South (Arora, 2024a); methodologically, by promoting cross-regional and reflexive research (D'Ignazio & Klein, 2020); and politically, by challenging Northern-centric AI governance narratives (Arora & Natale, 2025). This framework deepens critiques of dominant models of AI governance and provides a basis for examining the political economy of deepfake content regulation. More importantly, it sets the stage for introducing reconstructive approaches such as "geographies of hope" (Hazlewood et al., 2023), which reframe governance beyond risk and harm toward reimaginations of care, agency, and future visions.

2.3. The Political Economy of Deepfake Technology: Who Controls AI-Generated Content?

A feminist political economy allows us to examine how NCSII emerges from and reinforces intersecting hierarchies of gender, class, and global power, especially for women and marginalized digital users (Rao & Akram-Lodhi, 2021). The development and governance of deepfake technologies are concentrated among tech companies, open-source developer networks, and online communities, mainly located in the Global North, leaving Global South countries with little autonomy in setting platform accountability or technical standards (Viola & Voto, 2023). These actors control both content production tools and the infrastructures that distribute and monetize NCSII content, including cloud services, algorithms, and hosting platforms. This ownership structure excludes Southern communities from decisions over their data, digital identities, and

participation in online spaces (Paris, 2021). In particular, women and gender-diverse users in the Global South often lack control over how their identities are used or commodified. Furthermore, the technical labor of AI and deepfake tools is dominated by male engineers, developers, and entrepreneurs in the Global North (Mishra et al., 2024). Meanwhile, women, especially in the Global South, bear the social and emotional burden of NCSII. This unpaid, gendered labour, including coping with harm, seeking redress, protecting oneself online, and rebuilding safety, remains insufficiently recognized within many platform design and policy frameworks (Birrer & Just, 2024).

Despite generating economic value in advertising, entertainment, and politics, deepfake technologies operate within platform economies that tend to prioritize monetisation and innovation, and often at the expense of user safety and gender justice (Paris, 2021). Regulatory frameworks, too, are tilted toward maintaining the existing power structures and logics, rather than being centred on user wellbeing and social justice. For instance, in the US, Section 230 of the Communications Decency Act shields platforms from liability; while the EU's Digital Services Act centers on a "notice-and-action" mechanism that emphasizes post-event governance and user rights, providing large technology companies with legitimacy and operational space due to its complex compliance system. China's Deep Synthesis Regulation imposes proactive pre-publication review and platform accountability to maintain social stability and public opinion security (Okolie, 2023; Zheng et al., 2025). Whether driven by commercial logic or state-oriented control, these models fail to center user safety and gender equality, which indirectly exacerbate the vulnerability of women and gender diversity groups in the Global South. By contrast, some countries have implemented specific legislation aimed at protecting victims from NCSII content, including the UK Online Safety Act 2023 and Australia's eSafety Commissioner regulations (Broinowski & Martin, 2024; Romero Moreno, 2024). These developments demonstrate that regulatory practices vary significantly across the globe, underscoring the need for global standards that are rights-based, feminist, and yet are situated in the lived realities, local cultures, and social structures, to ensure standards can become enforceable.

From the perspective of feminist political economy and decolonial theory, NCSII is a systemic phenomenon driven by platform capitalism, patriarchal structures, and global hierarchies. Global North-controlled infrastructure and algorithms can reinforce capital accumulation, transforming the identities and experiences of women and gender diversity groups in the Global South into exploitable data resources. Economically, platform capitalism shifts the costs of safety and emotional labor onto female users, while benefiting from the guise of neutral regulation to mask these inequities. Moreover, colonial power dynamics have historically produced gendered "accumulation by dispossession," making the structural marginalization of Global South women in digital spaces almost inevitable (Harvey, 2004). Therefore, the proposed Gen(der) AI Safety framework must go beyond content moderation and include those excluded from the design and regulation of digital technologies.

2.4. Cross-Cultural Policy Approaches and Challenges for Gender AI Safety in the Global South

Dominant global AI governance models are shaped by frameworks led by the EU, the US, and China, each reflecting distinct institutional priorities. The EU's AI Act establishes a risk-centric regulatory framework focusing on quantifiable AI risks (e.g., unacceptable or high) and emphasizing conformity assessments and transparency obligations. However, it offers limited attention to deeply rooted social and intersectional harms, such as TFGBV, which are difficult to quantify (Fabuyi et al., 2024; Valeriani & Polito, 2025).

In contrast, US AI policy tends to privatize safety through legislation like Section 230 of the Communications Decency Act, leaving survivors of digital abuse to navigate opaque moderation systems with limited recourse (Paris, 2021). China's model allows for rapid intervention, particularly in the governance of synthetic media, but prioritizes state security and social stability. Policies like the Provisions on the Administration of Deep Synthesis Internet Information Services prohibit harms such as NCSII; however, these bans often serve state control rather than gender justice (Birrer & Just, 2024). While these frameworks differ in approach, they define safety through institutional and geopolitical logics rather than the lived experiences of survivors, limiting their capacity to address the complex social, cultural, and legal realities of TFGBV, especially in Global South contexts.

The NCSII challenges in the Global South reveal the structural limitations of Global North models in addressing gender AI safety. Cultural logics and intersecting social identities shape both the forms and severity of harm. For instance, South Africa has criminalized NCSII and incorporated it into data protection law (Cybercrimes Act and the Protection of Personal Information Act), but its legal framework still reflects Western-leaning priorities emphasizing free speech and intellectual property, making enforcement difficult due to weak accountability and delayed implementation (Gotora, 2024). In Senegal, women with public visibility and political engagement are more likely to be targeted by deepfakes (Miliza et al., 2025). These cultural logics interact with dimensions such as age, caste, religion, sexual orientation, and disability, exacerbating the vulnerability of certain groups (Bansal et al., 2024). These examples demonstrate that Northern-centric governance standards cannot adequately reflect the complexities of Global South societies.

In addition, support for victims varies significantly across the Global South. In countries such as Namibia and Indonesia, survivors often face shaming, isolation, or indifference from law enforcement (Ferdinal & Bakir, 2024; Miliza et al., 2025). Structural barriers, including limited legal protections, weak enforcement, and cultural stigma, further discourage reporting in Ghana and Senegal. These challenges persist in many Global North contexts, where police responses to NCSII remain limited and inconsistent (Birrer & Just, 2024; Umbach et al., 2024).

Institutionally, many legal systems in the Global South lag behind technological developments, lacking clear definitions and regulatory mechanisms for harms such as NCSII or algorithmic bias. For instance, India lacks a legal definition of "deepfake" (Vig, 2024). Although South Africa passed the Cybercrimes Act in 2021, its narrow definition of TFGBV and the need to prove "intent to harm" create obstacles for prosecution (Gotora, 2024, p. 14). Mexico's Ley Olimpia represents a landmark legal response to TFGBV, including deepfakes, yet over 80% of cases go unreported (Escalera Silva et al., 2024).

Furthermore, the Global South faces systemic disadvantages in technological infrastructure, data sovereignty, judicial capacity, and cross-platform governance. Judicial institutions often lack the technical tools and training to identify AI-generated content in NCSII cases, complicating evidence collection and prosecution (Miliza et al., 2025). Meanwhile, NCSII frequently spreads across platforms and countries, testing the limits of traditional territorial jurisdiction, making it more difficult to hold people and organizations accountable (Batool et al., 2024). At the same time, due to the fragmented and inconsistent policies and tools used by technology companies to define and detect NCSII content, law enforcement agencies often remain dependent on platform-specific procedures for content removal and evidence extraction (Bioni et al., 2023).

Despite judicial inefficiencies, limited technical capacity, and transnational regulatory gaps, the Global South has developed a range of innovative, locally rooted practices that bridge the systemic gaps left by dominant governance frameworks in the EU, US, and China. In Brazil and Senegal, local actors pilot community-based AI ethics through Indigenous data governance models and open infrastructures (Bioni et al., 2023; Miliza et al., 2025). Without strong state-led regulation, locally rooted experimentations step in and offer alternative pathways. For instance, in Pakistan, NGOs collaborate with law enforcement to develop hybrid judicial models combining digital forensics with trauma-informed care, offering survivors both emotional support and evidentiary resources (Batool et al., 2024). In Latin America, the Olimpia chatbot provides AI-powered emotional support and legal guidance to TFGBV victims, addressing gaps in institutional care (AuraChat.Ai, 2025).

These localized examples challenge the assumption that effective AI governance must mirror models from the Global North. Instead, they highlight the agency of Global South communities in shaping alternative pathways through culturally grounded resistance, relational care, and local gendered experiences, and in making the Global South not a regulatory gap for AI safety, but a geography of hope.

3. Policy Pathways for Gen(der) AI Safety

3.1. The Geographies of Hope for Policymaking

As generative AI becomes deeply embedded in social and cultural systems, scholars are moving beyond singular risk-based governance approaches, shifting toward multidimensional frameworks that integrate ethical, social, and relational perspectives (Pawelec, 2024; Whittaker et al., 2023). This has led to calls for a geographic reorientation of technology governance, emphasizing the need to treat spatiality as a critical dimension in understanding and shaping AI safety policy contextually (Walker & Winders, 2021). In response, critical geography's concept of "geographies of hope" offers a bottom-up theoretical and practical pathway for reimagining AI governance (Hicks, 2018). Hazlewood et al.'s (2023) "geographies of hope-in-praxis" framework stresses that technology policy must be embedded in space, power relations, and lived experiences, asking who gets to imagine hope, in which spaces, and whose futures are included. This implies that policy development should move beyond a narrow focus on technical functionality and economic utility, instead grounding itself in broader socio-spatial contexts and attending to the experiences and aspirations of those marginalized by technology, such as the Global South communities.

Here, the Global South represents a political category rather than a homogenous geographic entity. Despite its temporal flattening, it enables solidarity across shared postcolonial trajectories of weak institutional enforcement, patriarchal norms, and limited infrastructures (Spivak, 1988). Following Spivak's notion of "strategic essentialism," we deploy the Global South not as a homogenous unit but as a solidaristic category for comparative insights and collective pathways to gender AI safety. This provides a basis to move beyond the North/South binary toward South–South and North–South strategies, insisting that AI protocols be intrinsically cross-cultural and equity-based, with geographies of hope co-constructed by historically underrepresented contexts and communities.

Hence, efforts to govern gender AI safety in the Global South must begin with understanding that hope in the Global South is not a product of naivety, but a persistent ethical stance and survival strategy emerging from

enduring harm, structural asymmetries, and historical injustice (Arora, 2024b). This method centers the fears and hopes of women, LGBTQ+, and sexual minority groups in digital environments, positioning tech policy as a potential site of transformative practice. Their insights reframe AI safety not as a technical challenge but as a lived, ongoing negotiation of harm, care, and hope. These grounded practices require governance frameworks that are participatory, intersectional, and responsive to the diverse realities of life in the Global South. Thus, constructing a "mapping of hope" is not merely a conceptual exploration, but a practical reconstruction of how policy knowledge is produced, who participates, and what social imaginaries do technologies carry (Geerts, 2022, p. 385). Hope, in this context, becomes a governance approach in the Global South: a commitment to expanding the conditions of possibility for gender digital flourishing.

3.2. Legislative Innovations for Gendered AI Safety in the Global South

In the Global South, feminist and citizen-led data governance communities are pioneering innovative approaches to AI safety regulation that center care and justice in response to the challenges of gender inequality and algorithmic harm. These innovations are fundamental, not merely experimental, serving as structural responses to the gendered harms deeply embedded in digital infrastructures. What's more, these efforts challenge the dominant AI governance paradigms shaped by the Global North.

Unlike the risk-management approaches that prioritize hazard mitigation and strict control, legislative innovations in the Global South view AI safety as inseparable from broader struggles for gender justice, data sovereignty, and technological self-determination. The Digital Rights Foundation exemplifies this shift by developing feminist digital security frameworks that foreground survivor-centric, trauma-informed, and gender-responsive policies (Digital Rights Foundation, 2025a). Its Digital Security Helpline, South Asia's first dedicated service for digital rights, has received over 20,000 cases since its launch in Pakistan, offering psychosocial and legal support, as well as policy recommendations grounded in the lived realities of digital harms. According to the foundation's executive director, Nighat Dad, "cultural nuance, emotional intelligence, and lived experience cannot be programmed" (Digital Rights Foundation, 2025b). The Digital Rights Foundation's work advocates for reimagining gender AI safety as contextual, relational, and care-oriented, resisting automation-driven responses to gendered harms such as deepfakes, algorithmic blackmail, or image-based sexual abuse.

This reorientation is not confined to institutional governance; it is expanded through grassroots feminist tech practices across the Global South, where communities mobilize data not merely to protect, but also to empower. The civic interventions are not confined to risk reduction; they proactively create structures that strengthen participation and decision-making. Activist projects like Mariéme Jamme's iamtheCODE Foundation in Senegal and María Salguero's femicide mapping in Mexico exemplify how data can be mobilized for justice and visibility rather than surveillance and exclusion (Johnson, 2022; McCormick, 2024). Jamme's coding camps equip marginalized girls with the tools to become digital creators, not just passive subjects of AI systems. Salguero's femicide maps have not only raised national awareness but have also informed legislative reforms and public policy debates in Mexico, exemplifying how feminist data activism can transform fragmented, grassroots evidence into powerful tools for shaping state accountability and responsive governance. These examples demonstrate a situated, hopeful approach that emerges not from pessimism but within the Global South's fragmented, censored, and often digitally mediated patriarchal conditions (Arora, 2024a).

Furthermore, localized regulatory mechanisms are emerging to formalize these grassroots practices. In Mexico, Ley Olimpia, a comprehensive set of legislative reforms enacted across various states, recognizes and penalizes digital violence, particularly offences that violate individuals' sexual privacy through digital means (de la Vega & Escalera Silva, 2025). Named after activist Olimpia Coral Melo, who led the "Ley Olimpia movement" after surviving image-based abuse, the law is now considered a model framework for combating digital gender violence in Latin America (de la Vega & Escalera Silva, 2025, p. 5). Although often contested and incomplete, these legislative strategies in the Global South indicate the potential for feminist engagement with state structures to reshape legal infrastructures around digital harm.

Most importantly, these interventions demonstrate the limitations of standardized regulatory modes that attempt a one-size-fits-all solution (Phelan, 2022). Instead, they advocate for pluriversal approaches that recognize and incorporate local contexts, legal traditions, and sociotechnical realities that are situated, context-specific, and co-created with communities, acknowledging multiple valid ways to understand technology, define justice, and live in the digital world. This includes cross-sectoral collaborations between civic organizations, government policymakers, technologists, and feminists to create governance models of safety, participation, and flourishing.

3.3. The ABCDE Framework for Gen(der) AI Safety: Mapping of Hope

As AI technologies increasingly shape the lived realities of billions through recommendation systems, automated content generation, virtual agents, and biometric surveillance, the question is no longer merely whether AI systems are safe, but safe for whom, by whom, and under what conditions (Bengio et al., 2025; Valeriani & Polito, 2025). Addressing this complex notion of AI safety demands a fundamental shift in approach. It cannot be resolved through technical fixes or regulatory mandates alone, as its inherent complexity is deeply embedded in social, cultural, and political factors (Valeriani & Polito, 2025). This requires an urgent shift from centralized, reactive, and technocratic governance models toward participatory, anticipatory, and justice-driven approaches that center cultural specificity, community agency, and the lived realities of marginalized groups (Arora, 2024a). This article proposes the ABCDE Framework for Gen(der) AI Safety to operationalize this transformative vision of inclusive, anticipatory, and feminist AI governance. Figure 1 shows the overall structure of the ABCDE framework and its five interdependent pillars. This inclusive framework is rooted in cross-cultural alliances, empirical co-design, collective rights, design justice, and empowerment. Each pillar of the framework directly addresses one or more of the questions above, providing a holistic strategy to building AI futures that are not only safe, but also fair, equitable, and inclusive.

3.3.1. Alliances Cross-Cultural and Cross-Sectoral

AI governance models fundamentally shape how safety is defined, how it operates, and whose interests are ultimately served (Paris, 2021). In a deeply mediatized AI ecosystem characterized by structural inequities and asymmetrical power flows, cross-cultural and cross-sectoral alliances are not subsidiary to governance. Rather, they are the architecture upon which inclusive AI futures must be built (Arora, 2024a). Even the most technologically advanced AI systems can replicate and exacerbate existing exclusions if their design, deployment, and governance are divorced from inclusive relations, social context, and the lived experiences of those they serve. This highlights that to address gender inequality and other deep social exclusions in AI systems, it is not enough to rely on technological development alone. Instead, building and relying on

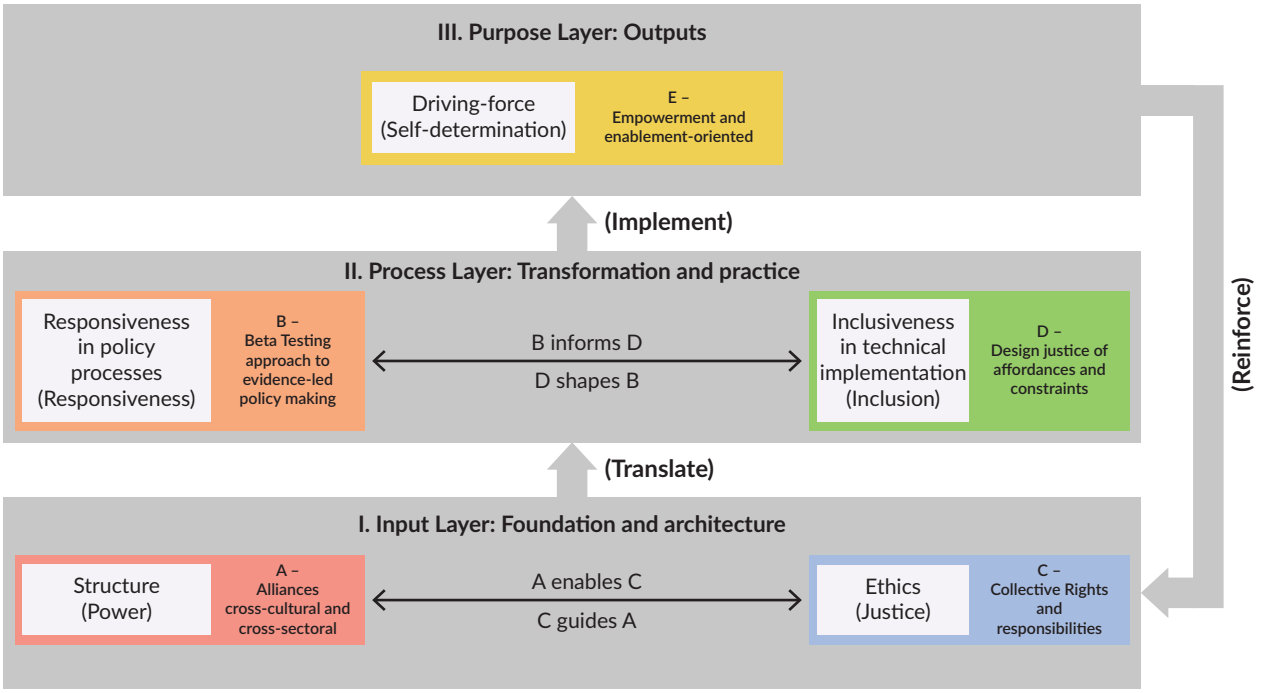

Figure 1. The ABCDE Framework for Gen(der) AI Safety.

cross-sectoral alliances is crucial to connect policymakers, feminist researchers, grassroots organizers, artists, technologists, and civil society actors. Moreover, the framework also insists on cross-cultural solidarity as a central governance mode, especially among Global South actors. These alliances must extend beyond a symbolic level of diversity and promote in-depth dialogue and transformation between feminist, decolonial, and Indigenous knowledge systems, which are not just consulted but structurally embedded.

According to Arora (2024a), decolonial frameworks require more than mere recognition. They need a fundamental redistribution of governance power, enabling agency of marginalized communities most impacted by AI systems. This process involves supporting institutional mechanisms traditionally excluded from the core of AI governance, including higher education translating critical knowledge into actionable insights, public institutions promoting fair and equitable legislation, and rights-based advocacy groups mediating local norms with global justice. This redistribution is both ethical and strategic: Cross-cultural and cross-sectoral alliances are more resilient to capture by Big Tech monopolies or authoritarian regimes and reflect governance logics already practiced in many Indigenous, feminist, and cooperative traditions across the Global South. These alliances decentralize power while institutionalizing accountability through legal reforms, public oversight, and platform compliance mechanisms, thereby translating power redistribution into enforceable governance. They further enhance survivor support systems, and improve platform procedures for content removal, data preservation, and evidence provision.

The intersectoral feminist alliances in Latin America offer an example (Ciolfi Felice et al., 2025). By promoting legal reforms and cross-regional networks, they have built a governance framework to challenge tech monopolies and algorithmic discrimination. Yet, despite advances in legislation and public advocacy, there remain challenges related to resource sustainability and limited leverage over large technology corporations. Persistent transformation of governance structures grounded in cross-sectoral and cross-cultural alliances can incrementally but meaningfully institute structural change.

3.3.2. Beta-Testing Approach to Evidence-Led Policymaking

Beta-testing is proposed as a foundational strategy for developing evidence-based policies that are responsive to the lived experiences of those most affected by AI harms in the Gen(der) AI Safety Framework. Rather than top-down and technocratic models, characterized by regulation and risk-based technical classifications (Birrer & Just, 2024), beta-testing calls for an iterative, participatory, and community-driven approach to AI safety governance. At its core, beta-testing reframes policymaking as an evolving process: one that relies on empirical testing, localized feedback mechanisms, and user-driven interventions (Arora, 2024a). It centers communities as co-creators of policy, not passive recipients of legal frameworks. This shift is especially critical in addressing emerging AI harms such as "deepfake pornography" where threats evolve more rapidly than legal systems can respond.

Responsive governance is made possible through mechanisms such as community-informed pilot programs, digital helplines, and survivor-led consultations. For example, the Digital Rights Foundation's helpline in Pakistan, designed to support victims of TFGBV, exemplifies how grassroots infrastructures can inform agile policy reform (Digital Rights Foundation, 2025a). Such models do more than collect complaints; they generate real-time insights that can be translated into concrete responsibilities, design interventions, and enforcement practices. Beta-testing can enable survivor-centered policy decisions, with continuous iteration for assessing its effectiveness across diverse cultural and contextual environments. The core contribution of the beta-testing approach lies in establishing an agile cycle linking practice, empowerment, and systemic strengthening, given the fast-changing nature of AI technologies. When feedback and response mechanisms are embedded in policymaking, governance no longer operates as a one-way directive and becomes a self-learning, self-correcting process.

3.3.3. Collective Rights and Responsibilities (Group Privacy vs. Individual Privacy)

US and EU AI safety and data governance models are deeply rooted in liberal legal traditions that prioritize individual rights, especially personal privacy and consent, as the normative foundation of ethical oversight (Fabuyi et al., 2024). However, in deeply mediatized environments of the Global South, data are fundamentally relational, cultural, and communal (Bhatia et al., 2025). Consequently, WEIRD models often obscure a critical reality: Harms are not exclusively personal. They are also collective, particularly for marginalized gender groups, racialized communities, and Indigenous populations whose data and representations are co-constituted through shared histories, identities, and systemic injustices (Arora, 2024b).

The collective rights and responsibilities pillar addresses this gap by centering group privacy, and cultural data sovereignty. It sidelines the Western paradigm of data protection, which is shaped by narrow, proprietary conceptions that prioritize the individual and marginalize relational and collectivist ethical frameworks (such as Ubuntu and Buddhist perspectives) that emphasize communal accountability and interconnected privacy (Goffi, 2021). This limitation is critical because AI systems, operating within data capitalism, routinely extract, process, and commodify data in ways that implicate entire communities, especially in the Global South, reinforcing systemic inequalities, eroding collective autonomy, and perpetuating digital colonialism (Medrado & Verdegem, 2024). For instance, when facial recognition algorithms misclassify racialized gender expressions, or when generative AI systems remix Indigenous symbols without consent, attribution, or accountability, they not only violate individual privacy but also

disrupt communal narratives, cultural continuity, and the intergenerational transmission of knowledge (Chateau et al., 2025; Zevop & Ballet, 2025).

Furthermore, this pillar redefines the architecture of justice. Shifting the focus from individual compensation, it advocates for community-led response systems and shared responsibility. This includes collective legal standing in content takedown processes, cooperative licensing models for cultural expression, and restorative justice measures for algorithmic harms disproportionately affecting gendered and cultural communities. These interventions reimagine AI safety as relational and systemic, not just procedural (Arora, 2024a). For example, African communal data governance practices demonstrate how local communities collaboratively manage sensitive genetic and health data, guided by philosophies such as Ubuntu and Ujamaa, which emphasize relational responsibilities, collective ownership, and shared accountability in decision-making and data use (Munung et al., 2024). These practices echo minority-led initiatives in Western countries, including the Sámi Data Sovereignty Initiative (Kukutai & Taylor, 2016). The Sámi Data Sovereignty Initiative enables the Sámi people to assert authority over their data, embedding cultural values and ethical principles into digital infrastructures. These cases demonstrate how marginalized communities across diverse regions are operationalizing justice-centered approaches to AI and data governance, while also revealing ongoing challenges such as resource constraints and the need for institutional support.

Ultimately, recognizing collective rights necessitates a radical shift in power—from extractive data regimes dominated by corporate platforms and state institutions toward community-determination. This means enabling communities to define the terms under which their data are gathered, circulated, and remediated, including who benefits and how harm is addressed. In feminist and decolonial contexts, this is essential to counteract ongoing epistemic violence, cultural appropriation, and gendered exploitation, often obscured by the rhetoric of "innovation." The significance of these practices is rooted in the empowerment they generate among participants. When communities exercise genuine control over their data, narratives, and identities, they shift from being objects of governance to active agents. This enhanced agency reinforces both the ethical foundations and operational structure, creating a self-reinforcing cycle resilient to external exploitation and grounded in cultural resources. Meaningful AI safety frameworks, therefore, must be embedded in sustained, community-driven, justice-oriented practice.

3.3.4. Design Justice of Affordances and Constraints

Design justice demands a shift from WEIRD paradigms toward participatory, transparent, context-sensitive approaches rooted in feminist, decolonial frameworks. It centers the idea that technologies, particularly AI systems, are never neutral in design (Bengio et al., 2025). They are not only embedded with human values, intentions, and social biases during their design and training, but they also actively structure the conditions of hope for user behavior, shaping what actions are enabled, constrained, or excluded. In the ABCDE Framework for Gen(der) AI Safety, design becomes the site where power is encoded: At the interface level, the framework surfaces who matters, what is normal, and what forms of agency are permitted or denied when attending to algorithmic decisions. To promote gender AI safety and flourishing in the Global South, design justice must center the affordances (what systems enable) and constraints (what systems restrict) from the perspective of the marginalized users (Arora, 2024a).

For instance, in Latin America, UN Women and the UN International Computing Centre co-developed a multilingual AI model to detect and flag sexist content on X (formerly Twitter), trained on Spanish-speaking culturally specific datasets (International Telecommunication Union, 2024). This initiative exemplifies design justice in action—embedding feminist principles into AI, while ensuring that content moderation reflects local realities rather than defaulting to WEIRD-centric norms. A comparable approach is seen in Western contexts, such as Costanza-Chock's (2020) design justice project that develops participatory methods to co-design digital systems with marginalized groups, ensuring the technologies reflect people's needs and values. These cases highlight how inclusive design, especially in moderation systems, can empower marginalized users. Extending this principle further, gender-diverse representation, multilingual moderation, and community-controlled data can expand agency and inclusion (Arora, 2024a). Restricting AI from generating hypersexualized, racialized, or non-consensual imagery is a safeguard against structural violence (Bengio et al., 2025).

Reframing safety as a relational and political concern implies building it into AI systems at the level of everyday user interaction, rather than relying solely on state regulatory frameworks. When embedded within the technical architecture, such design can empower users and communities, transforming them from passive subjects of protection into active agents in shaping the technological environment. This enhanced capability, in turn, generates a continuous drive and legitimacy for the safety governance system, advancing its continual evolution toward greater justice and inclusivity.

3.3.5. Empowerment and Enablement-Oriented

The final pillar of the ABCDE Framework—empowerment and enablement-oriented—repositions gender AI safety not as a perimeter of risk avoidance, but as a platform for hope, which is a spatial, social, and political commitment to empowerment and self-determination. This orientation moves beyond conventional, protectionist models of AI governance that treat users as passive recipients of harm mitigation, and instead affirms them as co-creators of digital futures grounded in justice, joy, and autonomy.

Mainstream AI safety discourses—dominated by the US, EU, and China—often flatten the spatial politics of AI, negating the lived realities and knowledge systems of those at the margins, particularly in the Global South. As Hazlewood et al. (2023) argue, empowerment is spatially situated and relationally constituted. It draws from critical geography and feminist theory to center the aspirations and creative capacities of communities that have historically been excluded from shaping technological imaginaries, particularly women, LGBTQ+, and other marginalized groups in the Global South. These are not just "end users" of AI; they are key actors whose unique knowledge and cultural visions must guide AI governance (Arora, 2024a).

Empowerment involves survivor-led policy design, gender-sensitive safety architectures, and community-driven governance mechanisms (Miliza et al., 2025). Examples of this empowerment in action are already visible: In Argentina, the feminist initiative AymurAI, developed by Data Género, equips judicial staff, particularly in criminal courts, with tools to anonymize and structure legal rulings on gender-based violence (Whitehead & Gandhi, 2024). By transforming opaque judicial records into open, privacy-protected datasets, AymurAI strengthens the capacity of institutional actors to make gender-based violence data accessible and actionable, while enabling feminist civil society to use these data for survivor-centered advocacy, policy reform, and systemic transparency. The significance of such practices derives from the ongoing drive for change they generate; they establish a reinforcing cycle: Enhanced institutional action and

stronger social oversight create a governance environment more responsive to gender justice, which in turn fosters further empowerment initiatives.

This pillar serves as a vital, self-reinforcing engine for the gender AI safety framework system, ensuring that governance transcends static regulation-making and continuously evolves through the strengthened agency of communities and institutions toward justice.

This ABCDE Framework offers a transformative roadmap for reimagining gen(der) AI safety through a feminist, justice-oriented, and Global South-led lens. This framework not only safeguards against harm, but also actively cultivates inclusive, culturally grounded, and empowering digital futures.

4. Rethinking AI Safety: From Harm and Care Toward a Hopeful Future

Feminist technoscience scholars argue that safety and harm in technology and design are not isolated risk control issues, but are shaped and determined by profound social, relational, and political factors (Costanza-Chock, 2020). Dominant regulatory approaches remain narrowly focused on detecting AI content that imitates reality, such as synthetic media or manipulated video, while this harm-centered perspective overlooks the structural exclusions embedded in AI design, deployment, and governance (Bengio et al., 2025; Fabuyi et al., 2024). Most regulatory systems in the Global South continue to rely on outdated models that treat such harms as incidental rather than systemic (Batool et al., 2024). This blind spot is further compounded by global governance structures in which standards, enforcement, and accountability are largely shaped by institutions in the Global North, often leaving the Global South structurally excluded (Okolie, 2023; Zheng et al., 2025).

This approach often overlooks the agency and motivations of marginalized users themselves. Global South users, including women and other marginalized groups, engage with digital platforms to access work, education, public visibility, and community building. As Arora (2019, 2024a) emphasizes, marginalized users frequently navigate challenging environments (such as patriarchal norms, surveillance, censorship, and bias). Hence, the digital can serve as one of the few places where they can assert autonomy, resist marginalization, and participate in public life, despite the digital harms and risks. Overlooking this dual reality of harm and hope results in policy approaches that restrict access rather than enabling safer, more equitable participation online. Feminist and justice-oriented AI governance must begin with this contradiction where risk and opportunity coexist, and shift from diagnosing harm to cultivating care and mobilizing hope.

Rethinking AI safety requires a decisive shift away from defensive, risk-centric paradigms toward frameworks that are co-constructed, spatially situated, and contextually grounded. Drawing on feminist theory and the concept of the geographies of hope (Hazlewood et al., 2023), we argue that care must be reframed not as digital paternalism but as a commitment to infrastructural support, mutual accountability, and enabling agency. It demands the development of co-designed AI systems that prioritize user dignity, participatory governance, and collective well-being (Arora, 2024a). Furthermore, we reposition the digital space through the geographies of hope. Grounding AI governance in the lived experiences and socio-spatial realities in the Global South enables policy frameworks that respond to the aspirations, fears, and hopes of marginalized communities. Through this reconceptualization, gen(der) AI safety offers not only a critique of Global North paradigms, but also a forward-looking approach for inclusive, proactive, and just AI futures.

Acknowledgments

This publication is part of the Google Awarded Gen(der) AI Safety project with Professor Payal Arora as the PI and Dr. Marta Zarzycka as Google liaison and partner. The views expressed are those of the authors and do not necessarily reflect the views of Google or its affiliates.

Funding

Publication of this article in open access was made possible through the institutional membership agreement between Utrecht University and Cogitatio Press.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

This study is based on secondary sources and publicly available policy reports. All data used in the analysis can be accessed through the references cited within the article.

LLMs Disclosure

NotebookLM was used to break down literature into modular, reusable segments, thereby supporting source tracking and perspectives' comparison.

References

AI for gender equality: UN Women AI School opens for changemakers. (2025, March 5). *UN Women*. https://asiapacific.unwomen.org/en/stories/announcement/2025/03/un-women-ai-school-opens-for-changemakers

Amatika-Omondi, F. (2022). The regulation of deepfakes in Kenya. *Journal of Intellectual Property and Information Technology Law*, *2*(1), 145–186.

Amnesty International. (2024). *Coded injustice: Surveillance and discrimination in Denmark's automated welfare state*.

Arora, P. (2019). *The next billion users: Digital life beyond the West*. Harvard University Press.

Arora, P. (2024a). *From pessimism to promise: Lessons from the Global South on designing inclusive tech*. MIT Press.

Arora, P. (2024b). The privilege of pessimism: The politics of despair towards the digital and the moral imperative to hope. *Dialogues on Digital Society*, *1*(1), 33–36. https://doi.org/10.1177/29768640241252103

Arora, P., & Huang, W. (2025). *Gender data: What is it and why is it important for the future of AI systems?* Friedrich-Ebert-Stiftung.

Arora, P., & Natale, S. (2025). Situating AI: Global media approaches to artificial intelligence. *Media, Culture & Society*, *47*(5), 1007–1011. https://doi.org/10.1177/01634437251341702

AuraChat.Ai. (2025). *Empowering victims of digital violence: How AuraChat.Ai's AI technology transforms support systems*. https://www.aurachat.ai/post/empowering-victims-of-digital-violence-how-aurachat-ai-s-ai-technology-transforms-support-systems

Bandhakavi, S. (2024, November 28). Lunit partners with Salud Digna to expand AI diagnostics in Latin America. *Medical Device Developments*. https://www.medicaldevice-developments.com/news/lunit-partners-with-salud-digna-to-expand-ai-diagnostics-in-latin-america

Bansal, V., Rezwan, M., Iyer, M., Leasure, E., Roth, C., Pal, P., & Hinson, L. (2024). A scoping review of

technology-facilitated gender-based violence in low- and middle-income countries across Asia. *Trauma, Violence, & Abuse*, *25*(1), 463–475. https://doi.org/10.1177/15248380231154614

Batool, A., Naseem, M., & Toyama, K. (2024). Expanding concepts of non-consensual image-disclosure abuse: A study of NCIDA in Pakistan. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Syste* (Article 398). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642871

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., . . . Wheeler, N. (2025). *International AI safety report* (DSIT 2025/001). AI Action Summit. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025

Benjamin, R. (2023). Race after technology. In W. Longhofer & D. Winchester (Eds.), *Social theory re-wired: New connections to classical and contemporary perspectives* (pp. 405–415). Routledge.

Bhatia, K. V., Pathak-Shelat, M., Sinha, S., & Mishra, T. (2025). Global influencers' content creation strategies: Negotiating with platform affordances to practice vernacular creativity. *Media, Culture & Society*, *47*(1), 130–153. https://doi.org/10.1177/01634437241276408

Bioni, B., Garrote, M., & Guedes, P. (2023). *Key themes in AI regulation: The local, regional, and global in the pursuit of regulatory interoperability*. Data Privacy Brazil Research Association. https://www.dataprivacybr.org/en/keythemes-in-ai-regulation-the-local-regional-and-global-in-the-pursuit-of-regulatory-interoperability

Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*, *27*(12), 6819–6838. https://doi.org/10.1177/14614448241253138

Broinowski, A., & Martin, F. R. (2024). Beyond the deepfake problem: Benefits, risks and regulation of generative AI screen technologies. *Media International Australia*. Advance online publication. https://doi.org/10.1177/1329878X241288034

Catalano, T., & Waugh, L. R. (2020). *Critical discourse analysis, critical discourse studies and beyond*. Springer.

Chateau, L., Arora, P., & Herman, L. (2025). Cross-cultural approaches to creative media content in the age of AI. *Media, Culture & Society*, *47*(5), 1012–1027. https://doi.org/10.1177/01634437251328188

Chowdhury, R., & Lakshmi, D. (2023). *"Your opinion doesn't matter, anyway": Exposing technology-facilitated gender-based violence in an era of generative AI*. UNESCO Publishing.

Ciolfi Felice, M., Feldfeber, I., Glasserman Apicella, C., Quiroga, Y. B., Ansaldo, J., Lapenna, L., Bezchinsky, S., Barriga Rubio, R., & García, M. (2025). Doing the feminist work in AI: Reflections from an AI project in Latin America. In N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 998). Association for Computing Machinery.

Confino, P. (2024, December 11). Melinda French Gates will donate $150 million toward women in the workplace—And one-third of it will go to AI. *Fortune*. https://fortune.com/2024/12/11/melinda-french-gates-150-million-donation-women-in-the-workplace-ai

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.

de la Vega, D. A. G., & Escalera Silva, L. A. (2025). Substantive analysis of digital violence in Mexico: Olimpia Law, a case study in Nuevo Leon. *TransAmerica Review*, *3*(1), Article e25003. https://doi.org/10.62910/transame25003

De Sousa Santos, B. (2015). *Epistemologies of the South: Justice against epistemicide*. Routledge.

Digital Rights Foundation. (2025a). *2024 digital security helpline annual report*. https://digitalrightsfoundation.pk/wp-content/uploads/2025/05/Digital-Security-Helpline-Annual-Report-2024.pdf

Digital Rights Foundation. (2025b, April 24). *Over 20,000 cases of technology-facilitated gender-based violence (TFGBV) received by Digital Rights Foundation's Helpline during 8 years of operation* [Press release]. https://digitalrightsfoundation.pk/subject-over-20000-cases-of-technology-facilitated-gender-based-violence-tfgbv-received-by-digital-rights-foundations-helpline-during-8-years-of-operation

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

Dunn, S. (2020). *Technology-facilitated gender-based violence: An overview* (Supporting a Safer Internet Paper No. 1). Centre for International Governance Innovation. https://www.cigionline.org/publications/technology-facilitated-gender-based-violence-overview

Dunn, S. (2021). Is it actually violence? Framing technology-facilitated abuse as violence. In J. Bailey, A. Flynn, & N. Henry (Eds.), *The Emerald international handbook of technology-facilitated violence and abuse* (pp. 25–45). Emerald Publishing.

Escalera Silva, L. A., Amador Corral, S. R., & Lara Hernández, Y. M. (2024). Digital gender-based violence from the experience of actors in the delivery of justice. *Sapienza: International Journal of Interdisciplinary Studies*, *5*(1), Article e24018. https://doi.org/10.51798/sijis.v5i1.741

Fabuyi, J., Olaniyi, O. O., Olateju, O., Aideyan, N. T., Selesi-Aina, O., & Olaniyi, F. G. (2024). Deepfake regulations and their impact on content creation in the entertainment industry. *Archives of Current Research International*, *24*(12), 52–74. https://doi.org/10.9734/acri/2024/v24i12997

Fearnley, L. C. A., Cairns, E., Stoneham, T., Ryan, P. M., Chubb, J. A., Iacovides, J., Iglesias Urrutia, C. P., Morgan, P. D. J., McDermid, J. A., & Habli, I. (2025). *Risk of what? Defining harm in the context of AI safety*. White Rose Research Online. https://eprints.whiterose.ac.uk/id/eprint/223407

Ferdinal, O., & Bakir, H. (2024). Legal protection efforts and policies to combat deepfake porn crimes with artificial intelligence (AI) in Indonesia. *Journal of Multidisciplinary Sustainability Asean*, *1*(6), 465–474.

Geerts, E. (2022). Navigating (post-)anthropocenic times of crisis: A critical cartography of hope. *CounterText*, *8*(3), 385–412. https://doi.org/10.3366/count.2022.0281

Goffi, E. (2021). Escaping the Western cosm-ethical hegemony: The importance of cultural diversity in the ethical assessment of artificial intelligence. *AI Ethics Journal*, *2*(2). https://doi.org/10.47289/AIEJ20210716-1

Gotora, N. T. (2024). Unmasking deception: Deepfake regulation in the context of South African law, could a rethinking of performers' protection rights be the answer? *International Journal of Law and Information Technology*, *32*(1), Article eaae026. https://doi.org/10.1093/ijlit/eaae026

Harvey, D. (2004). The 'new' imperialism: Accumulation by dispossession. *Socialist Register*, *40*, 63–87.

Hazlewood, J. A., Middleton Manning, B. R., & Casolo, J. J. (2023). Geographies of hope-in-praxis: Collaboratively decolonizing relations and regenerating relational spaces. *Environment and Planning E: Nature and Space*, *6*(3), 1417–1446. https://doi.org/10.1177/25148486231191473

Held, V. (2005). *The ethics of care: Personal, political, and global*. Oxford University Press.

Henry, N., Flynn, A., & Powell, A. (2020). Technology-facilitated domestic and sexual violence: A review. *Violence Against Women*, *26*(15/16), 1828–1854. https://doi.org/10.1177/1077801219875821

Hicks, D. (2018). Why we still need a geography of hope. *Geography*, *103*(2), 78–85. https://doi.org/10.1080/00167487.2018.12094041

Hussain, A. (2023, November 30). Pakistani girl killed after photos with boy's arm around her go viral. *Al Jazeera*. https://www.aljazeera.com/news/2023/11/30/pakistani-girl-killed-after-photos-with-boys-arm-around-her-go-viral

International Telecommunication Union. (2024). *Unveiling sexist narratives: AI approach to flag content on social media*. AI for Good. https://aiforgood.itu.int/event/unveiling-sexist-narratives-ai-approach-to-flag-content-on-social-media

Johnson, J. (2022, December 22). "Counting feminicide: What data scientists can learn from grassroots feminist activists" with Catherine D'Ignazio. *The NULab for Digital Humanities and Computational Social Science*. https://cssh.northeastern.edu/nulab/counting-feminicide-with-catherine-dignazio

Kalpokas, I., & Kalpokiene, J. (2022). *Deepfakes: A realistic assessment of potentials, risks, and policy regulation*. Springer. https://doi.org/10.1007/978-3-030-93802-4

Kukutai, T., & Taylor, J. (2016). Data sovereignty for Indigenous peoples: Current practice and future needs. In T. Kukutai & J. Taylor (Eds.), *Indigenous data sovereignty: Toward an agenda* (pp. 1–22). ANU Press.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

Li, Q., McDonald, A., Haimson, O. L., Schoenebeck, S., & Gilbert, E. (2024). The sociotechnical stack: Opportunities for social computing research in non-consensual intimate media. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), Article 375. https://doi.org/10.1145/3686914

McCormick, M. (2024, November 13). How lady Marième Jamme is bringing humanity into the boardroom. *Forbes*. https://www.forbes.com/sites/meghanmccormick/2024/11/13/how-lady-marime-jamme-is-bringing-humanity-into-the-boardroom

Medrado, A., & Verdegem, P. (2024). Participatory action research in critical data studies: Interrogating AI from a South–North approach. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241235869

Miliza, J., Gichanga, M., & Kiden, S. (2025). *Digital shadows: Deepfakes used as violence against women in journalism and politics during African elections*. Tanda Community Network. https://drive.google.com/file/d/1MD2OlME1SJbRMCpcMEnXac5odY0ucLbu/view?usp=sharing

Mishra, D., Ngoc Le, A., & McDowell, Z. (Eds.). (2024). *Communication technology and gender violence*. Springer.

Moroccan court upholds 30-month sentence for feminist activist over blasphemous t-shirt. (2025, October 6). *Le Monde*. https://www.lemonde.fr/en/le-monde-africa/article/2025/10/06/retrial-of-jailed-moroccan-feminist-activist-began-on-monday_6746158_124.html

Munung, N. S., Royal, C. D., de Kock, C., Awandare, G., Nembaware, V., Nguefack, S., Treadwell, M., & Wonkam, A. (2024). Genomics and health data governance in Africa: Democratize the use of big data and popularize public engagement. *Hastings Center Report*, *54*(S2), S84–S92. https://doi.org/10.1002/hast.4933

Okolie, C. (2023). Artificial intelligence-altered videos (deepfakes), image-based sexual abuse, and data privacy concerns. *Journal of International Women's Studies*, *25*(2), Article 11. https://vc.bridgew.edu/jiws/vol25/iss2/11

Paris, B. (2021). Configuring fakes: Digitized bodies, the politics of evidence, and agency. *Social Media + Society*, *7*(4). https://doi.org/10.1177/20563051211062919

Pawelec, M. (2024). Decent deepfakes? Professional deepfake developers' ethical considerations and their governance potential. *AI and Ethics*, *5*, 2641–2666. https://doi.org/10.1007/s43681-024-00542-2

Phelan, P. (2022). Are the current legal responses to artificial intelligence facilitated 'deepfake' pornography sufficient to curtail the inflicted harm? *North East Law Review*, *9*(2), 20–29.

Powell, A., Flynn, A., & Sugiura, L. (Eds.). (2021). *The Palgrave handbook of gendered violence and technology*. Springer.

Rao, S., & Akram-Lodhi, A. H. (2021). Feminist political economy. In G. Berik & E. Kongar (Eds.), *The Routledge handbook of feminist economics* (1st ed., pp. 34–42). Routledge.

Romero Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, *38*(3), 297–326. https://doi.org/10.1080/13600869.2024.2324540

Sheikh, M. M. R., & Rogers, M. M. (2024). Technology-facilitated sexual violence and abuse in low and middle-income countries: A scoping review. *Trauma, Violence, & Abuse*, *25*(2), 1614–1629. https://doi.org/10.1177/15248380231191189

Spivak, G. C. (1988). Can the subaltern speak? In C. Nelson & L. Grossberg (Eds), *Marxism and the interpretation of culture* (pp. 271–313). University of Illinois Press.

The Economist Intelligence Unit. (2021). *Measuring the prevalence of online violence against women*. https://onlineviolencewomen.eiu.com

Thomasen, K., & Dunn, S. (2021). Reasonable expectations of privacy in an era of drones and deepfakes: Expanding the Supreme Court of Canada's decision in 'R v Jarvis.' In J. Bailey, A. Flynn, & N. Henry (Eds.), *The Emerald international handbook of technology-facilitated violence and abuse* (pp. 555–576). Emerald Publishing.

Tronto, J. (1993). *Moral boundaries: A political argument for an ethic of care*. Routledge.

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-consensual synthetic Intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 779). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642382

UN. (2024a). *Intensification of efforts to eliminate all forms of violence against women and girls: Technology-facilitated violence against women and girls* (Report of the Secretary-General A/79/500).

UN. (2024b). *Mind the AI divide: Shaping a global perspective on the future of work*.

UN Women. (2022). *Technology-facilitated violence against women: Towards a common definition. Report of the meeting of the Expert Group 15-16 November 2022, New York, USA*. https://www.unwomen.org/sites/default/files/2023-03/Expert-Group-Meeting-report-Technology-facilitated-violence-against-women-en.pdf

Valeriani, M., & Polito, C. (2025). Artificial intelligence and political risk analysis. In C. E. Sottilotta, J. Campisi, J. Leitner, & H. Meissner (Eds.), *The Routledge handbook of political risk* (1st ed., pp. 143–155). Routledge.

Vig, S. (2024). Regulating deepfakes: An Indian perspective. *Journal of Strategic Security*, *17*(3), 70–93.

Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese*, *201*(1), Article 30. https://doi.org/10.1007/s11229-022-04012-2

Walker, M., & Winders, J. (2021). Where is artificial intelligence? Geographies, ethics, and practices of AI. *Space and Polity*, *25*(2), 163–166. https://doi.org/10.1080/13562576.2021.1985869

White, G. R. T., Samuel, A., Jones, P., Madhavan, N., Afolayan, A., Abdullah, A., & Kaushik, T. (2024). Mapping the ethic-theoretical foundations of artificial intelligence research. *Thunderbird International Business Review*, *66*(2), 171–183. https://doi.org/10.1002/tie.22368

Whitehead, H., & Gandhi, A. (2024, March 4). Feminist AI research network: Combatting gender-based violence with artificial intelligence innovations. *International Development Research Centre*. https://idrc-crdi.ca/en/research-in-action/feminist-ai-research-network-combatting-gender-based-violence-artificial

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation*, *125*, Article 102784. https://doi.org/10.1016/j.technovation.2023.102784

Zevop, A., & Ballet, S. (2025). Computing the face: From coloniality to control. *International Migration*, *63*(2), Article e70007. https://doi.org/10.1111/imig.70007

Zheng, G., Shu, J., & Li, K. (2025). Regulating deepfakes between Lex Lata and Lex ferenda—A comparative analysis of regulatory approaches in the U.S., the EU and China. *Crime, Law and Social Change*, *83*(1), Article 1. https://doi.org/10.1007/s10611-024-10197-z

About the Authors

Weijie Huang is a PhD candidate in media and culture studies at Utrecht University, affiliated with the Inclusive AI Lab. She has contributed to research on gendered digital harms and AI safety in the Global South, employing feminist, decolonial, and policy-focused approaches to advance inclusive AI governance.

Payal Arora, professor at Utrecht University and co-founder of the Inclusive AI Lab, is an award-winning digital anthropologist with two decades of Global South research. Listed among 100 Brilliant Women in AI Ethics (2025), she's lauded by Forbes as the "next billion champion" for her work on inclusive tech.

Marta Zarzycka (PhD) is a senior user experience researcher at Google, specializing in protecting users from egregious digital harms, including child safety and non-consensual intimate imagery. She develops victim-centric solutions in collaboration with abuse prevention teams, NGOs, and academic partners, advancing safe and inclusive technology design.