SYNTHETIC VIOLENCE: DEEPFAKES, GENDER HARM, AND THE LIMITS OF AI GOVERNANCE

Aradhya Jain, Advocate, District Courts, Patiala

ABSTRACT

Generative AI and synthetic media have rapidly transformed from experimental tools into core infrastructure for digital life, but this paper argues that their most urgent implications lie not in efficiency gains but in the escalation of gendered and epistemic harm. Focusing on deepfakes and other forms of synthetic media, it shows how these technologies intensify technology-facilitated gender-based violence (TFGBV), including non-consensual sexual imagery, sextortion, impersonation, and targeted harassment, while simultaneously eroding the conditions for trusting information, a phenomenon conceptualized as epistemic pollution and the "truth recession."

Drawing on Fricker's framework of testimonial and hermeneutical injustice, the analysis demonstrates how AI-enabled abuse disproportionately targets women and marginalized communities, undermining their credibility, silencing their testimony, and denying them the conceptual resources needed to name and contest new forms of violence. The paper situates these harms within broader structures of synthetic violence, algorithmic bias, and platform capitalism: biased datasets and engagement-driven recommender systems reproduce racial and gender hierarchies, while platform economies and moderation failures turn social media intermediaries into active amplifiers of misogyny and disinformation rather than neutral hosts. It then critiques contemporary AI and platform governance as a set of fragmented "regulatory islands": the EU's risk-based AI Act, the United States' sectoral patchwork, the UK's principles-based model, China's security-oriented deep synthesis rules, and regulatory uncertainty elsewhere, which collectively fail to address cross-border TFGBV and digital misogyny.

In response, the paper advances a gender-responsive AI governance agenda grounded in survivor-centred frameworks, gender-sensitive moderation, safety-by-design, transparency, accountability, stronger consent and data protections, digital literacy, and intersectional oversight across the AI lifecycle. Taken together, these interventions reframe AI governance as a project of redistributing epistemic and digital power, positioning the prevention of gendered synthetic violence as a central, rather than peripheral, objective of global AI regulation.

Keywords: Generative AI; Deepfakes; TFGBV; Epistemic Injustice; Algorithmic Misogyny; Platform Governance; Feminist AI Governance; Digital Harm.

1. Introduction

The evolution and adoption of generative artificial intelligence (GenAI) represents the most significant technological advancement since the early 1990s. While the concept of AI originated in the 1950s, it garnered sporadic mainstream interest until November 2022, when OpenAI launched ChatGPT-3.5, making advanced GenAI readily available to the public and dramatically increasing awareness of its capacity to revolutionize both business and daily life.[1]

The figures are, quite frankly, astonishing. Within a span of just over two years, OpenAI's valuation surged to over US $300 billion by April 2025.[2] Meanwhile, Nvidia's market capitalization skyrocketed from below US $400 billion in late 2022 to around US $5 trillion by 2026. Additionally, Microsoft, Google, Meta, and Amazon together invested more than US $246 billion in AI infrastructure in 2024 alone.[3] By 2025, this amount was anticipated to surpass US $320 billion, with expenditures in 2026 expected to rise even further, primarily fueled by an unquenchable demand for GPUs and data center capacity. The present situation resembles less a competitive business landscape and more a technological land grab, driven by a shared and urgent conviction that whoever builds the best AI infrastructure today will define the economy of tomorrow.[4]

GenAI's capability to generate hyper-realistic images, videos, and text that are nearly indistinguishable from reality addresses a core concern: the human tendency to trust what we observe and read. When this instinct is manipulated on a large scale, the ramifications go far beyond individual deceit. Misinformation proliferates, public confidence erodes, and the integrity of democratic systems faces significant challenges. The issues this raises are as much ethical and legal as they are technological, Foremost among these concerns is the question of who is responsible when artificially generated content inflicts harm. As the distinction between

[1] Bertalan Mesko, *The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals*, 25 J Med Internet Res e48392 (2023).
[2] *OpenAI Closes Deal That Values Company at $300 Billion - The New York Times*, https://www.nytimes.com/2025/03/31/technology/openai-valuation-300-billion.html
[3] Antonio Pequeño IV, *Nvidia Sets New Record With Nearly $5.3 Trillion Value After AI Darling Surges 4%*, Forbes (Apr. 27, 2026), https://www.forbes.com/sites/antoniopequenoiv/2026/04/27/nvidia-sets-new-record-with-nearly-53-trillion-value-after-ai-darling-surges-4/.
[4] The BIRM Grp., *AI Infrastructure Construction: The Next $400B Boom in 2026*, The BIRM Group, https://thebirmgroup.com/ai-infrastructure-construction-the-next-400b-boom-in-2026/.

the genuine and the artificial becomes difficult to identify, these are matters that lawmakers, institutions, and society as a whole are just beginning to confront.[5]

1.1. Convergence of Gen AI and Deepfake : A Global Threat

The real-world implications of deepfakes are extensively documented and affect multiple sectors. In 2024, a finance employee at Arup's Hong Kong office was tricked into transferring US $25.6 million through a fake video conference,[6] while political figures utilized synthetic audio to sway Slovakia's 2023 election[7]. Women, celebrities, and public figures have been subjected to widespread non-consensual explicit imagery, Taylor Swift ranked first on McAfee's 2025 list of most-deepfaked celebrities, and in India alone, the number of deepfake incidents has increased by 550% since 2019.[8] GenAI has equipped nearly anyone with tools of remarkable destructive capability, and the legal system, institutions, and detection technologies are all struggling to keep pace.

The societal damage of GenAI and Deepfake technologies is extensive and profoundly personal. Women represent 99% of victims of non-consensual intimate imagery, while celebrities, politicians, and everyday individuals are susceptible to explicit deepfakes that can now be created in less than a minute at minimal expense. Targeted harassment, frequently infused with racial and gender-based hostility, has affected victims ranging from schoolgirls, to female journalists across the globe[9]. The production of AI-generated child sexual abuse material increased fourfold between 2023 and 2024, and a UNICEF report from February 2026 revealed that 1.2 million children had their images altered into sexualised deepfakes in 2025, a statistic that highlights the magnitude of harm to an almost unfathomable degree[10]. Studies also

[5] Enzo Maria Le Fevre Cervini & María Victoria Carro, *An Overview of the Impact of GenAI and Deepfakes on Global Electoral Processes*, Italian Inst. for Int'l Pol. Stud. (Oct. 25, 2024), https://www.ispionline.it/en/publication/an-overview-of-the-impact-of-genai-and-deepfakes-on-global-electoral-processes-167584

[6] Kathleen Magramo, Chris Lau & Joyce Jiang, *Finance Worker Pays Out $25 Million After Video Call With Deepfake 'Chief Financial Officer'*, CNN (Feb. 4, 2024), https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk.

[7] Pranshu Verma, *AI Voice Clones Mimic Politicians and Celebrities, Reshaping Reality*, Wash. Post (Oct. 13, 2023), https://www.washingtonpost.com/technology/2023/10/13/ai-voice-cloning-deepfakes/

[8] Brooke Seipel, *The World's Most Deepfaked Celebrities Revealed*, McAfee (Nov. 13, 2025), https://www.mcafee.com/blogs/internet-security/the-stars-scammers-love-most-mcafee-reveals-worlds-most-deepfaked-celebs/

[9] Felipe Romero-Moreno, *Deepfake Detection in Generative AI: A Legal Framework Proposal to Protect Human Rights*, 58 Comput. L. & Sec. Rev. 106162 (2025), https://doi.org/10.1016/j.clsr.2025.106162

[10] Deepfake Abuse Is Abuse, UNICEF (Feb. 4, 2026), https://www.unicef.org/press-releases/deepfake-abuse-is-abuse.

indicate that racial bias is ingrained in deepfake technology itself, exacerbating the dangers of harassment and misidentification for communities of colour.[11]

This paper argues the nature of GenAI as a double edged sword that makes deepfakes particularly challenging to manage. Instruments such as DALL-E and Stable Diffusion, which are designed to learn and replicate intricate patterns in image, audio, and video, are the same tools that allowed activists in HBO's Welcome to Chechnya to be digitally masked for their safety, but they have also been misused to carry out multi-million dollar fraud.[12] The technology itself is neither inherently good nor bad; it is simply extraordinarily powerful. Detection has evolved into a race of its own, with machine learning systems encompassing image, audio, video, and text analysis being utilized in a continuous effort to keep up with increasingly convincing synthetic media. However, a significant gap persists: although the technical literature on deepfake detection is extensive, the legal and regulatory frameworks governing the development and deployment of such technologies across multiple jurisdictions remain fragmented and insufficiently examined. This research seeks to address that gap through a feminist and governance-oriented legal analysis.[13]

1.2. Generative AI and Technology-Facilitated Gender-Based Violence

Technology-facilitated violence against women is not new, even prior to the introduction of GenAI, women were disproportionately subjected to online harassment, doxing, stalking, and the distribution of non-consensual imagery. The primary changes lie in the scale, speed, and the manner with which these abuses can now be perpetrated. AI has rendered such abuse more anonymous, more persuasive, and considerably more difficult to detect or address.[14]

This paper argues that disinformation continues to be one of the most widespread instruments of this violence: 65% of women in a particular study reported having experienced or observed abuse rooted in disinformation, and AI now allows for the creation and spread of false

[11] Atin Jindal, *Misguided Artificial Intelligence: How Racial Bias Is Built Into Clinical Models*, 2 Brown J. Hosp. Med. 38021 (2022), https://doi.org/10.56305/001c.38021.

[12] Rebecca Heilweil, *How Deepfakes Could Actually Do Some Good*, Vox (June 29, 2020), https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya

[13] Reza Babaei, Samuel Cheng, Rui Duan & Shangqing Zhao, *Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis*, 14 J. Sensor & Actuator Networks 17 (2025), https://doi.org/10.3390/jsan14010017

[14] Kalliopi Mingeirou, Yeliz Osman & Raphaëlle Rafin, *The Impact of Artificial Intelligence on Violence Against Women and Girls*, Stimson Ctr. (Feb. 26, 2026), https://www.stimson.org/2026/the-impact-of-artificial-intelligence-on-violence-against-women-and-girls/

narratives and fabricated content on an unprecedented scale.[15] There is particular concern regarding the surge in image-based abuse, it is estimated that 98% of all deepfake content available online is non-consensual and pornographic, with 99% of the victims being women, merging sexual violence, reputational damage, and enduring psychological trauma into a single act.[16]

The rapid proliferation of GenAI and deepfake technologies has intensified concerns surrounding misinformation, epistemic pollution[17], and technology-facilitated violence against women and girls. By enabling the mass creation and dissemination of synthetic yet highly convincing content, GenAI increasingly undermines public trust, distorts knowledge systems, and amplifies digital harms such as non-consensual explicit imagery, impersonation, sextortion, and targeted harassment.[18] These harms disproportionately affect women, exposing the gendered dimensions of AI-enabled abuse and revealing critical inadequacies within existing legal and regulatory frameworks.

This paper therefore examines the intersection between GenAI, deepfakes, and gendered digital violence, with particular emphasis on the ways in which contemporary legal systems struggle to address emerging forms of epistemic and technological harm.

2. Synthetic Media and the Architecture of Digital Harm

Synthetic media denotes digital content, including images, videos, audio, and text that are generated or modified by artificial intelligence instead of being created through conventional human methods. Driven by neural networks that have been trained on extensive datasets, it includes deepfakes, AI-generated avatars, voice cloning, text produced by large language models, and computer-generated imagery. Its uses extend across entertainment, advertising, education, and corporate communications, facilitating hyper-personalized content and immersive virtual experiences. The increasing sophistication and accessibility of GenAI tools have intensified concerns regarding the erosion of digital authenticity; particularly as such

[15] Alexander Romanishyn, Olena Malytska & Vitaliy Goncharuk, *AI-Driven Disinformation: Policy Recommendations for Democratic Resilience*, 8 Frontiers in Artificial Intelligence 1569115 (2025), https://doi.org/10.3389/frai.2025.1569115.

[16] Why Women Can't Get Protection from AI Deepfake Abuse, U.N. News (Mar. 21, 2026), https://news.un.org/en/story/2026/03/1167174

[17] Aurélien Acquier & Jozef Cossey, *Generative AI, Academic Deepfakes, and Epistemic Pollution*, 65 Bus. & Soc'y (2025), https://doi.org/10.1177/00076503251406457.

[18] Felipe Romero-Moreno, *Generative AI and Deepfakes: A Human Rights Approach to Tackling Harmful Content*, 38 Int'l Rev. L. Computers & Tech. 297 (2024), https://doi.org/10.1080/13600869.2024.2324540.

technologies are increasingly weaponized against vulnerable and historically marginalised groups[19]. However, the very features that render synthetic media commercially attractive also render it ethically problematic, especially regarding its ability to generate convincing misinformation on a large scale.[20]

2.1. Understanding DeepFakes and Generative AI

The term "deepfake" originated on Reddit in 2017, a portmanteau of "deep learning" and "fake", when a user initiated the practice of inserting celebrity faces into pornographic videos. What began as a niche and unsettling internet phenomenon transformed, once the foundational code was made publicly available, into a technology that became accessible to the general public.[21]

By 2025, the number of deepfakes online increased from approximately 500,000 in 2023 to around 8 million, a sixteen-fold increase within just two years.[22] Today, deepfakes operate at an ecosystem level: fabricated videos amplified by fake accounts, hosted on fringe websites, and weaponized through coordinated misinformation campaigns. In 2023, manipulated audio falsely portrayed London Mayor Sadiq Khan dismissing Armistice Day among extremist networks, while deepfakes of Keir Starmer depicting him mistreating staff circulated during the Labour Party conference.[23] In January 2024, explicit deepfakes of Taylor Swift garnered 47 million views on X before being removed, prompting calls for criminalization that led to the US TAKE IT DOWN Act of May 2025, the first federal legislation explicitly criminalizing non-consensual AI-generated intimate imagery.[24]

---

[19] Igor Calzada, Géza Németh & Mohammed Salah Al-Radhi, *Trustworthy AI for Whom? GenAI Detection Techniques of Trust Through Decentralized Web3 Ecosystems*, 9 Big Data & Cognitive Computing no. 3, 62 (2025), https://doi.org/10.3390/bdcc9030062

[20] D-ID. *What Is Synthetic Media? Benefits & Applications*. D-ID. https://www.d-id.com/resources/glossary/synthetic-media/

[21] Canadian Security Intelligence Service, *The Evolution of Disinformation: A Deepfake Future* (Oct. 1, 2023), https://www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future/the-evolution-of-disinformation-a-deepfake-future.html

[22] Siwei Lyu, *Deepfakes Leveled Up in 2025 — Here's What's Coming Next*, The Conversation (Dec. 29, 2025), https://theconversation.com/deepfakes-leveled-up-in-2025-heres-whats-coming-next-271391

[23] Dan Sabbagh, *Faked Audio of Sadiq Khan Dismissing Armistice Day Shared Among Far-Right Groups*, The Guardian (Nov. 10, 2023, 1:32 PM), https://www.theguardian.com/politics/2023/nov/10/faked-audio-sadiq-khan-armistice-day-shared-among-far-right

[24] Imran Rahman-Jones, *Taylor Swift Deepfakes Spark Calls in Congress for New Legislation*, BBC News (Jan. 27, 2024), https://www.bbc.com/news/technology-68110476

As evidenced in *Clarkson v. OpenAI*[25], synthetic media technologies have evolved from tools of reputational manipulation into mechanisms capable of enabling disinformation campaigns, coercive exploitation, and unauthorized informational intrusion, raising serious concerns regarding the erosion of epistemic trust and the destabilization of democratic communication structures.

Deepfake technologies typically use AI techniques like GANs or autoencoders and require significant computing power (e.g., GPUs) and data.

- **Gathering source and target data**: At first, Deepfake collects thousands of high-quality images or video frames of the target person (whose face/voice you want to replicate), a source video (the actor/body whose movements will be used) and aims for varied angles, expressions, lighting, and poses for better training.

- **Preprocessing the data**: Followed by cleaning and aligning the footage, extracting faces using tools like facial landmark detectors (e.g., DLib or MTCNN), resizing images to uniform dimensions (e.g., 256x256 pixels), and normalizing lighting/color to reduce inconsistencies.

- **Training the AI model**:

  For GANs: Setting up a generator (creates fake faces) and discriminator (spots fakes); training them adversarially on the dataset until the generator produces realistic outputs.

  For autoencoders: Encoding the source and targeting faces into compact latent representations, then training separate decoders to reconstruct one face onto the other.

- **Face swapping or synthesis**: Applying the trained model frame-by-frame: Mapping the target's facial features, expressions, and movements onto the source actor's body. Handling voice if needed: Using tools like WaveNet or Tortoise-TTS trained on audio samples to clone and sync the voice.

- **Refining and smoothening artifacts**: Post-processing is done with blending techniques (e.g., Poisson image editing), fixing lighting mismatches, removing glitches like unnatural

[25] IM Human, *Unveiling the Legal Battle: OpenAI Faces Lawsuit Over Data Collection Practices* (Aug. 29, 2023), https://www.imhuman.ai/blog/unveiling-the-legal-battle-openai-faces-lawsuit-over-data-collection-practices

blinking or edge blurring, and upscaling resolution for realism.

- **Finally, rendering and exporting**: The rendering and export phase entails assembling the processed frames into a unified video output, aligning the accompanying audio, and assessing the realism of the finished product. Numerous software applications, such as DeepFaceLab, Faceswap, and Roop, assist in and automate substantial parts of this procedure.[26]

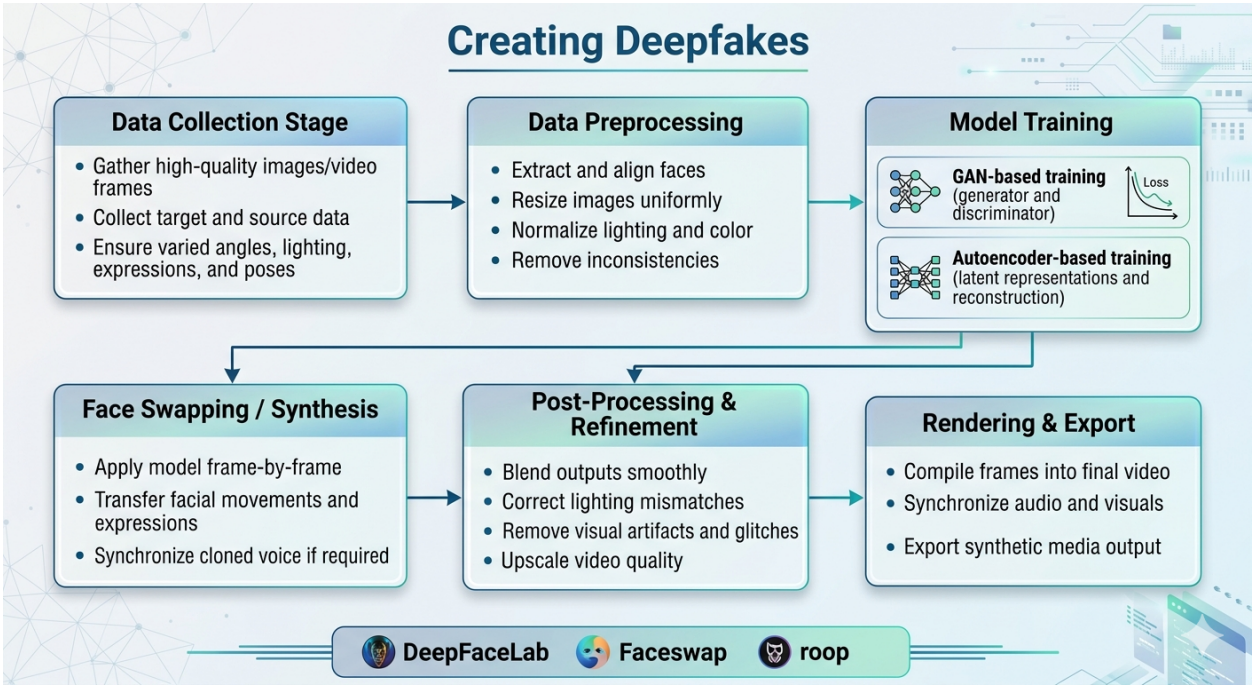

Figure 1: General Workflow for Synthetic Media/Deepfake Generation[27]

2.2. DeepFakes, Epistemic Pollution, and the Crisis of Authenticity

Beyond these immediate dangers lies a more profound and insidious crisis. Scholars have described the rise of deepfakes as an 'infopocalypse': a condition in which distinguishing between authentic content and fabrication becomes increasingly difficult, thereby undermining the evidentiary reliability of digital media.[28] In this context, deepfakes are viewed as epistemic pollutants within the digital information ecosystem thereby undermining collective ability to

[26] Florinel-Alin Croitoru et al., *Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook* (2024), arXiv, https://doi.org/10.48550/arXiv.2411.19537.
[27] Google Gemini, generative AI tool by Google, used March 2026 to assist in formatting and visualizing author-provided research content.
[28] Nina Schick, *Deep Fakes and the Infocalypse: What You Urgently Need to Know* (Monoray 2020).

discern reality.

The democratization of deceptive tools has hastened what some theorists describe as a "truth recession," where photographic and audiovisual evidence, once regarded as the gold standard for evidence, can no longer be relied upon.[29] Perhaps most significantly, deepfakes undermine what scholars term the 'epistemic backstop': the historical role of recordings as a foundation of social trust. It is this final aspect that renders deepfakes fundamentally different from previous forms of misinformation: not merely because falsehoods can be created, but because truth itself can be rendered deniable.[30]

The deepest epistemic threat posed by deepfakes is deceptively simple: they lead people to believe fabricated realities. A forged video, indistinguishable from an authentic one, serves as the foundation for belief, indignation, and subsequent actions.

A third dimension of this epistemic issue may be the most damaging of all. Deepfakes not only create false beliefs or weaken the justification for true ones, they can also hinder individuals from developing reliable beliefs entirely. As fake videos become increasingly prevalent, audiences tend to become broadly skeptical of all audiovisual content, including authentic footage. Consequently, legitimate journalism and genuine evidence may be dismissed not due to their falsehood, but because the overall information landscape has rendered trust epistemically irresponsible. As noted by Chesney and Citron (2019), this widespread distrust of credible media sources is, in itself, a primary goal of media manipulation, a tactic aimed at generating uncertainty rather than replacing one truth with another. In this regard, deepfakes do not simply disseminate misinformation; they undermine the essential conditions necessary for any type of information to be reliably assessed.[31]

The severity of this threat is primarily due to its accessibility. Tools like FakeApp and Zao have made it possible to create convincing deepfakes without the need for technical skills or specialized equipment, a smartphone and internet connection are often sufficient. As Johnson

[29] Adam Rose, *The Real Threat of AI Is the Collapse of Trust: Why Journalism Needs to Prove Which Images Are Authentic — Not Just Label Deepfakes*, Poynter (Feb. 2, 2026), https://www.poynter.org/commentary/2026/the-real-threat-of-ai-is-the-collapse-of-trust-deepfakes/

[30] Anne Hausknecht, *The Impact of Deepfakes on Trust in User-Generated Evidence*, in *Deepfakes and the Law: Challenges, Responses, and Critique* (Taylor & Francis 2025).

[31] Shiv Shankar Das, Mehul Agarwal, Sruthi Rajan & Anwesha Padhi, *Synthetic Realities: Youth Media Literacy and Trust in the Age of Digital Deception*, Information Dev. (OnlineFirst 2025), https://doi.org/10.1177/02666669251374658

points out, deception is not a novel concept; however, what is unprecedented is the widespread availability of deceptive tools, which poses a significant risk to the social fabric of trust that underpins public life.[32]

Although the most convincing deepfakes still necessitate considerable resources for their creation, the rapid pace of technological advancement is swiftly reducing that barrier. A related issue is the "liar's dividend," which refers to the diminishing trust in authentic media that deepfakes generate merely by their existence. Various responses are being implemented across platforms, with Facebook, TikTok, YouTube, and others establishing disclosure mandates, content prohibitions, and authenticity tagging systems, while news organizations have initiated training programs for fact-checkers. It is important to note, however, that expert fact-checkers caution that deepfakes, despite their seriousness, may pose a lesser immediate threat compared to "cheap fakes": decontextualized or reframed genuine content that is simpler to create and arguably more challenging to identify.[33]

Nevertheless, this same technology possesses legitimate and beneficial applications. It can recreate historical events that were never captured on film, enhancing the accuracy and engagement of educational content. It facilitates smoother dubbing across different languages, thereby expanding the reach of content to broader audiences. Additionally, it has been employed to safeguard the identities of whistleblowers and activists, enabling vulnerable individuals to voice their concerns without the fear of being exposed. Furthermore, it can produce realistic video lectures from a mere audio recording and a few images. Thus, the issue lies not within deepfake technology itself, but rather in the lack of effective governance regarding its application.[34]

The existence of legitimate uses does not diminish the scale of the threat. The advantages are tangible yet limited; the negative impacts are extensive and increasing. The aim here is not to determine if deepfakes result in an overall loss of knowledge, but to affirm that the epistemic danger they present is significant enough to require immediate focus.

---

[32] Amir Javadpour, Forough Ja'fari, Tarik Taleb, Mohammad Shojafar & Chafika Benzaïd, *A Comprehensive Survey on Cyber Deception Techniques to Improve Honeypot Performance*, 140 Computers & Security 103792 (2024), https://doi.org/10.1016/j.cose.2024.103792.

[33] Fatmaelzahraa Eltaher et al., *Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms* (2025), arXiv, https://arxiv.org/abs/2505.11160.

[34] Max Kalmykov, *Positive Applications for Deepfake Technology*, DataArt (Nov. 28, 2023), https://www.dataart.com/blog/positive-applications-for-deepfake-technology-by-max-kalmykov

3. Gendered Harm and Technology-Facilitated Violence Against Women

Technology-facilitated gender-based violence (TFGBV) encompasses any act that is committed, aided, or intensified through digital means, resulting in, or likely to result in, physical, sexual, psychological, social, political, or economic harm. Although TFGBV affects all women who use digital technologies, certain groups experience heightened vulnerability, particularly journalists, activists, academics, politicians, human rights defenders, and young women.[35]

The harms of TFGBV are further intensified by intersectionality. Women with disabilities, women of color, migrant women, LGBTIQ+ individuals, and those facing overlapping forms of discrimination experience disproportionate levels of online abuse shaped by race, class, sexuality, religion, age, and socioeconomic status.[36]

TFGBV manifests through a wide range of harmful behaviours, including sextortion, cyberstalking, doxxing, online harassment, impersonation, hate speech, and the non-consensual distribution of intimate content. Additionally, it acts as a catalyst for offline violence, which encompasses intimate partner violence and trafficking.

The most frequently reported types include misinformation and defamation (67%), cyber harassment (66%), hate speech (65%), and impersonation (63%).[37] The magnitude of the issue is considerable: as reported by the EU Agency for Fundamental Rights, one in ten women in the European Union has faced cyber-harassment since turning fifteen, which includes receiving unsolicited sexually explicit messages and unwelcome advances on social media platforms.[38]

[35] U.N. Women, *Accelerating Efforts to Tackle Online and Technology-Facilitated Violence Against Women and Girls* (2022), https://www.unwomen.org/sites/default/files/2022-10/Accelerating-efforts-to-tackle-online-and-technology-facilitated-violence-against-women-and-girls-en_0.pdf

[36] Shaffa Hameed, Babalwa Tyabashe-Phume, Eunice Tunggal, Xanthe Hunt, Lieketseng Ned & Karen Soldatić, *Technology-Facilitated Gender-Based Violence Against Women with Disabilities in Low- and Middle-Income Countries: A Scoping Review Protocol*, 15 BMJ Open e093988 (2025), https://doi.org/10.1136/bmjopen-2024-093988

[37] U.N. Regional Information Centre for Western Europe, *How Technology-Facilitated Gender-Based Violence Impacts Women and Girls* (Nov. 29, 2023), https://unric.org/en/how-technology-facilitated-gender-based-violence-impacts-women-and-girls

[38] European Parliament, *Cyberviolence Against Women: What Is It and How to Prevent It?* (Dec. 6, 2024), https://www.europarl.europa.eu/topics/en/article/20241205STO25880/cyberviolence-against-women-what-is-it-and-how-to-prevent-it

Region/Source	Statistic	Year
Global (OHCHR)	Fastest-growing form of violence against women/girls[39]	2026
US/NY (Provider survey)	97% of GBV cases involve tech abuse[40]	Recent
Australia	99.3% of GBV situations include TFA	2020
Philippines (NCR)	32% of 93+ TFGBV cases [41]	2025
Multi-country (Pan Int'l)	58% of girls experienced online harassment [42]	2020

Table 2: Global Statistics on Technology-Facilitated Gender-Based Violence (TFGBV)

3.1. Digital misogyny and the Consequence of TFGBV

Digital misogyny reinforces patriarchal structures through algorithms, platform cultures, and AI systems that amplify hostility toward women online. GenAI further intensifies these harms through deepfakes and non-consensual synthetic pornography. Studies indicate that women constitute the overwhelming majority of victims, while reports of online violence against women have risen sharply in recent years.[43] The 'manosphere', including incel, MGTOW, and pickup artist communities, normalizes anti-feminist ideologies and reinforces harmful masculinities across digital spaces.[44]

The consequences of TFGBV extend well beyond the immediate act of abuse. Survivors face elevated risks of depression, anxiety, PTSD, suicidal ideation, unintended pregnancies, and

[39] Office of the United Nations High Commissioner for Human Rights, *Technology Facilitated Gender-Based Violence* (2025), https://www.ohchr.org/sites/default/files/documents/issues/women/genderandequality/2025-tool-technology-facilitated-gbv.pdf

[41] Foundation for Media Alternatives, *Technology-Facilitated Gender-Based Violence (TFGBV) in the Philippines: Year-End Data Mapping Report* (Feb. 27, 2026), https://fma.ph/technology-facilitated-gender-based-violence-tfgbv-in-the-philippines-year-end-data-mapping-report/

[42] Ayodeji, Uzoma Maryjane. 2025. "Examining The Impact of Technology-Facilitated Gender-Based Violence on the Mental Health and Wellbeing of Adolescents". Current Journal of Applied Science and Technology 44 (5):66-77. https://doi.org/10.9734/cjast/2025/v44i54537.

[43] U.N. News, *Abuse of Women Journalists Made 'Easier and More Damaging' by AI* (Apr. 30, 2026), https://news.un.org/en/story/2026/04/1167416

[44] Selenia Anastasi, *Misogyny Beyond Borders: A Cross-Linguistic Corpus Assisted Analysis of Transnational Incel Communities* (Ph.D. dissertation, Univ. of Genova 2025), https://hdl.handle.net/20.500.14242/218817

sexually transmitted infections, effects that can persist across a lifetime. The psychological impact is particularly severe, with numerous survivors indicating that the breach of their digital safety directly correlates with declining mental and physical health.[45]

The broader societal implications are equally severe, as TFGBV silences women online, limits public participation, and reinforces patriarchal power structures. Consequently, TFGBV reinforces patriarchal systems and norms, acting not only as individual harm but also as a systemic obstacle to gender equality and the achievement of the Sustainable Development Goals.[46]

In response, reform efforts are gaining momentum: the EU's Digital Services Act and the Violence Against Women Directive are urging platforms to enhance their accountability, while UN Women,[47] UNFPA, and UNDP are promoting multi-stakeholder frameworks focused on digital literacy, safety-by-design, and support for survivors. At the heart of these reform agendas is the call for AI governance that tackles the unique risks associated with TFGBV, such as the reproduction of bias, gendered harm in automated systems, and the necessity for moderation practices that are authentically gender-sensitive.[48]

3.2. Epistemic injustice, in the context of technology-facilitated gender-based violence (TFGBV)

Fricker's concept of epistemic injustice, particularly in relation to technology-facilitated gender-based violence (TFGBV), arises when digital tools and platforms are employed to silence, discredit, or diminish the knowledge and experiences of marginalized individuals. This form of injustice often involves the dismissal or discrediting of survivors' experiences of online abuse, reinforcing existing power inequalities through digital platforms.[49]

---

[45] Caroline Stein et al., *The Health Effects Associated with Physical, Sexual and Psychological Gender-Based Violence Against Men and Women: A Burden of Proof Study*, 9 Nature Hum. Behav. 1201 (2025), https://doi.org/10.1038/s41562-025-02144-2

[46] United Nations Population Fund, *Technology-Facilitated Gender-Based Violence*, https://www.unfpa.org/TFGBV.

[47] U.N. Women, *16 Days of Activism 2025: End Digital Violence Against All Women and Girls*, https://www.unwomen.org/en/what-we-do/ending-violence-against-women/unite/theme

[48] United Nations Population Fund, *Creating a Safer Digital Future Free from Gender-Based Violence*, https://www.unfpa.org/updates/creating-safer-digital-future-free-gender-based-violence

[49] *Epistemic Injustice in the Digital Age: Social Media, Silencing, and the Politics of Credibility*, 7 J. Hum. & Educ. Dev. 18 (2025), https://doi.org/10.22161/jhed.7.3.4

Key Elements of Epistemic Injustice and TFGBV

- *Testimonial Injustice in Digital Spaces:* Individuals subjected to online harassment, especially women and members of minority groups, frequently find their allegations of abuse dismissed or downplayed due to bias. For instance, discussions on Twitter (now X) have revealed that black women endure significant levels of abuse, yet their voices are suppressed through organized efforts, undermining their status as credible "knowers".

- *Hermeneutical Injustice*: This occurs when marginalized communities lack the collective tools or language necessary to interpret or articulate their experiences of online violence. The rapid evolution of AI-generated abuse means many victims initially lack the language or conceptual tools needed to identify and report their experiences.[50]

- *AI as a Mechanism of Epistemic Injustice*: Generative AI is increasingly employed to produce deepfakes and non-consensual material, which directly undermines the victim's credibility and social reputation. These systems are often trained on biased datasets, resulting in the perpetuation of existing stereotypes and further marginalization of certain groups.[51]

- *Anonymity and Misinformation:* The anonymity afforded by online platforms exacerbates the challenges of epistemic injustice, as it enables offenders to conceal their identities or misrepresent themselves to evade responsibility or disseminate harmful misinformation.

Marginalized communities experience heightened epistemic injustice and the effects of TFGBV, where overlapping biases such as race, disability, and sexuality undermine their credibility and access.

- *Expanded Impacts Gender-Based Violence:* TFGBV, including cyberstalking and deepfakes, disproportionately affects women, with 38% impacted globally, leading to an escalation of offline physical harm.

[50] Arianna Falbo, *Hermeneutical Injustice*, in *The Blackwell Companion to Epistemology* (Jonathan Dancy, Ernest Sosa, Matthias Steup & Kurt Sylvan eds., 3d ed. forthcoming).
[51] Ana María Marín-López & María Isabel Pérez-Ramos, *The Gendered Dynamics of Trust and Artificial Intelligence: Implications for Human–AI Interaction*, 4 Frontiers in Human Dynamics 1790324 (2026), https://doi.org/10.3389/fhumd.2026.1790324

- *Disability and Technology*: Women with disabilities encounter risks that are 2-3 times greater; the neglect of accessibility in design results in epistemic dismissal, exacerbating their isolation and dependence.

- *Digital Exclusion:* A significant 76% of individuals are driven to self-censor or disconnect due to fear, which disrupts their information and support networks as well as their civic engagement.

- *Intersectional Mental Health*: Minority groups, including Black, Indigenous, and LGBTQ+ individuals, experience PTSD and shame stemming from testimonial silencing, while algorithms exacerbate these issues through biased profiling.

- *Economic/Social Isolation*: Victims often withdraw from society, resulting in job and educational losses; disabled and rural populations encounter severe obstacles due to the lack of inclusive reporting.[52]

These Technology-Facilitated Gender-Based Violence (TFGBV) harms are not isolated acts of individual misconduct, they emerge from broader systems of technological and structural power. The harms associated with technology-facilitated violence against women cannot be understood merely as isolated incidents of online abuse or individual misconduct. Rather, they reflect broader structural inequalities embedded within digital ecosystems, platform economies, and AI systems themselves. The increasing reliance on algorithmic systems trained on historically biased datasets risks reproducing existing hierarchies of gender, race, and social exclusion, while simultaneously amplifying both the visibility and scale of such harms. Consequently, synthetic violence must be understood not only as a technological issue, but as a manifestation of deeper asymmetries of power, access, and representation within contemporary digital environments.

4. Synthetic Violence, Power, and Structural Inequality

Synthetic violence refers to the use of AI-generated or manipulated content to produce,

[52] Chan Ho Park, Zhefan Rao, Liya Ji & Qifeng Chen, *(Re)mediators of Epistemic Injustice: Generative AI and Hermeneutic Resource Provision in Intimate Partner Violence*, in *Companion Proceedings of the 2024 ACM Conf. on Computer-Supported Cooperative Work and Social Computing* (CSCW Companion '24) 1 (2024), https://doi.org/10.1145/3772318.3791548

amplify, or simulate forms of digital harm and violence.[53]

The same synthetic media tools utilized for public safety also possess a highly effective destructive potential: disinformation. Generative AI is increasingly used to create deepfakes and non-consensual synthetic imagery, including the impersonation of public figures and the fabrication of explicit content. This synthetic abuse has particularly become a means of intimidation for women in public life, aimed at driving them out of both digital and democratic arenas. [54]

On a larger scale, synthetic media threatens the very fabric of social cohesion. AI-generated videos and images can be weaponized to exacerbate ethnic tensions, provoke communal violence, and create the illusion of atrocities that never took place. In environments prone to conflict, this is not merely a theoretical concern; it is an emerging operational reality, representing a modern and particularly insidious means through which synthetic content can act as a catalyst for mass violence.[55]

4.1. Algorithmic Bias and the Reproduction of Inequality

Algorithmic bias occurs when automated systems, which are trained on flawed or historically biased data, generate systematic errors that lead to unjust outcomes, consistently putting certain groups at a disadvantage. Rather than functioning as neutral systems, AI technologies often reproduce existing inequalities in areas such as healthcare, employment, lending, and criminal justice.

Sociological theories such as intersectionality[56], Critical Race Theory[57], and social stratification provide crucial insights into the ways AI systems engage with and often reinforce prevailing social hierarchies. Instead of functioning as unbiased mediators, algorithmic systems

[53] SIpEIA, *Book of Abstracts: Conference 2026* (Sapienza Università di Roma, Feb. 2–3, 2026), https://assets.super.so/169fae3b-be63-4d10-9489-e22abce56def/files/457c3b9f-a016-477b-b0b6-b2e8b0ec1ecb/sipeia2026_book_of_abstracts_new.pdf

[54] Ctr. for News, Tech. & Innovation, *Synthetic Media & Deepfakes*, https://cnti.org/issue-primers/synthetic-media-deepfakes/

[55] Matteo E. Bonfanti, *The Weaponisation of Synthetic Media: What Threat Does This Pose to National Security?*, Elcano Royal Inst. (July 14, 2020), https://www.realinstitutoelcano.org/en/analyses/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security/

[56] Kimberlé Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, 1989 U. Chi. Legal F. 139.

[57] Adam Augustyn, *Critical Race Theory*, Encyclopaedia Britannica, https://www.britannica.com/topic/critical-race-theory

tend to incorporate and exacerbate structural inequalities. Buolamwini and Gebru (2018) illustrated that commercial facial recognition technologies exhibited significantly elevated error rates for women with darker skin tones; the COMPAS recidivism assessment tool produced higher false-positive rates for Black defendants; healthcare algorithms underestimated the needs of Black patients by relying on cost as a surrogate for health conditions; and mortgage pricing mechanisms continued to disadvantage Black and Latinx borrowers, even with standardized lending practices.[58] In the realms of criminal justice, healthcare, employment, and credit, the evidence remains consistent: the notion of algorithmic objectivity is predominantly a facade, and without intentional measures, AI is at risk of perpetuating existing inequalities while masquerading as neutral.[59]

Algorithmic bias directed against women, referred to as algorithmic misogyny, takes place when artificial intelligence systems yield outcomes that consistently discriminate against women, thereby reproducing or exacerbating existing gender disparities. Such biases, which originate from biased historical training datasets and the insufficient representation of women in technology design, adversely impact recruitment processes, healthcare services, credit lending practices, and media portrayals.[60]

AI systems are only as objective as the data on which they are trained. Biases enter AI systems at multiple stages. These include unrepresentative datasets, historically biased data, human annotation practices, and measurement proxies that disguise inequality as neutrality. The repercussions are well documented. ChatGPT consistently links leadership with masculine characteristics, associates professional roles with male identifiers, and produces recommendation letters for women that highlight social attributes over professional ones.[61] Microsoft's chatbot confidently identified men as more intelligent when questioned directly. Predictive policing tools utilize arrest rates, which are themselves a result of racial bias, as indicators of safety. In each instance, bias is not incidental but structural, cyclical, and self-

[58] MIT News, *Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems* (Feb. 12, 2018), https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

[59] Jha, R. (2024). Algorithmic Bias and Social Inequality in AI Decision-Making Systems from a Sociological Perspective. ShodhKosh: Journal of Visual and Performing Arts, 5(6), 3151–3156. doi: 10.29121/shodhkosh.v5.i6.2024.6167

[60] U.N. Women, *How AI Reinforces Gender Bias—and What We Can Do About It* (Feb. 5, 2025), https://www.unwomen.org/en/news-stories/interview/2025/02/how-ai-reinforces-gender-bias-and-what-we-can-do-about-it

[61] Maya Indira Ganesh, *The Limits of Machine Learning in Automating Racialized Surveillance*, 12 Soc. Sci. 435 (2023), https://doi.org/10.3390/socsci12080435.

reinforcing.[62]

Algorithmic bias originates not only from training data, but also from the design and optimization logic of digital platforms. For example, social media algorithms are designed to maximize engagement, often rewarding outrage and antagonistic content over meaningful interaction. The gendered implications of this phenomenon are particularly significant.[63]

Mauro and Schellmann illustrated that AI classifiers identified images of women in bras as 'racy' while leaving similar images of shirtless men unaddressed, a seemingly technical differentiation that embeds profoundly sexist assumptions into automated content moderation on a large scale. Consequently, these systems frequently suppress and over-sexualize women's content while imposing no equivalent standard on men.[64]

Key Impacts of Algorithmic Bias:

- *Biased Datasets*: Training data that is skewed due to historical underrepresentation (for instance, fewer dark-skinned women) results in subpar model performance, embedding systemic racism and sexism into the predictions.

- *Racial/Gender Bias*: Darker-skinned females are misidentified at an error rate of 34.7%, compared to just 0.8% for light-skinned males, which contributes to wrongful arrests and disparities in surveillance.[65]

- *Discriminatory Outputs:* Algorithms tend to favor majority groups, resulting in the denial of loans and jobs to minorities or the generation of stereotypical content that reinforces objectification.[66]

[62] Alessandra Giannini, *Criminal Behavior and Accountability of Artificial Intelligence Systems* (Doctoral thesis, Maastricht Univ. & Univ. of Florence 2023), https://doi.org/10.26481/dis.20231124ag.
[63] William J. Brady, Joshua Conrad Jackson, Björn Lindström & M.J. Crockett, *Algorithm-Mediated Social Learning in Online Social Networks*, 27 Trends Cognitive Sci. 947 (2023), https://doi.org/10.1016/j.tics.2023.06.008
[64] Jerlyn Q.H. Ho, Andree Hartanto, Andrew Koh & Nadyanna M. Majeed, *Gender Biases Within Artificial Intelligence and ChatGPT: Evidence, Sources of Biases and Solutions*, 4 Computers in Hum. Behav.: Artificial Humans 100145 (2025), https://doi.org/10.1016/j.chbah.2025.100145
[65] David Leslie, *Understanding Bias in Facial Recognition Technologies: An Explainer*, Alan Turing Inst. (2020), https://doi.org/10.5281/zenodo.4050457
[66] Digital Divide Data, *Bias in Facial Recognition Systems for Computer Vision*, https://www.digitaldividedata.com/blog/bias-in-facial-recognition-systems-for-computer-vision

- *Over/Underrepresentation:* White and male faces dominate datasets (for example, over 75%), resulting in poorer recognition outcomes for underrepresented groups.

- *Facial Recognition Bias:* When deployed in policing, it mislabels minorities as threats (for example, studies from 2025 indicate 10 times higher false positives for Black faces), which erodes trust and infringes on rights.[67]

- *Bias in Moderation Systems*: Platforms tend to over-censor the speech of Black and queer activists as "hate" while amplifying white and male extremism, thereby silencing dissent and exacerbating epistemic injustice.[68]

These dynamics are directly related to TFGBV and digital misogyny in your research, as AI perpetuates patriarchal and racial hierarchies, necessitating debiasing through diverse data and audits.

4.2. Platform Economies, Visibility, and Digital Power

Platform economies, including digital intermediaries like Uber, Amazon, and social media platforms, facilitate large-scale connections among users while leveraging network effects to extract value, thereby exerting significant control over the visibility of content online. Algorithms play a crucial role in deciding which content is highlighted and which is obscured, promoting viral or majority opinions while systematically diminishing others. Consequently, the visibility landscape is anything but neutral: minority communities, marginalized voices, and non-dominant viewpoints are inherently disadvantaged by the very systems that claim to merely reflect user preferences.[69]

Platform economies represent intricate ecosystems comprising algorithms, datasets, APIs, and business models that create value by means of data collection, network effects, and multi-sided markets. Frequently viewed through the perspective of 'platform capitalism',[70] in which data is

[67] Lee J. Curley, Emily Breese, James Munro, Catriona Havard, Faye Skelton & Graham Pike, *The Effects of Contextual Bias on Face Recognition Decisions*, J. Forensic Sci. (2025), https://doi.org/10.1111/1556-4029.70177.
[68] Rutambhara Nayak & Anindya Sircar, *Algorithmic Content Moderation, IP and Expression: Navigating the Tension Through Theories*, 3 NLUA J. Intell. Prop. Rts. 1 (2024).
[69] Biqi Li & Yanyuan Cheng, *Mapping the Research Landscape of Algorithmic Control on Workers: A Bibliometric Analysis*, 12 Heliyon e44403 (2026), https://doi.org/10.1016/j.heliyon.2025.e44403
[70] H. V. Telnova & T. V. Reshetnyak, *Theoretical and Methodological Foundations of Modern Industry Market Analytics*, 2 Efektyvna Ekonomika 43 (2026), http://doi.org/10.32702/2307-2105.2026.2.43.

treated as a commodity or a source of monopoly rent, value is generated through a variety of mechanisms that can be both diverse and at times contradictory, ranging from centralized enclosure to decentralized contribution systems. A constant factor is the concentration of power regarding visibility, access, and participation, which prompts essential inquiries concerning accountability and governance within the digital economy.

In the context of the economy's "platformization," several key archetypes emerge:

- *Social media platforms* (such as X, Instagram, and TikTok) connect content creators, advertisers, and audiences, monetizing attention and data through targeted advertisements and recommendation algorithms.

- *E-commerce and marketplace platforms* (including Amazon, Flipkart, and Alibaba) facilitate connections between buyers and sellers, managing search rankings, logistics, and reviews, while often competing against their own sellers.

- *Cloud computing platforms* (like AWS, Azure, and Google Cloud) deliver on-demand computing and storage, along with platform services such as databases and machine learning tools, which other companies utilize, positioning them as infrastructure-level entities.

- *Fintech platforms* (for instance, PayPal, Stripe, UPI applications, and neobanks) serve as intermediaries for payments, credit scoring, and financial services, increasingly utilizing APIs that other applications incorporate.

- *AI platforms* (including OpenAI, Anthropic, Google, Meta, and various MLOps platforms) provide models, APIs, and tools that enable others to create AI-enabled products, thereby centralizing control over model weights, data, and safety regulations.[71]

Platformisation refers to the dissemination of platform logics and architectures into areas that were previously managed by traditional markets and companies, encompassing mobility and housing, as well as healthcare, education, and government services. These platforms are increasingly governed by AI: recommender systems curate the content users encounter,

[71] Inge Graef & Friso Bostoen, *A Typology of Platform Power and Its Regulation*, 29 Info., Comm. & Soc'y 324 (2026), https://doi.org/10.1080/1369118X.2025.2512972

algorithms establish pricing and allocate gig workers, and automated tools manage fraud detection, risk assessment, and content moderation. The most notable change, however, is the rise of agentic AI interactions, where semi-autonomous software agents negotiate, route, and compose on behalf of users, with human oversight rather than direct involvement. In this context, platformisation signifies a fundamental shift from markets and firms to AI-governed data infrastructures that dictate the conditions under which individuals, businesses, and institutions can engage in the economy.

Platform economies transform digital infrastructures into privately governed ecosystems in which data extraction, algorithms, and recommendation systems mediate visibility, participation, and economic value. Through algorithmic visibility and ranking mechanisms, platforms determine whose voices are amplified and whose are marginalized, often restricting the reach of women, activists, and racialized communities while simultaneously promoting sensationalist, polarizing, or misogynistic content for engagement. These systems are further reinforced through opaque moderation policies, terms of service, and platform governance structures that regulate speech, safety, and participation with limited democratic oversight.[72]

The dependence of creators, small businesses, advocacy groups, and marginalized users on platform visibility for economic sustainability and public engagement further intensifies these power asymmetries, reproducing existing social hierarchies and structural inequalities within digital spaces.

5. The Limits of Contemporary AI Governance

Global AI governance remains fragmented and inconsistent. While the European Union has adopted a binding horizontal framework, countries such as the United States, United Kingdom, China, and India continue to rely on varied and often sector-specific approaches.[73]

- **EU AI Act**

The EU AI Act, which came into effect in August 2024 and will be fully enforceable from August 2026, represents the first comprehensive and binding cross-sector AI legislation. It bans

[72] Antonio A. Casilli & Paola Tubaro, *The Organization of Algorithmic Management: A Comparative Analysis of Digital Labour Platforms*, 45 Info. & Org. 100612 (2026), https://doi.org/10.1016/j.infoandorg.2026.100612
[73] European Commission, *AI Act*, Shaping Europe's Digital Future, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

systems deemed to pose unacceptable risks, such as social scoring, manipulative AI, extensive facial recognition databases, and real-time biometric identification in public areas. Additionally, it imposes stringent requirements on high-risk systems related to data governance, human oversight, cybersecurity, and Fundamental Rights Impact Assessments. Although the Act positions the EU as a global leader in AI regulation, its effectiveness remains limited by the borderless nature of digital technologies.[74]

- **United States: Fragmented, Sectoral Approach**

The United States does not have a unified federal law governing AI. Instead, governance is decentralized and tailored to specific sectors, relying on pre-existing frameworks related to privacy, discrimination, and consumer protection. This is further enhanced by executive orders, guidance from agencies, and voluntary commitments from the industry. Various states and cities have enacted their own regulations, addressing issues such as biometric data, automated hiring processes, and, most recently, non-consensual deepfakes through the TAKE IT DOWN Act (2025).[75]

This has resulted in a fragmented system that leaves considerable gaps, especially concerning cross-border issues like TFGBV. Although this approach maintains flexibility and emphasizes innovation, it falls short in providing the consistent rights protections or clearly defined responsibilities that the European Union's more organized framework offers.

- **UK: Principles-Based 'Pro-Innovation' Model**

The United Kingdom currently relies on a non-legally binding, conventions-oriented strategy for AI governance, assigning current regulators, including the ICO, CMA, FCA, and others, the responsibility of implementing five overarching principles: *safety, transparency, fairness, accountability, and contestability.* The framework remains largely non-binding, creating uncertainty regarding enforceable rights and remedies for victims of AI-related harms. Although the framework imposes strict labeling and traceability obligations, it remains heavily

[74] European Parliament, *EU AI Act: First Regulation on Artificial Intelligence* (June 8, 2023), https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[75] Tatevik Davtyan, *The U.S. Approach to AI Regulation: Federal Laws, Policies, and Strategies Explained* (Sept. 9, 2024), SSRN, https://doi.org/10.2139/ssrn.4954290.

state-centric and prioritizes political control over individual rights.[76]

- **China: Deepfake-Focused "Deep Synthesis" Regulation**

China's 2023 deep synthesis regulations represent one of the first comprehensive laws globally that specifically address deepfakes. These regulations mandate that providers label synthetic content, verify user identities, secure consent, and prevent the use of such content for disinformation or any material considered a threat to economic or national security.[77]

Although the framework is relatively stringent regarding traceability and labeling, it is primarily state-centric, emphasizing information control and political security at the expense of individual rights. Broad definitions of terms like "fake news" and "national security" provide authorities with considerable interpretive flexibility, complicating alignment with human rights-oriented or globally compatible governance standards.

- **India: Regulatory Uncertainty**

India currently lacks a specific law governing artificial intelligence, depending instead on the Digital Personal Data Protection Act 2023 and the regulations of the IT Act, which were not formulated to address the harms caused by AI. Liability for harms such as deepfakes and algorithmic discrimination remains unclear, leaving victims without effective remedies and developers without defined obligations. This situation highlights the pressing necessity for a comprehensive national framework for AI governance.[78]

- **Key Argument: No Coherent Global Framework**

Collectively, these frameworks create a highly fragmented governance environment, comprising a rights-based, risk-tiered model in the EU; a sectoral and predominantly voluntary approach in the US; a principles-based framework in the UK; a security-focused regime for deepfakes in China; and an emerging response from India. Cross-border harms such as

[76] Dep't for Sci., Innovation & Tech., *Implementing the UK's AI Regulatory Principles: Initial Guidance for Regulators* (Feb. 6, 2024), https://www.gov.uk/government/publications/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators

[77] Xiangxiang Ma, *Deep Synthesis Not Deepfake: How AI Compliance Works in China*, China L. Vision (Feb. 27, 2025),https://www.chinalawvision.com/2025/02/digital-economy-ai/deep-synthesis-not-deepfake-how-ai-compliance-works-in-china/

[78] Om Prakash Rai, *Artificial Intelligence, Data Protection and Digital Governance in India: A Contemporary Legal Analysis*, 5 J. Soc. Rev. & Dev. 76 (2026), https://doi.org/10.64171/JSRD.5.1.76-81

deepfake-enabled technology-facilitated gender-based violence (TFGBV), digital misogyny, and algorithmic discrimination consistently slip through the cracks of these systems, as definitions, responsibilities, and enforcement mechanisms vary significantly across different jurisdictions.[79]

This paper contends that the current state of AI and platform governance resembles a series of regulatory islands rather than a cohesive global framework, resulting in victims of transnational, technology-enabled abuse lacking consistent protection and meaningful recourse.

5.1. Platform Liability and Moderation Failures

Platforms such as X/Twitter, Meta, and Reddit function not merely as hosts of user content, but as algorithmic amplifiers that shape visibility and accelerate the spread of harmful material.[80]

Platform recommendation algorithms are designed to maximize engagement, which, as evidence consistently indicates, is fueled by outrage, hatred, and sensationalism. Consequently, harmful content such as doxxing, deepfake leaks, and coordinated harassment is not only tolerated on these platforms; it is systematically amplified, disseminated into feeds and trends well beyond its initial audience. The resulting harm is not merely incidental to the system; it is embedded within its engagement-driven architecture.[81]

Algorithms increasingly prioritize content based on real-time engagement signals such as shares, comments, and viewing duration. The emergence of "*For You*" feeds exemplifies this change: instead of presenting users with content from accounts they opt to follow, these systems emphasize whatever captures attention, consistently promoting sensational, polarizing, and emotionally charged content as it enhances viewing duration and, consequently, platform earnings.[82]

---

[79] Nithin Monteiro SJ & Vaishali Singh, *The Wheel of Artificial Intelligence Governance*, Sustainable Futures 101279 (2025), https://doi.org/10.1016/j.sftr.2025.101279

[80] Raffael Heiss & Isabelle Freiling, *Addressing Social Media Platforms' Influence on Academic Research*, 13 Humanit. & Soc. Sci. Commc'ns 192 (2026), https://doi.org/10.1057/s41599-026-06690-6.

[81] Marianna Spring & Mike Radford, *Meta and TikTok Let Harmful Content Rise After Evidence Outrage Drove Engagement, Say Whistleblowers*, BBC News (Mar. 16, 2026), https://www.bbc.com/news/articles/cqj9kgxqjwjo.

[82] Niels Van Doorn & Koen Leurs, *Reaction Videos as "Interpassive" Aesthetic Experiences: Understanding the Meaning of Diffuse Media Practices in the Platform Society*, Comm. Rev. (2026),

Delayed Takedowns and Moderation Failures

Research utilizing the transparency data from the EU Digital Services Act highlights a significant conclusion: the promptness of content removal is of utmost importance. Takedowns executed within hours significantly diminish the risk of harm, whereas delays extending into weeks render any intervention nearly pointless. Nevertheless, X/Twitter, Meta, and Reddit have all encountered ongoing criticism, both in tribunal hearings and civil society reports, for their sluggish responses to abuse, inconsistent enforcement of rules, and chronically underfunded trust and safety teams, especially in the wake of cost-cutting measures and policy reversals.[83]

Moderation systems themselves remain deeply flawed. Automated tools frequently fail to detect coded harassment while disproportionately censoring activists, feminists, and minority voices. The lack of transparency in moderation rules leads to outcomes that are both revealing and concerning: political and human rights content is removed, while misogynistic and extremist material persists until it is reported en masse.[84]

On Reddit, even after specific posts or subreddits face sanctions, recommendation algorithms continue to promote toxic content, indicating that the issue lies not just with individual pieces of content but with the amplification logic itself, which current moderation practices are unable to address.[85]

Traditional safe harbour frameworks were developed on the assumption that platforms function as passive intermediaries rather than active publishers. This assumption is no longer sustainable. Platforms that curate, rank, and algorithmically enhance content wield significant editorial influence, and critics contend that they should assume commensurate responsibility for foreseeable damages. The EU's Digital Services Act embodies this transition, shifting from complete immunity to due-diligence requirements, mandating platforms to evaluate systemic risks, invest in moderation efforts, and uphold transparent, prompt takedown procedures.[86]

https://doi.org/10.1080/10714421.2026.2647307.

[83] European Commission, *The Digital Services Act (DSA): Impact on Platforms*, Shaping Europe's Digital Future, https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms

[84] Manoj, *State of AI Content Moderation 2026*, Foiwe (Feb. 20, 2026), https://www.foiwe.com/state-of-ai-content-moderation-2026/

[85] Jennifer Cobbe & Jatinder Singh, *Regulating Recommending: Motivations, Considerations, and Principles*, 10 Eur. J.L. & Tech. no. 3 (2019), https://ejlt.org/index.php/ejlt/article/view/686/979

[86] Yukta Chordia & Kanika Chhajerh, *From Shield to Sword: How Safe Harbour Became the State's Tool of Platform Control*, RSRR (Mar. 6, 2026), https://www.rsrr.in/post/from-shield-to-sword-how-safe-harbour-became-the-state-s-tool-of-platform-control

Even with the invocation of safe harbour protections, platforms like X, Meta, and Reddit do not merely serve as passive hosts for user-generated speech. By employing algorithmic ranking, recommendations, and delayed moderation, they play an active role in shaping, amplifying, and normalizing harmful content. Empirical studies regarding takedown delays and automated moderation clearly demonstrate that the design choices made by these platforms have a direct impact on the extent and reach of online abuse, revealing the illusion of intermediary neutrality for what it truly is.[87]

5.2. Toward Gender-Responsive AI Governance

Gender responsive AI governance should move from abstract principles to concrete, enforceable duties that prevent harm, center survivors, and redistribute digital power instead of merely "managing risk."[88]

Survivor-Centered Frameworks

- Incorporate the perspectives and lived experiences of survivors into the design of AI policies, impact assessments, and redress mechanisms, which should include specific reporting channels and support services for TFGBV.

- Mandate that platforms and AI providers deliver confidential, trauma-informed reporting options, ensure the swift removal of abusive content (such as deepfakes), and create avenues for legal and psychosocial assistance.[89]

Gender-Sensitive Moderation

- Transition from prioritizing engagement-maximizing algorithms to a moderation approach that specifically addresses TFGBV, misogyny, and intersectional harassment, employing gender-sensitive policies and teams educated in feminist and human rights principles.

[87] Luca Ettore Perriello, *Blurred Realities: Legal Strategies for the Deepfake Era*, Int'l Rev. L., Computers & Tech. (OnlineFirst 2026), https://doi.org/10.1177/1023263X261433380.
[88] United Nations Population Fund & Gender in Digital Coalition, *Concept Note: AI for Good: Gender Inclusion in the AI Ecosystem* (Virtual Consultation, Apr. 23, 2026).
[89] United Nations Development Programme Pakistan, *Enabling Safe Digital Spaces Through a Survivor-Centred Response to Technology-Facilitated Gender-Based Violence in Pakistan*, https://www.undp.org/pakistan/projects/enabling-safe-digital-spaces-through-survivor-centred-response-technology-facilitated-gender-based-violence-pakistan

- Integrate enhanced machine detection capabilities with adequately supported human moderators and community-driven reporting mechanisms, guaranteeing that activists and marginalized voices are not unduly suppressed.

Safety-by-Design

- Adopt a "Prevention by design" approach: modify recommendations, reporting mechanisms, and privacy settings to mitigate TFGBV at its origin, instead of depending exclusively on post-event removal.

- Require risk assessments for new functionalities (such as generative tools and changes in recommendations), ensuring a thorough evaluation of TFGBV and gender-specific harms prior to implementation.[90]

Transparency and Accountability

- Mandate substantial AI transparency: comprehensive public documentation of models, categories of training data, and the logic behind content ranking, in addition to gender-disaggregated harm reporting and stress testing.

- Create enforceable accountability measures (including audits, penalties, corrective actions, and the suspension of high-risk systems) when organizations do not adequately address foreseeable gender-related harms.[91]

Digital Literacy and Empowerment

- Allocate resources towards gender-sensitive digital literacy initiatives: educate users, particularly girls, LGBTQ+ individuals, and marginalized groups, on the functioning of AI systems, the manifestations of TFGBV, and the avenues for seeking assistance and justice.

- Equip users with tools to manage their data, visibility, and interactions (for instance, safer default settings, detailed blocking options, shadow-banning controls, and consent

[90] Sexual Violence Research Initiative, *TFGBV Safety by Design*, https://www.svri.org/tfgbv-safety-by-design/
[91] Jelena Cupac, *The Gender Code: Gendering the Global Governance of Artificial Intelligence* (2025), arXiv, https://doi.org/10.48550/arXiv.2512.09570.

management dashboards).[92]

International Collaboration

- Harmonize national AI and platform regulations with human rights-oriented, feminist principles to ensure that cross-border issues such as TFGBV are addressed uniformly across different jurisdictions.

- Utilize global AI forums and UN initiatives to establish shared standards on TFGBV, which should encompass minimum responsibilities for platforms and model providers regarding consent, labeling, and redress.

Stronger Consent and Data Frameworks

- Implement stringent and detailed consent protocols for the utilization of biometric data and personal images during both training and deployment phases, with explicit bans on non-consensual sexualized materials and deepfake pornography.[93]

- Guarantee that survivors have the right to request the removal of non-consensual content and limit the reuse of their data, supported by legal recourse and obligations for cross-platform content removal.[94]

Intersectional Governance

- Mandate intersectional analysis as a requirement in AI governance, policies must be assessed for their impact on women who exist at the intersections of race, class, caste, disability, sexuality, and migration status.

- Empower regulators to postpone or prevent deployment in cases where intersectional inequities remain, even if it compromises model performance, acknowledging that safety

[92] Integrity Inst., *Prevention by Design: A Roadmap for Tackling TFGBV at the Source*, https://www.integrityinstitute.org/research/pwojclz543oq3n3n9uk7ro644y0kcm

[93] Women in Digital, *Embedding Gender Equality in the EU's Digital Future: From AI Bias to Actionable Policy Solutions* (Mar. 7, 2025), https://widigital.eu/embedding-gender-equality-in-the-eus-digital-future-from-ai-bias-to-actionable-policy-solutions/.

[94] Press Information Bureau, Gov't of India, *[Title of the Press Release]*, PIB (Apr. 2026), https://www.pib.gov.in/PressReleasePage.aspx?PRID=2238191&reg=3&lang=1.

encompasses both equality and epistemic justice.[95]

AI governance must be survivor-centered, feminist, and intersectional, recognizing gendered harms not as incidental side effects but as central risks requiring regulation throughout the AI lifecycle.

6. Conclusion

GenAI and deepfake technologies are intensifying already-existing structures of patriarchy, racism, and platform power, turning digital spaces into sites of synthetic, gendered, and epistemic violence that current governance regimes are not equipped to handle. GenAI and deepfakes dramatically scale technology-facilitated gender-based violence by making non-consensual sexual imagery, impersonation, sextortion, and targeted harassment cheap, fast, and anonymous, with women and girls, especially those who are racialized, disabled, queer, or otherwise marginalized, bearing the overwhelming burden of harm. These abuses are not isolated incidents but expressions of structural inequality: algorithmic bias encodes historical sexism and racism into "objective" systems, while platform economies and ranking algorithms actively amplify misogynistic and disinformation content, eroding victims' credibility (testimonial injustice) and even the possibility of naming their experiences (hermeneutical injustice).

At the same time, global AI and platform regulation is fragmented into "regulatory islands" the EU's rights-based AI Act, the US's sectoral patchwork, the UK's soft principles, China's security-oriented deepfake rules, and regulatory uncertainty elsewhere, which leaves cross-border TFGBV, digital misogyny, and algorithmic discrimination without coherent protection or redress. Because platforms like X, Meta, and Reddit operate as active amplifiers rather than passive hosts, safe-harbour models that assume neutrality no longer match technical reality, and their design choices (engagement-driven ranking, delayed takedowns, weak and biased moderation) directly shape the scale of harm.

AI governance must become survivor-centred, feminist, and intersectional, embedding survivor-centred frameworks, gender-sensitive moderation, safety-by-design, transparency, accountability, stronger consent and data protections, digital literacy, and international

[95] Swept AI, *Detecting Intersectional Unfairness in AI* (Feb. 7, 2026), https://www.swept.ai/post/detecting-intersectional-unfairness-in-ai.

cooperation as concrete, enforceable duties across the AI lifecycle. Only by treating gendered and intersectional harms as core design and regulatory risks, not side-effects, can AI and platform governance begin to repair epistemic injustice, redistribute digital power, and make synthetic media compatible with democracy and gender equality.

References

Acquier, A., & Cossey, J. (2025). Generative AI, academic deepfakes, and epistemic pollution. *Business & Society*, 65. https://doi.org/10.1177/00076503251406457

Anastasi, S. (2025). *Misogyny beyond borders: A cross-linguistic corpus assisted analysis of transnational incel communities* (Doctoral dissertation, University of Genova). https://hdl.handle.net/20.500.14242/218817

Ayodeji, U. M. (2025). Examining the impact of technology-facilitated gender-based violence on the mental health and wellbeing of adolescents. *Current Journal of Applied Science and Technology*, 44(5), 66–77. https://doi.org/10.9734/cjast/2025/v44i54537

Babaei, R., Cheng, S., Duan, R., & Zhao, S. (2025). Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks*, 14(1), 17. https://doi.org/10.3390/jsan14010017

Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, 27, 947. https://doi.org/10.1016/j.tics.2023.06.008

Calzada, I., Németh, G., & Al-Radhi, M. S. (2025). Trustworthy AI for whom? GenAI detection techniques of trust through decentralized Web3 ecosystems. *Big Data & Cognitive Computing*, 9(3), 62. https://doi.org/10.3390/bdcc9030062

Canadian Security Intelligence Service. (2023, October 1). *The evolution of disinformation: A deepfake future*. https://www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future/the-evolution-of-disinformation-a-deepfake-future.html

Casilli, A. A., & Tubaro, P. (2026). The organization of algorithmic management: A comparative analysis of digital labour platforms. *Information & Organization*, 45, 100612. https://doi.org/10.1016/j.infoandorg.2026.100612

Cervini, E. M. L. F., & Carro, M. V. (2024, October 25). An overview of the impact of GenAI and deepfakes on global electoral processes. *Italian Institute for International Political Studies*. https://www.ispionline.it/en/publication/an-overview-of-the-impact-of-genai-and-deepfakes-on-global-electoral-processes-167584

Chordia, Y., & Chhajerh, K. (2026, March 6). From shield to sword: How safe harbour became the state's tool of platform control. *RSRR*. https://www.rsrr.in/post/from-shield-to-sword-how-safe-harbour-became-the-state-s-tool-of-platform-control

Cobbe, J., & Singh, J. (2019). Regulating recommending: Motivations, considerations, and principles. *European Journal of Law and Technology*, 10(3). https://ejlt.org/index.php/ejlt/article/view/686/979

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 139.

Croitoru, F-A., et al. (2024). *Deepfake media generation and detection in the generative AI era: A survey and outlook*. arXiv. https://doi.org/10.48550/arXiv.2411.19537

Cupac, J. (2025). *The gender code: Gendering the global governance of artificial intelligence*. arXiv. https://doi.org/10.48550/arXiv.2512.09570

D-ID. (n.d.). What is synthetic media? Benefits & applications. https://www.d-id.com/resources/glossary/synthetic-media/

Das, S. S., Agarwal, M., Rajan, S., & Padhi, A. (2025). Synthetic realities: Youth media literacy and trust in the age of digital deception. *Information Development*. https://doi.org/10.1177/02666669251374658

Department for Science, Innovation & Technology. (2024, February 6). *Implementing the UK's AI regulatory principles: Initial guidance for regulators*. https://www.gov.uk/government/publications/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators

Digital Divide Data. (n.d.). Bias in facial recognition systems for computer vision. https://www.digitaldividedata.com/blog/bias-in-facial-recognition-systems-for-computer-vision

Eltaher, F., et al. (2025). *Protecting young users on social media: Evaluating the effectiveness of content moderation and legal safeguards on video sharing platforms*. arXiv. https://arxiv.org/abs/2505.11160

Epistemic injustice in the digital age: Social media, silencing, and the politics of credibility. (2025). *Journal of Humanities and Educational Development*, 7(3), 18. https://doi.org/10.22161/jhed.7.3.4

European Commission. (n.d.). The Digital Services Act (DSA): Impact on platforms. https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms

European Commission. (n.d.). AI Act: Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Parliament. (2023, June 8). EU AI Act: First regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

European Parliament. (2024, December 6). Cyberviolence against women: What is it and how to prevent it? https://www.europarl.europa.eu/topics/en/article/20241205STO25880/cyberviolence-against-women-what-is-it-and-how-to-prevent-it

Falbo, A. (forthcoming). Hermeneutical injustice. In J. Dancy, E. Sosa, M. Steup & K. Sylvan (Eds.), *The Blackwell companion to epistemology* (3rd ed.). Blackwell.

Fallis, D. (2004). On verifying the accuracy of information: Philosophical perspectives. *Library Trends*, 52(3), 463–487.

Foundation for Media Alternatives. (2026, February 27). *Technology-facilitated gender-based violence (TFGBV) in the Philippines: Year-end data mapping report*. https://fma.ph/technology-facilitated-gender-based-violence-tfgbv-in-the-philippines-year-end-data-mapping-report/

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Ganesh, M. I. (2023). The limits of machine learning in automating racialized surveillance. *Social Sciences*, 12, 435. https://doi.org/10.3390/socsci12080435

Giannini, A. (2023). *Criminal behavior and accountability of artificial intelligence systems* (Doctoral thesis, Maastricht University & University of Florence). https://doi.org/10.26481/dis.20231124ag

Graef, I., & Bostoen, F. (2026). A typology of platform power and its regulation. *Information, Communication & Society*, 29, 324. https://doi.org/10.1080/1369118X.2025.2512972

Hameed, S., Tyabashe-Phume, B., Tunggal, E., Hunt, X., Ned, L., & Soldatić, K. (2025). Technology-facilitated gender-based violence against women with disabilities in low- and middle-income countries: A scoping review protocol. *BMJ Open*, 15, e093988. https://doi.org/10.1136/bmjopen-2024-093988

Hausknecht, A. (2025). The impact of deepfakes on trust in user-generated evidence. In *Deepfakes and the law: Challenges, responses, and critique*. Taylor & Francis.

Heiss, R., & Freiling, I. (2026). Addressing social media platforms' influence on academic research. *Humanities & Social Sciences Communications*, 13, 192. https://doi.org/10.1057/s41599-026-06690-6

Heilweil, R. (2020, June 29). How deepfakes could actually do some good. *Vox*. https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya

Ho, J. Q. H., Hartanto, A., Koh, A., & Majeed, N. M. (2025). Gender biases within artificial intelligence and ChatGPT: Evidence, sources of biases and solutions. *Computers in Human Behavior: Artificial Humans*, 4, 100145. https://doi.org/10.1016/j.chbah.2025.100145

IM Human. (2023, August 29). Unveiling the legal battle: OpenAI faces lawsuit over data collection practices. https://www.imhuman.ai/blog/unveiling-the-legal-battle-openai-faces-lawsuit-over-data-collection-practices

Integrity Institute. (n.d.). *Prevention by design: A roadmap for tackling TFGBV at the source*. https://www.integrityinstitute.org/research/pwojclz543oq3n3n9uk7ro644y0kcm

Javadpour, A., Ja'fari, F., Taleb, T., Shojafar, M., & Benzaïd, C. (2024). A comprehensive survey on cyber deception techniques to improve honeypot performance. *Computers & Security*, 140, 103792. https://doi.org/10.1016/j.cose.2024.103792

Jha, R. (2024). Algorithmic bias and social inequality in AI decision-making systems from a sociological perspective. *ShodhKosh: Journal of Visual and Performing Arts*, 5(6), 3151–3156. https://doi.org/10.29121/shodhkosh.v5.i6.2024.6167

Jindal, A. (2022). Misguided artificial intelligence: How racial bias is built into clinical models. *Brown Journal of Hospital Medicine*, 2, 38021. https://doi.org/10.56305/001c.38021

Kalmykov, M. (2023, November 28). Positive applications for deepfake technology. *DataArt*. https://www.dataart.com/blog/positive-applications-for-deepfake-technology-by-max-kalmykov

Leslie, D. (2020). *Understanding bias in facial recognition technologies: An explainer*. Alan Turing Institute. https://doi.org/10.5281/zenodo.4050457

Li, B., & Cheng, Y. (2026). Mapping the research landscape of algorithmic control on workers: A bibliometric analysis. *Heliyon*, 12, e44403. https://doi.org/10.1016/j.heliyon.2025.e44403

Lyu, S. (2025, December 29). Deepfakes leveled up in 2025 — here's what's coming next. *The Conversation*. https://theconversation.com/deepfakes-leveled-up-in-2025-heres-whats-coming-next-271391

Ma, X. (2025, February 27). Deep synthesis not deepfake: How AI compliance works in China. *China Law Vision*. https://www.chinalawvision.com/2025/02/digital-economy-ai/deep-synthesis-not-deepfake-how-ai-compliance-works-in-china/

Magramo, K., Lau, C., & Jiang, J. (2024, February 4). Finance worker pays out $25 million after video call with deepfake 'Chief Financial Officer'. *CNN*. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

Manoj. (2026, February 20). State of AI content moderation 2026. *Foiwe*. https://www.foiwe.com/state-of-ai-content-moderation-2026/

Marín-López, A. M., & Pérez-Ramos, M. I. (2026). The gendered dynamics of trust and artificial intelligence: Implications for human–AI interaction. *Frontiers in Human Dynamics*, 4, 1790324. https://doi.org/10.3389/fhumd.2026.1790324

Mesko, B. (2023). The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. *Journal of Medical Internet Research*, 25, e48392.

Metz, C. (2025, March 31). OpenAI closes deal that values company at $300 billion. *The New York Times*. https://www.nytimes.com/2025/03/31/technology/openai-valuation-300-billion.html

Mingeirou, K., Osman, Y., & Rafin, R. (2026, February 26). The impact of artificial intelligence on violence against women and girls. *Stimson Center*.

https://www.stimson.org/2026/the-impact-of-artificial-intelligence-on-violence-against-women-and-girls/

MIT News. (2018, February 12). Study finds gender and skin-type bias in commercial artificial-intelligence systems. https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

Monteiro, N. SJ., & Singh, V. (2025). The wheel of artificial intelligence governance. *Sustainable Futures*, 101279. https://doi.org/10.1016/j.sftr.2025.101279

Nayak, R., & Sircar, A. (2024). Algorithmic content moderation, IP and expression: Navigating the tension through theories. *NLUA Journal of Intellectual Property Rights*, 3, 1.

Office of the United Nations High Commissioner for Human Rights. (2025). *Technology facilitated gender-based violence*. https://www.ohchr.org/sites/default/files/documents/issues/women/genderandequality/2025-tool-technology-facilitated-gbv.pdf

Park, C. H., Rao, Z., Ji, L., & Chen, Q. (2024). (Re)mediators of epistemic injustice: Generative AI and hermeneutic resource provision in intimate partner violence. In *Companion proceedings of the 2024 ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 1). https://doi.org/10.1145/3772318.3791548

Pequeño, A. IV. (2026, April 27). Nvidia sets new record with nearly $5.3 trillion value after AI darling surges 4%. *Forbes*. https://www.forbes.com/sites/antoniopequenoiv/2026/04/27/nvidia-sets-new-record-with-nearly-53-trillion-value-after-ai-darling-surges-4/

Perriello, L. E. (2026). Blurred realities: Legal strategies for the deepfake era. *International Review of Law, Computers & Technology*. https://doi.org/10.1177/1023263X261433380

Press Information Bureau, Government of India. (2026, April). https://www.pib.gov.in/PressReleasePage.aspx?PRID=2238191&reg=3&lang=1

Rahman-Jones, I. (2024, January 27). Taylor Swift deepfakes spark calls in Congress for new legislation. *BBC News*. https://www.bbc.com/news/technology-68110476

Rai, O. P. (2026). Artificial intelligence, data protection and digital governance in India: A contemporary legal analysis. *Journal of Social Review and Development*, 5(1), 76–81. https://doi.org/10.64171/JSRD.5.1.76-81

Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: Policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, 8, 1569115. https://doi.org/10.3389/frai.2025.1569115

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

Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Computer Law & Security Review*, 58, 106162. https://doi.org/10.1016/j.clsr.2025.106162

Rose, A. (2026, February 2). The real threat of AI is the collapse of trust. *Poynter*. https://www.poynter.org/commentary/2026/the-real-threat-of-ai-is-the-collapse-of-trust-deepfakes/

Sabbagh, D. (2023, November 10). Faked audio of Sadiq Khan dismissing Armistice Day shared among far-right groups. *The Guardian*. https://www.theguardian.com/politics/2023/nov/10/faked-audio-sadiq-khan-armistice-day-shared-among-far-right

Schick, N. (2020). *Deep fakes and the infocalypse: What you urgently need to know*. Monoray.

Seipel, B. (2025, November 13). The world's most deepfaked celebrities revealed. *McAfee*. https://www.mcafee.com/blogs/internet-security/the-stars-scammers-love-most-mcafee-reveals-worlds-most-deepfaked-celebs/

Sexual Violence Research Initiative. (n.d.). TFGBV safety by design. https://www.svri.org/tfgbv-safety-by-design/

Spring, M., & Radford, M. (2026, March 16). Meta and TikTok let harmful content rise after evidence outrage drove engagement, say whistleblowers. *BBC News*. https://www.bbc.com/news/articles/cqj9kgxqjwjo

Stein, C., et al. (2025). The health effects associated with physical, sexual and psychological gender-based violence against men and women: A burden of proof study. *Nature Human Behaviour*, 9, 1201. https://doi.org/10.1038/s41562-025-02144-2

Swept AI. (2026, February 7). Detecting intersectional unfairness in AI. https://www.swept.ai/post/detecting-intersectional-unfairness-in-ai

Telnova, H. V., & Reshetnyak, T. V. (2026). Theoretical and methodological foundations of modern industry market analytics. *Efektyvna Ekonomika*, 2, 43. http://doi.org/10.32702/2307-2105.2026.2.43

The BIRM Group. (2026). AI infrastructure construction: The next $400B boom in 2026. https://thebirmgroup.com/ai-infrastructure-construction-the-next-400b-boom-in-2026/

UNICEF. (2026, February 4). Deepfake abuse is abuse. https://www.unicef.org/press-releases/deepfake-abuse-is-abuse

United Nations Development Programme Pakistan. (n.d.). Enabling safe digital spaces through a survivor-centred response to technology-facilitated gender-based violence in Pakistan. https://www.undp.org/pakistan/projects/enabling-safe-digital-spaces-through-survivor-centred-response-technology-facilitated-gender-based-violence-pakistan

United Nations News. (2026, March 21). Why women can't get protection from AI deepfake abuse. https://news.un.org/en/story/2026/03/1167174

United Nations News. (2026, April 30). Abuse of women journalists made 'easier and more damaging' by AI. https://news.un.org/en/story/2026/04/1167416

United Nations Population Fund. (n.d.). Technology-facilitated gender-based violence. https://www.unfpa.org/TFGBV

United Nations Population Fund. (n.d.). Creating a safer digital future free from gender-based violence. https://www.unfpa.org/updates/creating-safer-digital-future-free-gender-based-violence

United Nations Population Fund & Gender in Digital Coalition. (2026, April 23). *Concept note: AI for good: Gender inclusion in the AI ecosystem* [Virtual consultation].

UN Regional Information Centre for Western Europe. (2023, November 29). How technology-facilitated gender-based violence impacts women and girls. https://unric.org/en/how-technology-facilitated-gender-based-violence-impacts-women-and-girls

UN Women. (2022). *Accelerating efforts to tackle online and technology-facilitated violence against women and girls*. https://www.unwomen.org/sites/default/files/2022-10/Accelerating-efforts-to-tackle-online-and-technology-facilitated-violence-against-women-and-girls-en_0.pdf

UN Women. (2025). How AI reinforces gender bias — and what we can do about it. https://www.unwomen.org/en/news-stories/interview/2025/02/how-ai-reinforces-gender-bias-and-what-we-can-do-about-it

UN Women. (n.d.). 16 days of activism 2025: End digital violence against all women and girls. https://www.unwomen.org/en/what-we-do/ending-violence-against-women/unite/theme

Van Doorn, N., & Leurs, K. (2026). Reaction videos as "interpassive" aesthetic experiences: Understanding the meaning of diffuse media practices in the platform society. *Communication Review*. https://doi.org/10.1080/10714421.2026.2647307

Verma, P. (2023, October 13). AI voice clones mimic politicians and celebrities, reshaping reality. *The Washington Post*. https://www.washingtonpost.com/technology/2023/10/13/ai-voice-cloning-deepfakes/

Women in Digital. (2025, March 7). Embedding gender equality in the EU's digital future: From AI bias to actionable policy solutions. https://widigital.eu/embedding-gender-equality-in-the-eus-digital-future-from-ai-bias-to-actionable-policy-solutions/