
Deepfake Abuse and Gendered Digital Creative Violence: Feminist AI Interventions from Mexico

Payal Arora, Ana María Miranda Mora, and Marta Zarzycka

Generative AI is transforming image-based sexual abuse and exposing fundamental limitations in existing trust and safety frameworks. While research has focused on the technical details of detection and platform governance, less attention has been paid to how generative AI changes the nature of harm itself. Drawing on an intersectional feminist study of interviews with 12 survivors, legal advocates, activists, technologists, and researchers in Mexico, this article introduces gendered digital creative violence to explain how generative AI weaponizes creative production to generate emotional, reputational, psychological, and structural harm. We show that deepfake abuse extends beyond synthetic content to encompass victim-blaming, institutional failure, evidentiary instability, and affective exhaustion, leaving survivors to shoulder the burden of digital safety. At the same time, feminist organizations have developed AI-enabled counter-infrastructures, most notably the survivor-support chatbot OlimpIA, demonstrating how AI can be redesigned around care, accompaniment, and collective protection. We argue that creative violence offers a transferable framework for understanding emerging forms of generative harm beyond deepfakes by shifting attention from reactive content moderation to the politics of creation itself. This perspective advances trust and safety scholarship by proposing feminist approaches to AI governance grounded in structural prevention, situated ethics, cross-sector collaboration, and survivor-centered design.

1 Introduction: When Generative AI Becomes a Technology of Creative Violence

Generative artificial intelligence (GenAI) has been heralded as a transformative force, unlocking new domains of social production, knowledge synthesis, and computational culture. Public and corporate discussions about GenAI emphasize innovation, speed,

scalability, and productivity. However, alongside these celebratory narratives lies a growing, deeply corrosive area of gendered harm: the synthetic production of “deepfake pornography” or non-consensual synthetic intimate imagery (NCSII) (Huang, Arora, and Zarzycka 2026). Unlike traditional image-based sexual abuse (IBSA), which relies on soliciting, leaking, hacking, or surreptitiously recording authentic intimate content, GenAI enables perpetrators to fabricate high-fidelity intimate content without needing prior access to private audio-visual material (Henry et al. 2020). This technological shift disrupts the foundational assumptions governing digital platforms. Much of the existing trust and safety infrastructure across tech companies is designed to moderate user-generated content by assuming that harmful material originates from pre-existing, real-world artifacts (Anciaux and Gramaccia 2025). Under traditional frameworks, once a non-consensual image is detected, it can be flagged, removed, and hashed using automated matching systems to prevent further spread. GenAI fundamentally challenges this paradigm in two ways. Technically, traditional perceptual hashing breaks down when confronted with the infinite generation of novel, unique content that lacks a prior digital signature. Socially and ethically, while the harm is compounded by exposure, circulation, and threats of sharing, the violation also originates intrinsically in the agentic act of creation itself (Tan et al. 2025).

This development requires a profound conceptual reorientation around the notion of “creativity.” In mainstream sociological and technological discourses, creativity carries overwhelmingly positive connotations linked to self-expression, artistic innovation, and economic empowerment. To capture this emergent form of weaponized production, this article introduces the concept of gendered digital “creative violence,” focusing specifically on its digital and generative AI-driven manifestations. We define gendered digital creative violence as the strategic deployment of generative technological capacity to fabricate sexualized, degraded, or weaponized representations that result in severe emotional, reputational, psychological, and structural damage. This violence is “creative” in a literal, generative sense: it brings new artifacts, narratives, and vectors of exposure into existence without pre-existing material.

Crucially, rather than suggesting that GenAI produces entirely unprecedented forms of misogyny, this article argues that generative systems reconfigure, automate, and intensify historically entrenched structures of gender-based violence through novel mechanisms of fabrication, cross-platform circulation, and profound evidentiary instability. At the same time, the synthetic nature of GenAI introduces distinct social dynamics around fault: because content is fabricated without requiring victims to have taken or shared intimate media, traditional victim-blaming narratives grounded in “reckless self-exposure” are partially disrupted. While survivors confront familiar patterns of victim-blaming, institutional neglect, and affective exhaustion, this shift can reduce self-blame and increase willingness to report incidents, even as generative tools lower the technical barriers to entry, enabling hyper-realistic harms to be manufactured at an industrial scale.

Mexico provides a critical, highly instructive site for examining these dynamics. Widely recognized as a global pioneer in the legal recognition of technology-facilitated sexual violence, Mexico adopted the landmark Ley Olimpia (Olimpia Law) reforms, a comprehensive legal framework that criminalizes violations of sexual privacy in digital spaces (Garza de la Vega and Escalera Silva 2025). However, institutional responses remain characterized by severe enforcement gaps, systemic impunity, and bureaucratic revictimization. In response to these state and platform failures, local feminist organizations and activists have built autonomous counter-infrastructures, most notably OlimpIA, a conversational AI chatbot designed to accompany survivors through complex legal, psychological, and emotional protocols (Huang, Arora, and Zarzycka 2026). The co-existence of pioneering legislation, profound institutional limits, and grassroots technological agency makes Mexico an essential case study for understanding the friction between generative harms and emergent modes of feminist resistance in the majority world.

This study investigates how different actors, including survivors, activists, legal advocates, technologists, and researchers, collectively interpret, navigate, and resist AI-enabled digital sexual violence within the Mexican sociolegal ecosystem. By combining a systematic qualitative approach with feminist epistemologies, this paper bridges the analytical gap between theoretical critiques of technology and the actionable demands of trust and safety practice.

2 Theoretical Framework

2.1 Creative Violence: Between (Re)creation and Insurgency

To unpack the political economy of generative harms, we build upon and bring into dialogue two distinct frameworks: Furtado's (2022) conceptualization of violence tied to an "economy of playfulness and pleasure," and Kraidy's (2016) theory of "creative insurgency." Furtado posits that violence should not be understood merely as an act of absolute destruction, obliteration, or physical injury. Instead, violence can operate as a highly productive, generative force, structuring spaces, consumer desires, and modes of life organized around entertainment, gratification, and playful consumption. In digital spaces, this manifests as an asymmetric economy of pleasure: perpetrators derive amusement, social capital, or sexual gratification from the creation and consumption of synthetic intimate imagery, secure in their anonymity, while simultaneously exposing marginalized subjects to radical insecurity, exposure, and structural harm.

Conversely, Kraidy's (2016) work on "creative insurgency" foregrounds aesthetic expression, media performance, and computational subversion as vital political tools used by marginalized groups to contest authoritarian and structural power. Building on these two frameworks, we argue that creativity is neither inherently emancipatory nor inherently violent; rather, it is a saturated and politically contested terrain of struggle. In the age

of GenAI, creativity becomes a double-edged sword. On one side, it weaponizes algorithmic systems to reproduce patriarchal dominance through synthetic image-based abuse. On the other hand, it enables the formation of feminist counter-infrastructures that deploy the same generative capacities for collective care, survivor accompaniment, and systemic disruption.

We explicitly address the tension inherent in using the term “creative” alongside “violence.” Critics may argue that applying creative to sexual violence risks romanticizing or softening the underlying brutality of the act. However, we contend that failing to recognize the generative dimension of GenAI harm leaves trust and safety paradigms stuck in a reactive loop. Creative violence demands that we analyze the politics of production: who possesses the infrastructure and capacity to create representations, whose bodily autonomy is treated as raw material for computational training data, and how algorithmic affordances lower the friction required to manifest deep-seated misogyny into hyper-realistic digital artifacts.

2.2 Contextualizing Digital Gender-Based Violence in Mexico

The scholarship on digital gender-based violence in Mexico has established that online harms are a direct extension of a historical continuum of systemic violence against women. Work by feminist media scholars and digital ethnographers (Vega Montiel et al. 2024; Salgado-Espinosa and Salgado-Espinosa 2022; Gómez Cruz 2023; Vega Montiel 2019) has demonstrated that practices like cyberstalking, trolling, and the unauthorized dissemination of intimate content are structural mechanisms used to police women’s visibility and speech in public and digital spheres. Within Mexico’s socio-political reality, where high rates of femicide represent the most extreme manifestation of patriarchal violence, digital spaces do not constitute a disconnected virtual reality. Rather, as the Comisión de Derechos Humanos de la Ciudad de México notes,¹ digital violence operates as an infrastructural amplification of real-world physical, psychological, and economic harm.

A parallel body of literature (Aguilar Pérez 2025; Garza de la Vega and Escalera Silva 2025) focuses on public policy and legislative interventions, analyzing the emergence and uneven implementation of the Ley Olimpia reforms. These scholars have charted the development of these legal frameworks, noting the historic victory of survivor-led cyberactivism in codifying digital harassment as a punishable offense. Yet, as Escalera Silva, Amador Corral, and Lara Hernández (2024) observe, the transition from formal law to judicial enforcement is severely hindered by institutional biases, bureaucratic inertia, and a widespread lack of digital literacy among prosecutors and judicial actors.

Emerging studies have begun mapping these risks—either conceptually (Pereira Hernández and Zatarain Avendaño 2024) or by empirically characterizing the technical capabilities, economics, and attacker tactics within deepfake creation ecosystems (Han et

1. <https://cdhcm.org.mx/>

al. 2025). However, the literature still lacks empirical, qualitative accounts of how survivors and local advocates navigate the specific epistemic and evidentiary challenges posed by synthetic fabrication. Furthermore, mainstream trust and safety research remains deeply Western-centric, focusing primarily on Global North platform dynamics while neglecting how local communities in the Global Majority build autonomous, AI-driven counter-tools. This article addresses these gaps by providing an empirically grounded analysis that links the unique affordances of GenAI abuse to localized feminist technological responses in Mexico.

2.3 Trust and Safety and the Limits of Content Moderation

For much of the early web era, trust and safety scholarship and industry practice positioned digital platforms as quasi-sovereign regulatory authorities that autonomously codified and enforced online behavioral norms (Van Dijck 2020; Huang, Arora, and Zarzycka 2026). In recent years, however, sweeping statutory interventions, most prominently the European Union's Digital Services Act (DSA), have fundamentally reconfigured this paradigm. As legal scholar Keller (2024) demonstrates, trust and safety has increasingly transitioned from voluntary self-regulation into an institutionalized corporate compliance function. Content moderation, automated flagging, hashing, and systemic risk assessments are now tightly bound to statutory auditability, reporting protocols, and legal liability, shifting platform governance from discretionary community management to regulated legal compliance. However, this established model operates primarily through a retrospective and reactive paradigm. Traditionally, platform intervention requires an offense to occur and content to be created before governance mechanisms are activated, whether through proactive automated filters that block violative content prior to user exposure or through retrospective user reporting and algorithmic flagging post-upload. While automated systems increasingly intercept and suppress known harms before human users ever encounter them, this paradigm remains fundamentally reactive to existing content patterns rather than predictive of novel, synthetic generation.

The advent of GenAI collapses this linear temporality. Synthetic media challenges traditional asset-based detection mechanics. When an abusive image can be generated with infinite variations in background, lighting, and composition, standard cryptographic hashing tools (such as PhotoDNA or MD5 hashes) fail, because each generated file represents a structurally distinct mathematical artifact despite conveying identical expressive harm. Non-consensual synthetic content is not simply a tech platform issue, but a reflection of global legal and cultural friction. Current platform policies already go beyond statutory baselines, yet struggle with competing extremes: standard definitions may overlook localized modesty violations, while accommodating hyper-conservative demands risks facilitating total censorship of women's depictions. Ultimately, effective governance requires country-specific legal frameworks that leverage local civil society organizations to facilitate enforceable, global takedown requests. For women in the Global South, digital safety cannot merely be a function of a "takedown button;" it

must encompass understandings of cultural norms and their relation to rights-based values, institutional reforms and enforcement capabilities, socio-linguistic accessibility and user behavior, and a fundamental realignment of platform architectures away from engagement maximization and toward expanding gender freedoms online.

3 Methods

3.1 Mexico as a Critical Case Study

3.1.1 Structural and digital gender-based violence in Mexico

Gender-based violence in Mexico is a widespread crisis driven by deep-rooted patriarchal norms, economic inequality, and high rates of impunity. Data from the 2021 National Survey on the Dynamics of Household Relationships (ENDIREH), published by the Instituto Nacional de Estadística y Geografía (INEGI),² reveals that an astonishing 70.1% of women aged 15 and older have experienced at least one form of violence over their lifetime. This violence permeates every social environment: 22.4% in community settings, 20.8% in the workplace, 20.2% in educational institutions, and 20.7% within intimate relationships.

Digital communication technologies have provided new vectors for these entrenched power dynamics. According to INEGI's 2024 Módulo sobre Ciberacoso (MOCIBA),³ more than 22.2% of women internet users in Mexico, representing approximately 10.6 million women, have been subjected to severe digital harassment. This cyber-harassment is profoundly gendered: while men frequently report text-based insults or threats related to fraud, women disproportionately face unsolicited sexual content and targeted attacks on their physical appearance. For young women aged 20 to 29, the digital space functions as a volatile terrain where everyday interactions are weaponized through practices of sexualized abuse that reinforce deep racialized and class-based hierarchies.

3.1.2 The Ley Olimpia reforms and the reality of enforcement

In response to grassroots mobilization led by activist Olimpia Coral Melo and the Frente Nacional por la Sororidad, Mexico passed federal legislative reforms in 2021 known collectively as the Ley Olimpia. Integrated into the Criminal Code and the Ley General de Acceso de las Mujeres a una Vida Libre de Violencia (LGAMVLV), this legal framework formally defines and criminalizes digital sexual violence (intentional creation, distribution, exhibition, or commercialization of intimate sexual content without explicit consent) as a distinct modality of gender-based harm (Aguilar Pérez 2025). The law primarily targets individual perpetrators through criminal penalties, including prison sentences of three to six years for non-consensual production, fabrication, or distribution of intimate

2. <https://www.inegi.org.mx/>

3. <https://www.inegi.org.mx/programas/mociba/2024/>

content. It also establishes protective measures that allow judicial authorities to issue administrative orders requiring digital platforms and online services to remove, block, or delete offending media. This framework aligns with regional standards, such as the Organization of American States' Inter-American Model Law to Prevent, Punish, and Eradicate Gender-Based Digital Violence against Women, which frames image-based sexual abuse as a direct violation of human dignity, autonomy, and freedom of expression.⁴

Despite this advanced legislative scaffolding, a stark asymmetry exists between formal criminalization and the material reality of enforcement. According to the Secretaría de las Mujeres de la Ciudad de México,⁵ between 2020 and 2024, authorities opened 3,205 investigation files for digital sexual violence. Yet approximately 95% of these investigations never reached the judicial stage. This systemic breakdown stems from an overreliance on a purely punitive framework that lacks parallel investments in forensic infrastructure, inter-institutional coordination, and gender-sensitive training for public prosecutors. Survivors navigating the judicial system face a bureaucratic gauntlet characterized by systemic victim-blaming, procedural delays, and the physical loss of digital evidence, leaving them trapped between the low responsiveness of global technology platforms and the structural failure of domestic legal institutions.

3.1.3 Research design

This study employs an intersectional feminist qualitative research design (Reinharz and Davidman 1992). Feminist methodology operates on the principle that knowledge is inherently situated, embodied, and bound up within operations of power. Rather than executing a top-down data extraction, this approach elevates the lived experiences of survivors and local practitioners as epistemically central to understanding socio-technical systems. Drawing on Crenshaw's (1991) framework of intersectionality, we examine how overlapping axes of gender, class, race, geography, and other vectors shape both exposure to digital harms and access to redress. In the Mexican context, for example, women journalists and activists outside major urban centers often encounter compounded risks of online abuse alongside weaker institutional protections and fewer reporting mechanisms. Our research group bridges the standard divide between critical academia and industrial application. The fieldwork was conducted by a Mexican woman researcher with extensive experience navigating local feminist networks and human rights frameworks. The Principal Investigator brings a long trajectory in transnational inclusive design, while our industry research partner works internally within a major technology company to advance victim-centered generative AI safety models. This positioning enabled a multi-layered analysis that accounts for grassroots survivor realities alongside structural platform capabilities.

4. <https://www.oas.org/en/>

5. <https://semujeres.cdmx.gob.mx/>

3.1.4 Participants, sampling, and sample characterization

Participants were recruited using purposive and snowball sampling strategies through trusted civil society intermediaries, human rights attorneys, and feminist accompaniment networks. This method was necessary given the sensitive and traumatizing nature of digital sexual abuse, ensuring that all participants entered the study voluntarily through a secure environment.

In total, 12 semi-structured interviews and 2 multi-layered expert consultations were conducted between February 2025 and September 2025. The participant cohort was explicitly disaggregated into clear stakeholder categories to provide context for the qualitative findings:

- **Victim-Survivors (n=6):** Women aged 19 to 34 residing in the metropolitan area of Mexico City and the State of Mexico. The sample included university students, a performing artist, and a professional dancer, reflecting a range of socioeconomic backgrounds and digital presence styles. Two of these survivors experienced high-fidelity AI-generated deepfake abuse and NCSII, while four experienced traditional image-based sexual abuse that was spread wider and faster by platform recommendation algorithms.
- **Civil Society and Feminist Digital Rights Organizations (n=2):** Senior advocates and technology policy leaders from prominent Mexican NGOs specializing in digital self-defense, cyberactivism, and tracking online misogyny.
- **Legal Advocates (n=2):** Defense attorneys specializing in strategic litigation under the Ley Olimpia, representing victims before public prosecutors in Mexico City.
- **Academic Experts (n=2):** University-based researchers focusing on digital sociology, critical masculinities, and the tracking of the digital “manosphere” in Latin America.
- **Expert Consultations (n=2):** Dedicated strategic consultation sessions held with a representative from an international human rights organization (UN Women Mexico) and a government official embedded in the Mexican administrative agencies’ task force on emerging technology and AI policy governance.

3.1.5 Data collection and interview infrastructure

Interviews lasted between 45 and 60 minutes and were conducted via encrypted communication platforms (e.g., Teams) or through safe, private, in-person settings chosen by the participants. Two distinct, semi-structured interview guides were developed to address the specific positionalities of the cohorts while maintaining an integrated thematic link.

The survivor-centric guide focused on: (a) patterns of routine digital interaction, (b) the

precise mechanics, temporality, and discovery of the digital violation, (c) the systemic evidentiary hurdles faced when attempting to document synthetic content, (d) interactions with automated platform reporting and state legal apparatuses, and (e) the overarching psycho-emotional, familial, and professional repercussions of the abuse.

The stakeholder-centric guide explored: (a) the systemic limitations of the Ley Olimpia, (b) corporate accountability, data opacity, and the structural limits of platform moderation, (c) the socio-technical affordances of emergent generative AI tools, and (d) concrete proposals to implement care-centered, trauma-informed trust and safety protocols. All interviews were recorded with authorization and transcribed verbatim in Spanish.

3.1.6 Ethics, data protection, and trauma-informed protocols

This study received formal institutional ethical approval and continuing review from the Research Ethics Committee at Utrecht University. Fieldwork was supported by an unrestricted academic award from Google. Participant recruitment and primary data coding were conducted independently by researchers at Utrecht University to maintain research integrity, while all co-authors contributed to the development and editing of the final manuscript.

Given the vulnerability of the participant group and potential for re-traumatization, the study implemented rigorous trauma-informed research protocols modeled on the cross-cultural safety principles outlined by Jessel, Beaulieu, and Rajaraman (2025). Rather than demanding explicit or graphic descriptions of the sexualized imagery, the interviews focused on the structural pathways of the abuse, institutional frictions, and coping responses. Written and verbal informed consent were obtained sequentially. Participants retained the right to skip any question, pause the session, or completely withdraw from the project at any stage without penalty.

All data were thoroughly anonymized during transcription. Real names, precise locations, employment details, and unique institutional markers were scrubbed and replaced with generic signifiers. To prevent unauthorized access or data leaks, all audio recordings and transcripts were stored using high-grade encryption protocols on Utrecht University's secure, access-controlled cloud architecture. In accordance with institutional regulations, the datasets will be securely retained for ten years before being permanently and automatically deleted.

3.1.7 Data analysis

The data were evaluated using Reflexive Thematic Analysis following the six-phase qualitative framework established by Braun and Clarke (2021). This methodology rejects the notion that qualitative themes simply “emerge” naturally from data; instead, it positions theme development as an active, deeply interpretive act executed by the researchers through iterative engagement with the texts.

The analytical process moved systematically from deep familiarization through verbatim reading to the creation of initial codes, then to the generation of themes, which were then reviewed against the entire dataset and formally defined. Coding operated through a dual structure: deductive codes derived from existing literature on trust and safety, platform architectures, and feminist legal critiques were applied alongside inductive codes that captured the lived realities, emotional syntax, and structural workarounds described by the participants. Reflexive research memos were updated after each phase to critically evaluate how our own transnational positionalities, institutional privileges, and theoretical backgrounds shaped the interpretation of the narratives.

4 Findings: Continuities and Transformations

4.1 Socio-Digital Infrastructures Reflect a Culture of Self-Responsibility and Self-Blame

A primary finding of this study is that contemporary digital infrastructures consistently mirror and enforce deep-seated cultural norms, shifting the burden of protection away from perpetrators and platforms onto victims. When GenAI or algorithmic spaces enable a violation, the immediate systemic and social response operates through an internalized logic of self-regulation and individual risk management. Survivors are forced to spend immense energy adapting their behaviors, hiding their identities, and policing their own digital footprints. One victim-survivor articulated this structural inversion of responsibility:

Sometimes, it's really frustrating to be the one who has to stay behind or find ways to avoid being violated. It's a bit shocking that I have to be self-conscious about it. I mean, I'm the one who needs to stop certain behaviors, right? Because their [Public Prosecutor Office] initial advice was about what I was doing. For example, when I was hiding on social media, they [Public Prosecutor Office] told me not to upload pictures of my body, only of my face, or to avoid sharing my location.

This narrative illustrates how trust and safety mechanisms and the social networks surrounding them rely on a self-restrictive logic of visibility. Instead of constraining the perpetrator's capacity to weaponize technology, the system demands that women shrink their digital presence. This dynamic is further reinforced by the victims' immediate family and social ecosystems. A prominent legal advocate specializing in Ley Olimpia cases explained that this internal culture of blame frequently undermines formal legal complaints:

They [survivor's family members] almost never accompany them [survivors] to file the complaints because there is always re-victimization by the family, saying it is their fault for taking the pictures. We have also had cases where

the moms, for the most part, tell them, ‘It’s okay,’ because in the end, you were trusting your partner or the person you were sharing with.

This finding shows that creative violence is not a bounded technological event; it is an extended socio-technical process. The initial generation of a deepfake or the unconsented distribution of an image triggers a secondary wave of structural violence managed through social judgment and oftentimes, institutional indifference. The moral economy of blame focuses on the victim’s original act of trusting someone or participating in digital culture, framing the generation of the image, rather than the violation of consent, as the true transgression.

This reality is not unique to Mexico; it reflects a broader global pattern where digital safety protocols mirror deeply entrenched patriarchal structures. Decades of cross-cultural research demonstrate that online gender-based violence is consistently met with victim-blaming logics that demand women restrict their digital expression, as documented in studies across India (Bhatia, Arora, and Pathak-Shelat 2021) and Saudi Arabia (Rabaan, Young, and Dombrowski 2021).

Platform designs formalize these boundary-setting dynamics through user-facing safety controls. Features such as locked profiles, granular privacy toggles, and safety nudges can play an essential role in user safety, providing targets of abuse with immediate agency and actionable controls over their online environments. Yet these features also entail trade-offs when they focus on containing the artifact of creative violence rather than the conditions that enable it. While locked profiles may reduce exposure, they also limit women’s visibility, voice, and participation, shifting the costs of safety onto those targeted by abuse. A structural tension arises if these user-side tools are treated as primary solutions rather than complementary safeguards. When safety architectures rely too heavily on target-side mitigation, the daily labor of protection risks being shifted onto potential victims, requiring women to continuously manage their self-presentation and treat digital public spaces as inherently hostile environments, even when platforms work to strengthen proactive, upstream safety measures.

4.2 Gender Harms Online Remain a Structural Gender Issue

The qualitative data demonstrates that the visibility, validation, and collective outrage surrounding digital abuse remain profoundly fractured along gender lines. While female and queer networks quickly recognize creative violence as a structural violation, male-dominated social and familial spaces often respond with public silence, minimization, or defensive indifference. A survivor detailed this asymmetry when describing how different family members responded after she shared her case publicly on social networks to warn her community:

My gay cousin is very angry; he told me, ‘I hate him [x], I hate him.’ ...I don’t know...he [the aggressor] needs to be called out [*quemarlo*] on social media.

Stand up and pursue justice.’ Since we’ve made this public, that cousin was the only one to share it with the men in the family. The other two, who are straight, one being my cousin’s brother, didn’t share anything. I think that reflects straight white male privilege, totally. It’s a disgusting reality that men have... no one has commented or shared the video. Someone even messaged me privately, but no one commented publicly or at all. It was like, ‘Gee, what’s happening? What’s going on?’ I believe it’s a privilege because, ultimately, it doesn’t impact them directly. They’ve never truly felt this violence firsthand, even if they might be empathetic. Every woman I know has experienced assault, abuse, or harassment by a man in some form.

This narrative underscores that the harm of digital abuse is determined by how it is received and validated by the victim’s social network. The straight male family members’ refusal to publicly engage with or share the denunciation reflects a protective privilege that insulates men from the digital continuum of violence. Their silence functions as a form of passive normalization, leaving the structural weight of the harm to be carried by the survivor and her allies.

This social normalization is often masked as strategic advice within the family. Multiple survivors reported that when they approached their parents or male peers for help after discovering abusive messages or synthetic content, they were systematically told to minimize the issue. As one survivor recounted:

I told my mother what had happened; it was about a year ago. My mother replied that I should ignore it. My mother always wanted me to maintain a firm stance, as if I didn’t look vulnerable to him, acting as if I didn’t care. Believe me, though, it was very difficult, because those messages would come suddenly, just when I least expected them, whether it was a text or a call like that.

The social advice to “ignore it” or “act as if you don’t care” functions as a double-edged sword: while intended to deny the perpetrator satisfaction, it requires survivors to absorb ongoing harassment. While these tools offer essential, immediate friction, they also illustrate how creative violence extends beyond content creation, requiring survivors to continually manage its effects and highlighting the need for continued advancement in automated detection, particularly for novel, synthetic harms for which pre-existing image hashes do not exist. Within platform safety architectures, this dynamic intersects with how users navigate available interventions. Major platforms provide dedicated reporting mechanisms alongside cross-industry proactive measures such as StopNCII’s hash-matching system to mitigate non-consensual intimate imagery. However, empirical research shows that target behavior often favors immediate user-level controls like muting or blocking to regain agency and avoid administrative delay. While these tools offer essential, immediate friction, relying primarily on user-driven mitigation highlights the need for continued advancement in automated detection, particularly for novel,

synthetic harms for which pre-existing image hashes do not exist.

4.3 Gendered Platforms and (Non)Reactive Governance

The architectures of mainstream generative software and social media platforms contain distinct affordances that channel user behavior in gendered directions, lowering the technical friction required to sexualize women's bodies while maintaining default protections for male subjects (Toupin 2024). A civic activist detailed how their organization's digital forensic investigation into popular AI-enabled "clothes-off" modification applications revealed a deeply biased design structure:

When we began our investigation, we found that the platform's advertising was primarily aimed at altering women's bodies. Although AI could modify bodies regardless of gender, the promotional content specifically targeted women, with messages like 'Undress women, place them in various sexual positions, and alter their bodies.' There was also a question: 'Can I modify men?' Yes, men could be modified too, changing the color of their suits, adding ties, or changing their shoes. It upset me to realize that, despite being an app for modifying advertisements, it had a clear bias—women's bodies were sexualized, while men's modifications were limited to simple changes like suit color or shape.

Under our framework of creative violence, these tools do not merely reflect pre-existing societal misogyny; they actively automate and scale it. The design logic transforms women's bodies into modifiable objects available for digital consumption, while restricting modifications of men to professional adjustments like suit color or styling.

This algorithmic objectification contrasts with the historically reactive, slower responses of platforms, affecting trust and safety governance. While platforms have long deployed proactive, automated matching systems to detect terrorism and child sexual abuse material (CSAM) under strict statutory mandates, they historically treated gender-based deepfakes and NCSII with far less institutional urgency. Moreover, while platform governance operates within distinct national legal frameworks, the technical infrastructures through which Mexican users experience safety are largely designed by global platform companies and deployed across markets. Consequently, improvements to automated detection, reporting workflows, and evidence preservation developed for one jurisdiction can shape protections for users well beyond the country in which they originate. An academic expert tracking the Latin American manosphere and its link to extreme violence shared an instructive case from Michoacán that illustrates how creative violence escalates into physical harm within less moderated platform spaces:

I worked on a case in 2019 in Morelia, Michoacán. A student named Diego Ulrich had a girlfriend named Marisela. Diego killed her violently and discarded her body in a plastic bag... But then some of the evidence that led to his trial

and conviction became semi-public. In a private Facebook group with his university classmates before Marisela was murdered, they shared memes with her face in which they talked about her murder before it happened. There's a particularly disturbing one that shows her face on a plastic bag, with the words 'Marisela all dead' written on it, implying she will be murdered soon. I believe that these interactions [such as creating images with AI and sharing or distributing them] occur in closed spaces. They occur in forums that they organize themselves. These Facebook groups have both a visible side and an invisible side.

This case underscores that violence is frequently a creative, calculated, and collaborative act of intimidation nurtured within semi-private or encrypted digital infrastructures. Perpetrators use these spaces to share modified images, coordinate harassment campaigns, and desensitize communities to violence. Also, this reveals how globally standardized platform architectures can shape local experiences of harm. Meta's commitment to protecting encrypted communication creates semi-private spaces that can facilitate the coordination and normalization of creative violence while limiting the visibility required for timely trust and safety interventions. The challenge remains for national regulators: how to respond to harms arising from platform infrastructures designed and governed beyond national borders. Another civic activist described a similar case where digital tools were used to build a coordinated campaign of psychological torture against a sexual assault survivor:

We once had a case where they made a video game about a girl who was raped in real life. The game featured the girl as the avatar. They created an avatar of that girl and her peers who attacked her and had them participate in a game where they raped, tortured, and mutilated her. They sent the screenshots to her WhatsApp group, saying, 'Look, today we're going to rape you.'

These targeted attacks exploit critical features of mainstream messaging and social media platforms that practitioners perceive as difficult to moderate, document, or preserve as evidence. Rather than challenging formal content policies, activists emphasized the practical limitations of platform interfaces and response times when supporting survivors through reporting and legal processes, as one civic activist shares:

Much depends on the community's criteria and standards regarding content removal. Since sharing intimate and sexual images is banned, removal is generally straightforward. However, the primary challenge is often the time it takes to actually remove the content. For example, if it's an image uploaded in a post or a story, it falls under stories. Instagram Stories, in particular, do not have reporting forms; you can report them, but Meta will not take action faster than within the 24 hours they are visible. Additionally, there is no option to copy the story's URL, so you can't save it as evidence for a future complaint. This is why stories are more often used to commit this type

of aggression—they disappear after 24 hours, leaving no trace. Without any trace, it's impossible to detect what was done or not done, making verification impossible.

From the perspective of practitioners, ephemeral platform affordances amplify creative violence by undermining survivors' ability to preserve evidence and anticipate timely intervention. These perceptions, whether they accurately mirror platform implementation, shape protective practices and reinforce the sense that platform governance lags behind the temporal dynamics of abuse.

Also, feminist tech activists face immense frustration when trying to engage with corporate platforms to fix these loopholes. A representative from a local digital rights organization noted that engagement with major companies is often superficial:

We have been in contact with Meta regarding their gender initiatives, but we believe these efforts are more of a simulation. Although they launched pilot programs, these initiatives have not progressed and are now seemingly nonexistent.

Similarly, stakeholders described Telegram as “possibly the most opaque entity world-wide,” arguing that its limited cooperation with local authorities and reporting infrastructure makes digital justice practically inaccessible, although this perception predates the 2024 arrest of co-founder Pavel Durov. While platforms that host unencrypted public spaces often remain insulated from the material consequences of digital abuse, the governance of private and encrypted spaces presents a critical tension. Although encrypted channels can shield perpetrators, they also provide vital privacy and safety protections for survivors, activists, and advocates navigating increasingly authoritarian contexts.

4.4 Affective Exhaustion, Self-Censorship, and the Spiral of Distrust

The continuous labor of navigating fragmented legal systems and defensive platforms causes severe emotional depletion for survivors. The impact of creative violence extends far beyond the moment an image is circulated; it triggers a profound erosion of personal trust, a forced withdrawal from public and digital life, and an existential desire to become invisible. A survivor detailed how the trauma reordered her basic practices of motherhood and digital self-expression:

I went an entire year without sharing any updates and stayed off social media. The only thing I shared was the case; we made it public, anonymizing it... with my baby, I would never upload pictures now. I believe I would have been a different mother before; if I hadn't experienced this, I probably would have shared photos of my baby.

Another survivor described the exhausting, hyper-vigilant routine she had to adopt to manage her personal visibility online after being targeted:

I stopped uploading many things. Sometimes, I would upload something and check who viewed it. If I saw a profile I didn't recognize, I would block it. Currently, I have about 200 blocked profiles, simply from checking who has seen my posts.

This state of constant alertness shows that, under a reactive governance paradigm, survivors are forced to serve as the manual moderators of their own lives. The psychological harm alters how they inhabit physical spaces and present their bodies to the world. One survivor explained how the violation destroyed her confidence and forced her to change her physical wardrobe to avoid public judgment:

My confidence was also affected by how I dressed—before, I would wear loose tops and shorts, feeling freer. People even said that I cover myself more now. I changed my wardrobe to wear sweaters and pants because I didn't want to be seen or pointed out. So, yes, I lost a great deal of confidence.

The trauma of synthetic exposure creates an overwhelming sensation of digital nakedness, where survivors feel as though their bodies have been rendered permanently visible and modifiable by the public eye. A legal advocate who accompanies survivors of technology-facilitated violence in Mexico City explained this behavioral shift:

Many women withdraw from their social circles and alter their routines; they ignore messages from certain numbers and stop using their phones, claiming they communicate only with family. These changes in their environment and lifestyle result from violence, which affects how they live... They feel as if everything about them is visible, whether it's varicose veins or stretch marks, making everything seem to come out into the open. This sensation can be so overwhelming that it makes them want to disappear from the world.

In its most acute manifestations, this structural paranoia leads survivors to treat their everyday communication devices as active threats. One survivor recalled, "I became really scared and wrapped my phone in aluminum foil. I told my whole family that if we talked about it [the case], we should hide all our phones."

This spiral of distrust shatters close personal relationships, creating an enduring barrier to intimacy and community. As another participant summarized, "I don't trust men anymore; not even my boyfriend inspires confidence in me." Under our framework of creative violence, this total destabilization of identity and social connections is the core structural effect of the abuse. Synthetic violence re-scripts the survivor's future, compelling her to shrink her presence, self-censor her expression, and live in a state of isolation to manage the risk of future exposure.

4.5 “Cyber-Attacks” and the Politics of Intentionality and Casualization

Victims do not experience deepfake generation as random online trolling or an accidental byproduct of a neutral technology. Instead, they experience them as targeted attacks designed to destroy their identity, professional legitimacy, and public authority. This framing challenges mainstream platform discourses that individualize digital harm as mere interpersonal disputes or simple violations of content guidelines (Lazard et al. 2025).

The creation of a deepfake is a deliberate act of representation that uses the power of high-fidelity fabrication to silence specific voices. One digital rights activist noted that hate speech and synthetic manipulation scale rapidly when targeting women who organize around public or political issues: “We receive a lot of hate speech on the pages of the networks we run. A lot. This is especially true when topics about gender diversity are posted.” The intentionality behind these attacks often manifests in direct psychological harassment, where perpetrators explicitly use the synthetic nature of the image to torment the victim. A survivor described how an abuser downloaded normal public photos from her profile, applied an automated filter, and immediately weaponized the result:

They [pictures] were the ones that he downloaded from the networks. He probably just applied a filter to them, but it looked like he took a screenshot. The next day, he texted me and asked, ‘How does it feel to know that I’m now molesting minors?’

However, the deliberate targeting by perpetrators is consistently met with a process of institutional casualization by state authorities. While the attack is precise and calculated, the judicial system treats the violation as a low-priority administrative issue or an inconvenience. A survivor described a shocking failure in the chain of custody for a digital case that demonstrates this structural indifference:

[...] the treatment of the prosecutors is very poor. They also lost the USB drive containing sensitive information, despite knowing how delicate it was. With a casual attitude, they claim they can’t find it and ask for another copy. How impressive is that? You understand what’s inside; it can’t be lost if it’s truly in the chain of custody, as they claim to be. So, the reality is, yes, it’s been quite victimizing.

This casual attitude by state officials diminishes the political gravity of the abuse, reducing a violation of privacy to a series of bureaucratic errors. By losing sensitive digital evidence or forcing survivors to repeatedly submit explicit materials to un-trained clerks, the state can actively revictimize the survivor. Intentional violence and institutional delay work together: perpetrators exploit platform features to create untraceable harms, while state institutions use bureaucratic slowness to avoid accountability. Creative violence is thus sustained not just by the perpetrators who click “generate,” but by the slow, and at times

Table 1: Continuities and critical transformations between traditional forms of gendered digital violence and those enabled by GenAI.

	Traditional Digital Violence	GenAI Abuse (Deepfakes/NCII)
Source Material & Consent	Requires an authentic intimate image obtained via hacking, theft, secret capture, or broken trust.	Requires no prior authentic intimacy; fabricated out of public, mundane facial/body data.
Evidentiary Instability	Verification is stable; the artifact is a real file that can be hashed and tracked across networks.	High evidentiary instability; variations in pixel composition defeat standard hashing tools.
Scalability of Harm	Limited by the availability of authentic media assets possessed by the perpetrator.	Scaled infinitely; variations can be generated repeatedly with little technical skill.
Socio-Cultural Continuities	Perpetuates victim-blaming, structural misogyny, and institutional neglect.	Perpetuates the same structural misogyny, victim-blaming, and institutional neglect.

indifferent networks of response that treat gendered digital harm as an inevitable feature of the modern internet.

4.6 Comparing Generative AI Abuse and Non-Generative Digital Violence

To address whether GenAI harms require a new analytical framework, this study outlines clear continuities and critical transformations between GenAI abuse and traditional, non-generative digital violence (Table 1).

This comparison demonstrates that while the underlying cultural motivations and structural inequalities remain unchanged, GenAI introduces a qualitative transformation in the speed, scale, and epistemic uncertainty of digital harm.

4.7 Feminist Counter-Infrastructures: OlimpIA

In response to the systemic failures of both platform moderation and state legal systems, Mexican feminist movements have developed autonomous technological interventions, contributing to a growing ecosystem of digital victim-support AI chatbots designed to assist with gender-based violence (Henry, Witt, and Vasil 2025; Tan et al. 2025). The most prominent example is OlimpIA (Figure 1), a conversational tool deployed on WhatsApp through a collaboration between the Ley Olimpia movement, Digital Defenders, and AuraChat.Ai. Designed around trauma-informed accompaniment rather than corporate data collection, OlimpIA operates across four core axes: psycho-emotional containment, accessible legal orientation, digital safety guidance, and community assistance. Rather than replacing human care, its community support function acts as a warm referral bridge, triaging survivors and connecting them directly with vetted local feminist advocate networks, legal clinics, and peer support groups. Human legal and psychological experts regularly audit its outputs to mitigate algorithmic bias and maintain safety.

OlimpIA can be understood as a form of “creative insurgency” (Kraidy 2016), or what we

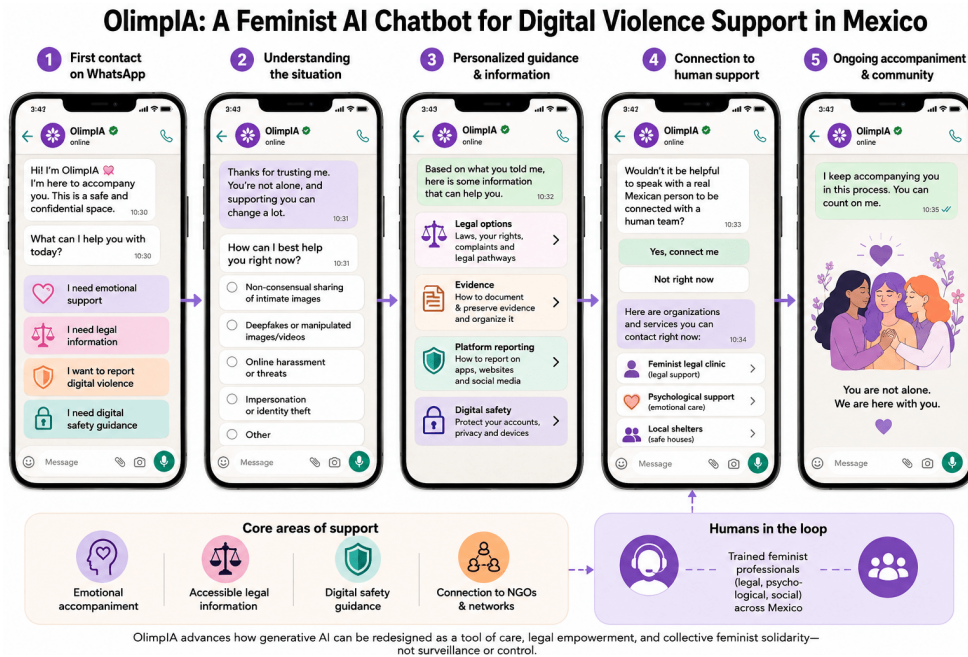


Figure 1: Illustrative Representation of Olimpia's Interface and User Flow. This figure is intended for illustrative purposes and does not depict the application's exact interface or workflow.

call “feminist counter-creativity.” It proves that generative technological capacities are not inherently tied to objectification or surveillance; instead, the same conversational and computational tools used to automate harassment can be redesigned to provide collective care and navigate legal systems. One survivor shared how finding this network of digital solidarity transformed her recovery process:

I've received substantial support from many people, including Olimpia, who shared my situation. She kept me updated, and I responded: without YOU, our dreams wouldn't be possible. I'm grateful to those I haven't even met personally but who have shared my story on social media without hesitation, offering their support. I've also noticed that some close friends didn't share anything, which shows who truly is part of my tribe. Sometimes, those you least expect are the ones who share and advocate in the comments, not necessarily those who've been with you all your life.

This testimony demonstrates that autonomous feminist technology works because it is embedded within a broader social infrastructure of solidarity. It helps survivors build alternative spaces of trust, a digital “tribe,” that counteracts the isolation caused by abuse.

However, these feminist counter-infrastructures still face structural barriers built into mainstream platform code. When activists and survivors attempt to use commercial networks to raise awareness, share safety resources, or denounce perpetrators, they

run directly into algorithmic suppression. As another digital activist observed:

Many times, we try to publish content, such as videos, and we have realized that when we use words related to sexual issues or violence, the volume automatically decreases, or it doesn't receive as many views.

This automatic downranking reveals a fundamental flaw in platform moderation. In their attempt to suppress explicit material, automated corporate filters censor the speech of survivors and the educational content of digital rights organizations. This algorithmic suppression silences survivors, blocking their efforts to find public support and legal redress. By showing how corporate platforms limit grassroots activism, our analysis proves that safety cannot be solved by automated moderation alone; it requires a political reconfiguration of technology design centered around transparency, survivor agency, and care.

5 Discussion and Conclusion

5.1 From Reactive Moderation to Feminist Trust and Safety Protocols

Our findings demonstrate that prevailing trust and safety paradigms remain structurally ill-equipped to address generative digital creative violence. Existing governance models remain predominantly reactive, focusing on removing content after harm has occurred rather than on disrupting the infrastructures that enable abuse. We argue that generative AI requires a shift from reactive moderation to structural containment, embedding friction throughout the content lifecycle and replacing claims of algorithmic neutrality with situated, context-sensitive governance that recognizes how gendered harms are shaped by local cultural, legal, and linguistic contexts. Effective platform governance, therefore, requires care-centered ecosystems that strengthen collaboration among platforms, legal institutions, and feminist civil society while redistributing the burden of safety from survivors onto platforms.

Beyond the Mexican context, we propose creative violence as a transferable framework for understanding how generative AI transforms harm from isolated content violations into dynamic, participatory, networked processes. Drawing on Crenshaw's (1991) intersectional framework, it recognizes that intersecting inequalities shape exposure to AI-mediated harms and access to redress. We therefore propose four shifts, from reactive removal to structural containment, from abstract neutrality to situated ethics, from corporate isolation to cross-sector integration, and from profit-centered optimization to care-centered design, offering a feminist trust and safety framework that is both analytically useful and practically actionable for platform designers, policymakers, researchers, and civil society seeking to build more equitable AI governance.

5.2 Epistemic and Policy Contributions

This study contributes to transnational feminist technology studies and trust and safety research by introducing gendered digital creative violence as a framework for understanding how AI's generative capacities reproduce structurally gendered harms. Rather than focusing on algorithmic bias or data privacy, it shifts attention to the politics of generation and the weaponization of computational creativity.

Our findings challenge the assumption that harms from GenAI can be addressed through content moderation alone. Deepfake violence is an extended process sustained by victim-blaming, institutional casualization, and platform architectures that prioritize engagement over safety. The Mexican case demonstrates that legal reform requires structural transformation across state institutions and digital platforms. Finally, OlimpIA offers an empirical model of feminist counter-creativity, showing how generative AI can be reoriented toward care, accompaniment, and collective solidarity.

5.3 Limitations and Future Research Directions

This study has several limitations. First, it draws on a small, purposively selected qualitative sample (12 interviews and 2 expert consultations), providing depth but not statistical representativeness. Second, it was limited to fieldwork in Mexico City and the State of Mexico, limiting the transferability of the findings. Third, the study does not examine perpetrators, consumers of synthetic media, or the social networks that perpetuate victim-blaming. Finally, it could not assess platform algorithms, moderation systems, or internal decision-making.

GenAI embodies a socio-technical paradox: the computational capacities that enable creativity also facilitate new forms of violence. The Mexican case demonstrates that deepfake abuse cannot be addressed through content moderation alone, as harm is rooted in systems of design, circulation, and scale. Digital safety therefore requires infrastructural accountability grounded in structural containment, situated ethics, and care-centered design. Drawing on Global Majority innovations such as OlimpIA, future AI governance should prioritize human dignity, collective care, and systemic responsibility over reactive harm mitigation.

References

- Aguilar Pérez, Eliud. 2025. "Transformación Jurídica ante la Violencia Digital contra la Mujer en México: Ley Olimpia y Protección de sus Derechos Fundamentales." *Ciencia Latina Revista Científica Multidisciplinar* 9 (6): 8584–605. https://doi.org/10.37811/cl_rcm.v9i6.21981.
- Anciaux, Arnaud, and Julie A. Gramaccia. 2025. "Imagining Markets and Crafting Value: The Emergence of an AI-Generated Pornographic Content Ecosystem." *Porn Studies*, 1–16. <https://doi.org/10.1080/23268743.2025.2492332>.
- Bhatia, Kiran Vinod, Payal Arora, and Manisha Pathak-Shelat. 2021. "Good Girls Don't Go Online: Unpacking the Quotidian Playful Resilience Influencing Girls' Social and Digital Engagements." *International Journal of Communication* 15:4755–73. <https://ijoc.org/index.php/ijoc/article/view/17552>.
- Braun, Virginia, and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide*. SAGE.
- Crenshaw, Kimberlé. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color." *Stanford Law Review* 43 (6): 1241–99. <https://doi.org/10.2307/1229039>.
- Escalera Silva, Luz Alejandra, Sandra Rubí Amador Corral, and Yuri Marisol Lara Hernández. 2024. "Digital Gender-Based Violence from the Experience of Actors in the Delivery of Justice." *Sapienza: International Journal of Interdisciplinary Studies* 5 (1). <https://doi.org/10.51798/sijis.v5i1.741>.
- Furtado, Henrique Tavares. 2022. "(Re)Creative Violence." *Critical Studies on Security* 10 (2): 106–9. <https://doi.org/10.1080/21624887.2022.2111839>.
- Garza de la Vega, Daniel Alberto, and Luz Alejandra Escalera Silva. 2025. "Substantive Analysis of Digital Violence in Mexico: Olimpia Law, a Case Study in Nuevo León." *Trans America Review* 3 (1). <https://doi.org/10.62910/transame25003>.
- Gómez Cruz, Brenda Magali. 2023. "Lo digital es político: Universitarias frente a la violencia digital hacia las mujeres." *Revista Pueblos y Fronteras Digital* 18:1–28. <https://doi.org/10.22201/cimsur.18704115e.2023.v18.640>.
- Han, Catherine, Anne Li, Deepak Kumar, and Zakir Durumeric. 2025. "Characterizing the MrDeepFakes Sexual Deepfake Marketplace." In *34th USENIX Security Symposium (USENIX Security 25)*, 5169–88.
- Henry, Nicola, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Adrian J. Scott. 2020. *Image-Based Sexual Abuse: A Study on the Causes and Consequences of Non-Consensual Nude or Sexual Imagery*. Routledge. <https://doi.org/10.4324/9781351135153>.

- Henry, Nicola, Alice Witt, and Stefani Vasil. 2025. "A 'Design Justice' Approach to Developing Digital Tools for Addressing Gender-Based Violence: Exploring the Possibilities and Limits of Feminist Chatbots." *Information, Communication & Society* 28 (11): 1884–907. <https://doi.org/10.1080/1369118X.2024.2363900>.
- Huang, Weijie, Payal Arora, and Marta Zarzycka. 2026. "Geographies of Hope: Rethinking Deepfake Harms and Gender AI Safety in the Global South." *Media and Communication* 14. <https://doi.org/10.17645/mac.10969>.
- Jessel, Joshua, Lauren Beaulieu, and Adithyan Rajaraman. 2025. "How to Conduct More Culturally Responsive and Trauma-Informed Functional Assessment Interviews: A Tutorial." *Behavior Analysis in Practice* 18 (4): 1267–82. <https://doi.org/10.1007/s40617-025-01047-y>.
- Keller, Daphne. 2024. "The Rise of the Compliant Speech Platform." *Lawfare*, October 16, 2024. <https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform>.
- Kraidy, Marwan M. 2016. *The Naked Blogger of Cairo: Creative Insurgency in the Arab World*. Harvard University Press.
- Lazard, Lisa, Rose Capdevila, Emma L. Turley, Kathryn Gilfoyle, and Nelli Stavropoulou. 2025. "Deepfake Technology and Gender-Based Violence: A Scoping Review." *Trauma, Violence, & Abuse*, <https://doi.org/10.1177/15248380251384271>.
- Pereira Hernández, María Luisa, and Virginia Mirella Zatarain Avendaño. 2024. "Pornografía deepfake en la era de la IA: nuevos desafíos para la educación de género, humanística y tecnológica." *RECIE. Revista Electrónica Científica de Investigación Educativa* 8. <https://doi.org/10.33010/recie.v8i0.2337>.
- Rabaan, Hawra, Alyson L. Young, and Lynn Dombrowski. 2021. "Daughters of Men: Saudi Women's Sociotechnical Agency Practices in Addressing Domestic Abuse." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW3): 1–31. <https://doi.org/10.1145/3432923>.
- Reinharz, Shulamit, and Lynn Davidman. 1992. *Feminist Methods in Social Research*. Oxford University Press.
- Salgado-Espinosa, Laura Anahí, and Mayra Lizeth Salgado-Espinosa. 2022. "Violencia digital contra las mujeres en México: Caracterización, efectos, experiencias y redes." *Femeris* 7 (3): 29–42. <https://doi.org/10.20318/femeris.2022.7150>.
- Tan, Yuying, Karolien Poels, Heidi Vandebosch, Sara Pabian, Nicola Henry, and Alice Witt. 2025. "'It's not that I don't trust the bot, but can it actually help?' Listening to Young Adults' Experience with the Transparency Designs of an Image-Based Sexual Abuse Victim Support AI Chatbot." In *Proceedings of the International Conference on Information Technology for Social Good (GoodIT '25)*, 60–69. <https://doi.org/10.1145/3748699.3749775>.

- Toupin, Sophie. 2024. "Shaping Feminist Artificial Intelligence." *New Media & Society* 26 (1): 580–95. <https://doi.org/10.1177/14614448221150776>.
- Van Dijck, José. 2020. "Governing Digital Societies: Private Platforms, Public Values." *Computer Law & Security Review* 36. <https://doi.org/10.1016/j.clsr.2019.105377>.
- Vega Montiel, Aimée. 2019. *Ciberviolencia contra las mujeres y discurso de odio sexista*. Ciudad de México: Instituto Electoral de la Ciudad de México.
- Vega Montiel, Aimée, Daniela Carolina Esquivel Domínguez, Adina del Carmen Barrera Hernández, and Carolina Pacheco Luna. 2024. "La violencia sociodigital contra las mujeres." *Atlánticas. Revista Internacional de Estudios Feministas* 9 (1): 1–31. <https://doi.org/10.17979/arief.2024.9.1.9985>.

Authors

Payal Arora (PhD) is a Professor of Inclusive AI Cultures at Utrecht University and Founder of the Inclusive AI Lab. She is a digital anthropologist and author of award-winning books, including *The Next Billion Users* (Harvard Press) and *From Pessimism to Promise* (MIT Press). Email: p.arora@uu.nl.

Ana María Miranda Mora (PhD) is an Assistant Professor of Gender and Postcolonial Studies at Utrecht University, working at the intersection of law and violence, power and the state, gender-based violence, care work, and feminist–Marxist thought, with a particular interest in Hegel and Marx. Email: a.m.mirandamora@uu.nl.

Marta Zarzycka is a Research Lead at Google, specializing in AI safety, abuse prevention, and high-risk online harms. Her expertise spans GenAI safety, violent extremism, non-consensual intimate imagery, child safety, and research with victim-survivors and at-risk users. Email: martazarzycka@google.com.

Acknowledgments

This publication is part of the Google Awarded Gen(der) AI Safety project with Professor Payal Arora as the Principal Investigator, Dr. Ana María Miranda Mora as the Mexico Field Lead, 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. We sincerely thank the survivors and stakeholders who shared their experiences, reflections, and expertise with us. Their trust and knowledge made this research possible and profoundly shaped the understandings developed in this article.

Data Availability Statement

The qualitative interview data are not publicly available because they contain sensitive and potentially identifying information relating to experiences of gender-based and sexual violence. Participants did not consent to public release of the interview transcripts, and sharing the underlying data could pose risks of re-identification or re-traumatization. Anonymized excerpts are included in the article where appropriate. The interview protocols are made available as supplementary material.

Funding Statement

Fieldwork was supported by an unrestricted academic award from Google.

Ethical Standards

This study received formal institutional ethical approval and continuing review from the Research Ethics Committee at Utrecht University.

Keywords

Deepfakes; NCSII; gendered digital creative violence; gender-based violence; Mexico.