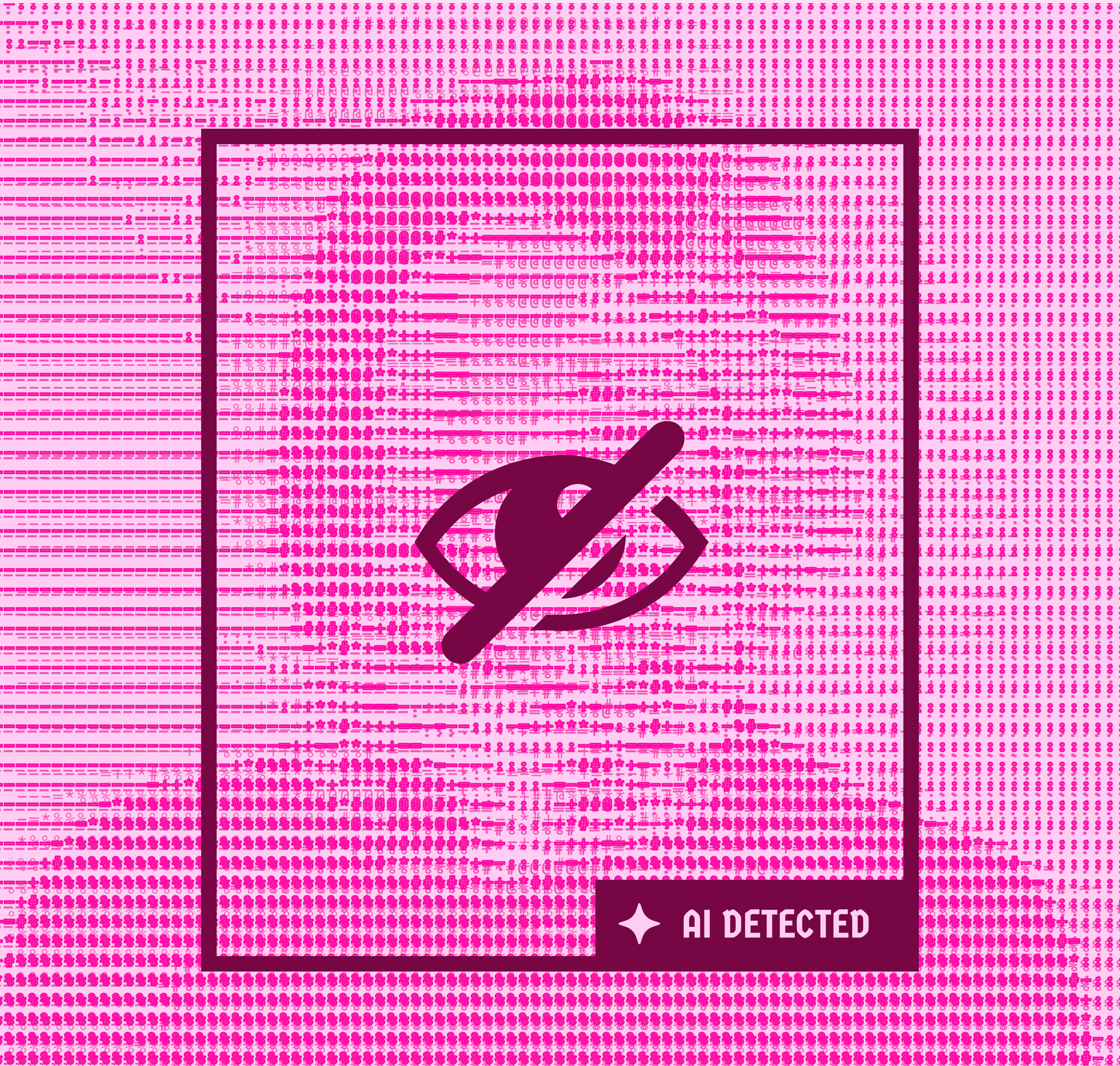
Tattle RATI

REPORT 2025

Make It Real: Mapping AI-Facilitated Gendered Harm



Make It Real: Mapping AI-Facilitated Gendered Harm

Tattle RATI

Date
November 3, 2025

Authors
Siddharth Pillai, Tarunima Prabhakar and Kaustubha Kalidindi

Editor
Tanima Saha

Designer
Yatharth

Acknowledgements
This report builds on the expertise of the entire teams at Rati Foundation and Tattle Civic Tech. In particular, we are grateful to Salman Mujawar, Sameer Parmar, Ritu Jain and Ruta Sawant, counselors at Rati Foundation, whose frontline work with survivors forms the basis of much of the evidence presented here. We would also like to thank Cheshta Arora & Baarish for reviewing an earlier draft of the report.

Contents

Glossary and Acronyms

Image-based Sexual Abuse (IBSA): An act where an individual shares or threatens to share non-consensual intimate images or videos of a person.

Deepfakes: Media where a person's face, body or voice is digitally altered to falsely imply that they said or did something that they did not do.

Artificial Intelligence: In the context of this report, AI refers to computational systems that are able to generate images, audio, video or text in response to prompts. These can be used to create entirely new media or make alterations to existing media.

Tech-facilitated Gender-based Violence (TFGBV): An act of violence against an individual based on their gender using technology.

Online Gender-based Violence (OGBV): Targeted violence against persons based on their gender in online space.

Deep learning: A method of machine learning that enables computers to process large datasets and analyse complex information.

Non-consensual Intimate Imagery (NCII): Intimate images created and distributed without the consent of the person depicted in them.

Impersonation: Using a real person's identity to send or post vicious or embarrassing material to/about others.

Doxxing: Publishing private or identifying information about a person on the internet, typically without their consent and with malicious intent.

Introduction

In the recent few years, there have been rapid advancements in AI systems being able to produce realistic visual and audio content. This has led to concerns around a spike in near-realistic but inaccurate content that can be used to bully, defame and target individuals. In fact, it has become evident in the last three years that a vast majority of AI-generated content is used to target women and gender minorities.

Even in its longer trajectory, the production and consumption of AI is deeply sexualized. The earliest use of AI in video generation was for pornographic content. The term "deepfake" originated in a Reddit thread in 2017, which posted videos that used AI to insert celebrities' likenesses into existing pornographic videos.[1] In 2019, 96% of all deepfake content was pornographic.[2] With the advancements in AI, more forms of manipulation have continued to emerge. For example, advances in capabilities and accessibility of image diffusion models led to a spike in applications that could manipulate existing photos and video footage of real individuals to make them appear nude without their consent.[3] AI tools have also adapted to the varying cultural expressions of intimacy.[4] Some applications, when provided with images of two individuals, produce morphed images of them hugging or kissing. In certain contexts, such manipulated images can be sufficient to stigmatize women in their communities.

Meri Trustline

Meri Trustline is an India-based helpline launched by Rati Foundation that supports individuals facing online risks by ensuring response and redressal through content takedown, mental health counselling and providing social and legal support. Individuals can reach the Trustline through phone, WhatsApp, email and through an online form. Since its inception in 2022, the Trustline has handled more than 482 cases. A case is defined as a report from a victim-survivor that requires intervention through one of the aforementioned ways. A case may involve a combination of attacks, such as hacking and impersonation. It may also involve more than one perpetrator and/or on more

[1] More specifically, deep learning techniques.

[2] Henry, A., Giorgio, P., Francesco, C., and Laurence, C. (2019). 'The State of Deepfakes: Landscape, Threats, and Impact'. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[3] Santiago, L., (2023). 'A Revealing Picture: AI generated "Undressing" images move from niche pornography discussion forums to a scaled and monetised online business'. Graphika. https://graphika.com/reports/a-revealing-picture.

[4] Gilani, M. (2025, January 22). 'Pakistan chief minister targeted by AI "hug" video. AFP FactCheck. https://factcheck.afp.com/doc.afp.com.36TP83J

than one account. While a majority of those who reach out to the Trustline identify themselves as women or gender and sexual minorities, approximately 25% of the cases involve men as victim-survivors.

In total, cases from over 26 states in India have been reported. Around a quarter of the reports come from Maharashtra, followed by Uttar Pradesh and Delhi. Other states with high reporting include Karnataka and West Bengal. A small proportion, less than 5%, originated outside India and involved either an Indian-origin victim or perpetrator.

In 2022, following a steep rise in incidents involving altered images, the Trustline began tracking cases where content was digitally manipulated.[5] The initial cases included manipulation by adding suggestive clipart, juxtaposing with sexualised imagery or text or editing the picture through Photoshop. As reports of deepfakes began to appear, the same term was extended to include those cases as well. Roughly 10% of the cases reported to the Trustline involved such digitally manipulated material. But in response to those, the Trustline team escalated and successfully helped in the take down of over 150 offending accounts across different platforms. A detailed analysis of the cases reported on the Trustline can be found in Meri Trustline's annual reports.[6]

Scope of Report

A key trend observed in the cases reported to the Trustline is the rise in AI-generated content and threats of creating 'deepfakes' for harassment.[7] While media coverage tends to focus on the use of AI to target public figures such as celebrities and politicians, the Trustline provides insight into the more private experience of online harassment, which, due to the stigma and trauma associated with harassment, are not reported to family, law enforcement or media. Unlike public figures, who are targeted primarily as a symbol of the gender and sexual identity group they represent and in the public sphere, many of the survivors who approach the Trustline are targeted by

[5] Digitally Manipulated is the terminology that Rati Foundation uses internally to describe situations where perpetrators modify an existing media item to harass someone.

[6] Rati Foundation. (n.d.). Meri Trustline. Retrieved August 4, 2025 from https://ratifoundation.org/meri-Trustline/.

[7] That is marked as AI-generated content.

acquaintances with an intent to specifically target them.[8,9,10] In some cases of one-on-one harassment, people might be targeted because of existing socio-cultural vulnerabilities but the domain of harassment remains private. While the social status of celebrities, politicians and public figures makes them an easy target for public harassment, it also affords a degree of resilience and mediated control over reputation.

The aim of this report is to humanize the less visible and more private experiences of online harassment. It brings forward unique evidence on the impact of AI on ordinary women and gender minorities navigating the online world. Through the lens of AI, the report reflects on the suitability, or the lack thereof, of existing avenues to protect and empower victims of online harassment.

Note on Terminology

This report is concerned with how AI is interweaving into the complex terrain of online harassment. There are numerous terms that are used to describe aspects of online harassment and the role of technology in it. A common overarching term used to describe online harassment involving visual content is Image-based Sexual Abuse (IBSA). While many of the cases that come to the Trustline are IBSA in nature, this isn't a term that the Trustline uses for internal cataloguing. The Trustline has also observed some cases of audio, digitally manipulated or otherwise, being used to harass victims. For images and videos authentically recorded and shared without consent, the term Non-consensual Intimate Imagery (NCII) has been used. Images, videos and audio that have some visible manipulation have been referred to as digitally manipulated content. AI-generated content is a subset of digitally manipulated content.

This report uses the terms "victim" and "survivor" interchangeably, and sometimes even "victim-survivors". "Victim" highlights the severity of harm and ongoing vulnerability while also countering narratives that blame the harmed. "Survivor" signals agency, resilience and the refusal to be defined solely by the experience of harm. The combined and interchangeable use of these terms reflects the complex and evolving realities of those affected by digital abuse.[11]

[8] Bureau, (2024, January 21) 'Delhi police arrest techie from Andhra Pradesh for Rashmika Mandanna deepfake video'. The Hindu. https://www.thehindu.com/news/cities/Delhi/delhi-police-arrest-techie-from-andhra-pradesh-for-rashmika-mandanna-deepfake-video/article67760419.ece

[9] Southern, Rosalynd and Harmer, Emily. (2021). 'Twitter, Incivility and "Everyday" Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament'. Social Science Computer Review. Sage Journals. 39. 259–75. 10.1177/0894439319865519.

[10] Bureau, (2024, January 21)

[11] Lakshané, R. (2024, August 29). 'The Crying Shame of Image-based Abuse'. Factory Daily. https://factordaily.com/the-crying-shame-of-image-based-abuse/

Literature Review

Theorizing Online Gendered Harassment

Emerging research has highlighted the pervasiveness of online gendered harassment across countries.[12,13,14] Building on this evidence, there have been attempts to categorize experience of gender harassment based on the kind of act; the relationship between victims and target; and platform affordances.[15,16] Harassment can also be analysed based on the motivations for harassment. Harassment, especially in one-on-one acts, is generally attributed to an intent to harm or inflict suffering on targets. But scholars have also placed harassment as a part of "complex social processes among the hate messengers themselves". Social gratification from online communities, rather than a desire to harm the victim, can also be the driver of the act of harassment.[17] Finally, online harassment is also embedded in an online economy, where content can be used to extract money or sexual favors from a victim.[18] In some cases, such as in the Gamergate harassment campaign, these different motivations comingle. What starts as a targeted attack from a known individual can snowball into a coordinated attack by strangers online.

Understanding the impact of online harassment is more complex. The impact depends on a number of contextual factors, such as the vulnerability and agency of the target, the availability of psycho-social and institutional support and the scale and kind of attack. Scheruman et al., document four kinds of harms from online content: physical harm; emotional harm; relational harm; and financial harm. They propose nine dimensions against which the severity of these harms can be assessed. These include the perceived intent of the act by the victim; the agency with the victim; the scale of the attack; and the medium of the act amongst others.[19]

[12] Dunn, S., Vaillancourt, T. and Brittain, H. (2023). Supporting Safer Digital Spaces. Centre for International Governance Innovation. https://www.cigionline.org/programs/supporting-safer-digital-spaces/

[13] Im, J., Schoenebeck, S., Iriarte, M., Grill, G., Wilkinson, D., Batool, A., Alharbi, R., Funwie, A., Gankhuu, T., Gilbert, E., and Naseem, M. (2022). 'Women's Perspectives on Harm and Justice after Online Harassment'. Proceedings of the ACM on Human-Computer Interaction. Association for Computing Machinery. 6(CSCW2).

[14] Social Development Direct. (2023). 'Technology-facilitated gender-based violence: preliminary landscape analysis'. The Global Partnership for Action on Online Gender Based Abuse. https://www.gov.uk/government/publications/technology-facilitated-gender-based-violence-preliminary-landscape-analysis

[15] UNFPA. (n.d.) 'The Background'. Retrieved August 4, 2025 from https://www.unfpa.org/thevirtualisreal-background#glossary

[16] Tong, S.T. (2024). 'Foundations, definitions, and directions in online hate research', in Social Processes of Online Hate, eds. Walther, J.B., and Rice, E.R. (pp. 36). (2024). Routledge. https://doi.org/10.4324/9781003472148

[17] Walther, J. (2024). Making a Case for a Social Processes Approach to Online Hate Research', in Social Processes of Online Hate, eds. Walther, J.B., and Rice, E.R.(pp 9–36). Routledge. https://doi.org/10.4324/9781003472148-2

[18] Lakshané, R. (2024, August 29). 'The Crying Shame of Image-based Abuse'. Factory Daily. https://factordaily.com/the-crying-shame-of-image-based-abuse/

[19] Scheuerman et al. (2021). 'A Framework of Severity for Harmful Content Online'. Arxiv. https://arxiv.org/abs/2108.04401

Dimensions of Severity of Online Harm Perceived by Survey Respondents* Scheuerman et al. (2021)

Dimension	Description
Perspective	The severity of harm varied depending on whose perspective is being taken to rank the harm. The authors note three perspectives: the target, viewer and perpetrator. The harm is perceived to be higher by the target, and lower by the perpetrator.
Intent	The severity of harm is dependent on whether the harm was intentional or not. The harm is perceived to be proportional to the intention to harm.
Agency	The agency of the person harmed—whether they had a choice to participate in either the harm or circumstances leading up to the harm—was influential in the perception of harm. If the individual had lower agency, the harm was perceived to be higher. But harms associated with a lack of choice were considered more severe.
Experience	Personal experience with a certain category of harm led people to consider them more severe.
Scale	Harm depended on the number of people impacted or the number of actors dedicated to harming an individual or group. The larger the scale, the more is the perceived harm.
Urgency	The level of urgency or time-sensitivity with which action needed to be taken is perceived to be correlated with harm. The greater the time sensitivity, the more harmful is the content perceived to be.
Vulnerability	If the target is vulnerable such as children or populations that have lesser privilege, the harm is perceived to be greater.
Medium	The harm is also perceived to be linked to the medium. For example, harm from visual content was perceived to be worse than textual content.
Sphere	The sphere where harm took place—on public posts or in private direct messages—also affects the perception of harm. For example, targeting through private messages can often make those targeted feel more alone.

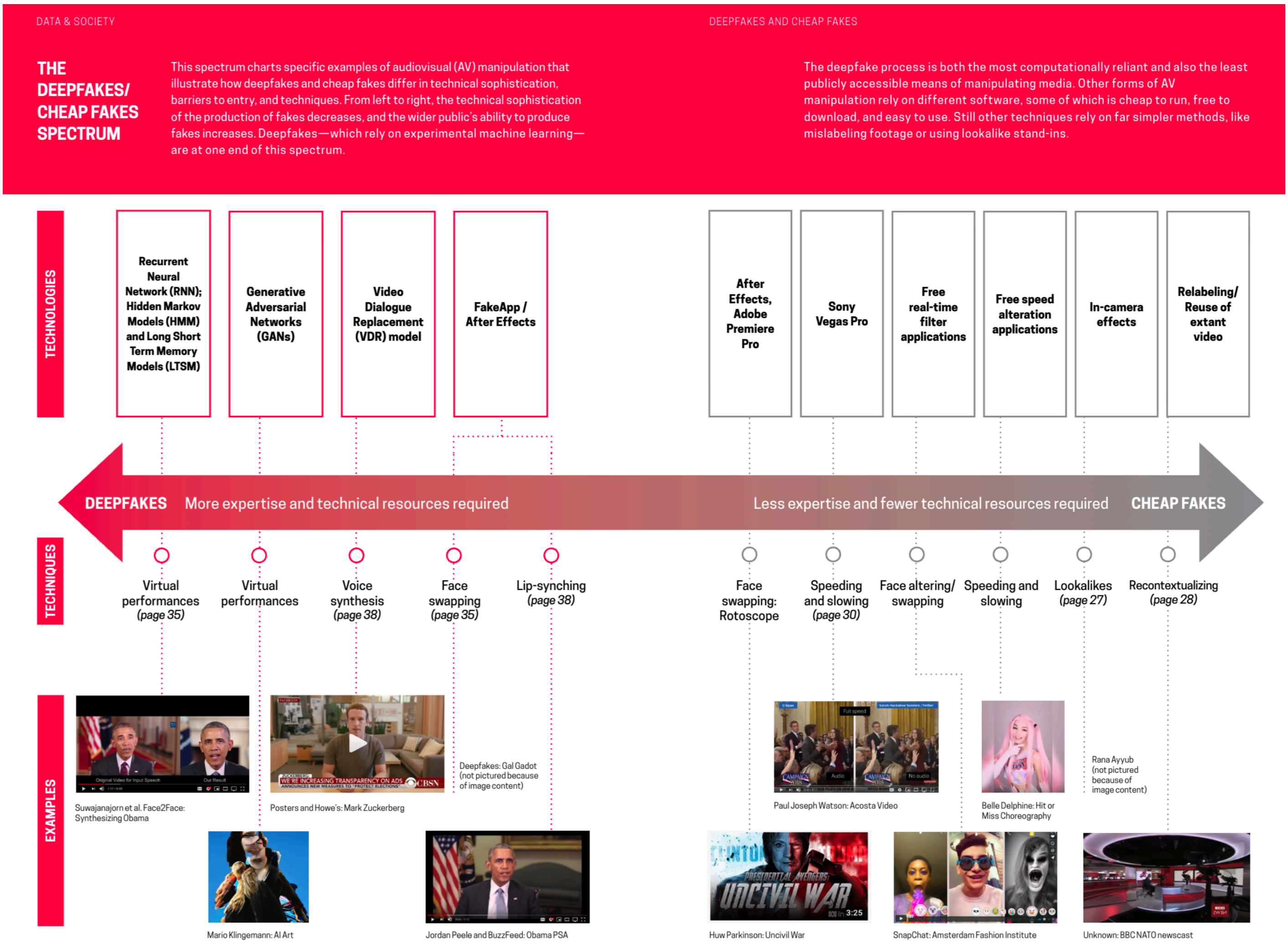
Specifically in the context of Meri Trustline, the literature on intimate partner violence (IPV) is also pertinent. Often harassers known to targets use tactics designed to cause shame in their targets, "... including denigrating their dignity, undermining their autonomy, or harming their reputation." Similar to IPV survivors, "... victims of online harassment may come out with an abiding sense of shame as a result of their victimization—from a lost sense of self, to self-blame, to fear of (or actual) social judgment."[20]

[20] Camp, R. (2022). 'From Experiencing Abuse to Seeking Protection: Examining the Shame of Intimate Partner Violence'. UC Irvine Law Review. https://escholarship.org/content/qt6jr1d4sq/qt6jr1d4sq.pdf

Theorizing AI-Generated Content and 'Deepfakes'

The rise of Generative AI models has eased the creation of digital content, such as videos, images, music, and natural language. With innovation within Generative AI, the diversity of AI Generated Content (AIGC) has increased. AIGC is a subset, though potentially the dominant form of, synthetic content. The level of manipulation through AI, and the likeness of AIGC to a real-world media item, falls on a spectrum.[21,22] The most sophisticated manipulations, such as "virtual performances" that show women in pornographic acts that they did not shoot or politicians giving speeches that they never did, are nearly entirely AI-generated. Similarly, AI can also be used to mimic individuals' voices to generate entirely synthetic audio. Lower in the spectrum is the use of AI to make modifications within a media item such as creating lip-syncs of a video and changing the complexion or clothing of a person.

Figure: Paris B., and Donovan, J. (2019). 'Deepfakes and Cheapfakes: The Manipulation of Audio and Visual Content'. Data & Society. https://datasociety.net/library/deepfakes-and-cheap-fakes/



DATA & SOCIETY

DEEPFAKES AND CHEAP FAKES

THE DEEPFAKES/ CHEAP FAKES SPECTRUM

This spectrum charts specific examples of audiovisual (AV) manipulation that illustrate how deepfakes and cheap fakes differ in technical sophistication, barriers to entry, and techniques. From left to right, the technical sophistication of the production of fakes decreases, and the wider public's ability to produce fakes increases. Deepfakes—which rely on experimental machine learning—are at one end of this spectrum.

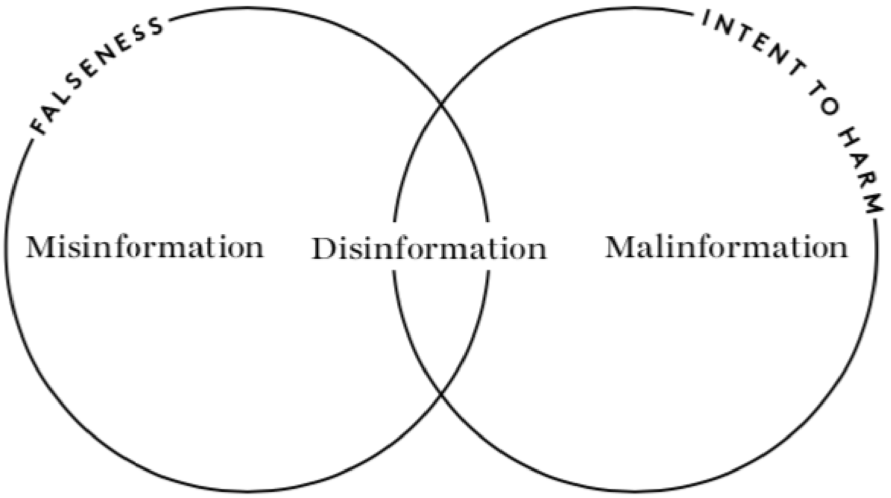
The deepfake process is both the most computationally reliant and also the least publicly accessible means of manipulating media. Other forms of AV manipulation rely on different software, some of which is cheap to run, free to download, and easy to use. Still other techniques rely on far simpler methods, like mislabeling footage or using lookalike stand-ins.

TECHNOLOGIES
Recurrent Neural Network (RNN); Hidden Markov Models (HMM) and Long Short Term Memory Models (LTSM)
Generative Adversarial Networks (GANs)
Video Dialogue Replacement (VDR) model
FakeApp / After Effects
After Effects, Adobe Premiere Pro
Sony Vegas Pro
Free real-time filter applications
Free speed alteration applications
In-camera effects
Relabeling/ Reuse of extant video
DEEPFAKES
More expertise and technical resources required
Less expertise and fewer technical resources required
CHEAP FAKES
TECHNIQUES
Virtual performances (page 35)
Virtual performances
Voice synthesis (page 38)
Face swapping (page 35)
Lip-synching (page 38)
Face swapping: Rotoscope
Speeding and slowing (page 30)
Face altering/ swapping
Speeding and slowing
Lookalikes (page 27)
Recontextualizing (page 28)
EXAMPLES
Suwajanajorn et al. Face2Face: Synthesizing Obama
Posters and Howe's: Mark Zuckerberg
Deepfakes: Gal Gadot (not pictured because of image content)
Mario Klingemann: AI Art
Jordan Peele and BuzzFeed: Obama PSA
Paul Joseph Watson: Acosta Video
Belle Delphine: Hit or Miss Choreography
Rana Ayyub (not pictured because of image content)
Huw Parkinson: Uncivil War
SnapChat: Amsterdam Fashion Institute
Unknown: BBC NATO newscast

[21] Paris B., and Donovan, J. (2019). 'Deepfakes and Cheapfakes: The Manipulation of Audio and Visual Content'. Data & Society. https://datasociety.net/library/deepfakes-and-cheap-fakes/

[22] Human-AI Spectrum (n.d.). 'VerifiedHuman™ Human-AI Collaborative Spectrum'. Retrieved August 5, 2025, from https://www.iamverifiedhuman.com/human-ai-spectrum

A number of research efforts have tried to map the risks of AI-generated content.[23,24] Many of these risks are seen as a continuation or supercharging of risks of older forms of digital manipulation. For example, there is a concern that AIGC leads to a dilution of the authenticity of content.[25] Scholars, for example, have attempted to assess the believability of AI-generated content vis-a-vis human created media items, or more traditional ("photoshop") forms of manipulation.[26,27] Some other scholars have proposed that challenges of AIGC to authenticity are more foundational with it "... not just being indistinguishable from human production but it reshaping the very grounds upon which we understand authenticity and experience."[28] AIGC, seen from the perspective of authenticity, or its lack thereof, falls neatly in the existing framing of an "infodemic" emerging from the intersection of content that is false or intended to harm.



Beyond the lens of authenticity, AIGC is also increasingly deployed to produce memes and outright satirical content.[29,30] AI has eased the creation of visual content through simple text-based prompts. Memes cannot be easily classified as content that is intended to harm. Operating on a "logic of lulz", memes enable participatory collectives through "... a detached and dissociated amusement at others' distress."[31] Memes serve as the ammo of trolling, which can be "equal opportunity laughter" but still disproportionately target minorities and women.[32]

[23] Wittenberg, C., Epstein, Z., Berinsky, A.J., and Rand, D.G. (March 27, 2024). 'Labeling AI-Generated Content: Promises, Perils, and Future Directions.' An MIT Exploration of Generative AI. https://doi.org/10.21428/e4baedd9.0319e3a6

[24] Guo, D., Chen, H., Wu, R., and Wang, Y. (2023). 'AIGC challenges and opportunities related to public safety: A case study of ChatGPT'. Journal of Safety Science and Resilience, 4(4), 329–39. doi:10.1016/j.jnlssr.2023.08.001

[25] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan. (2023) 'A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions'. IEEE Open Journal of the Computer Society, 4, 280–02. doi:10.1109/OJCS.2023.3300321

[26] Duestad, L.L., Foss, H.C., Toth, J., and Gleasure, R. (2025). 'Pictures or It Didn't Happen! How the Use of the Generative AI Images Impacts the Perceived Believability of News Headlines', in Information Systems and Neuroscience, eds F.D. Davis, R. Riedl, J. vom Brocke, P. M. Léger, A. B. Randolph, & G. R. Müller-Putz (Eds.). (pp. 29–35). Springer. https://doi.org/10.1007/978-3-031-71385-9_4

[27] Hameleers, M. (2024). 'Cheap Versus Deep Manipulation: The Effects of Cheapfakes Versus Deepfakes in a Political Setting'. International Journal of Public Opinion Research, 36(1). https://doi.org/10.1093/ijpor/edae004

[28] Berry, D. (2025). 'Synthetic media and computational capitalism: towards a critical theory of artificial intelligence'. AI & Society. https://doi.org/10.1007/s00146-025-02265-2

[29] Singler, B. (2020). 'The AI Creation Meme: A Case Study of the New Visibility of Religion in Artificial Intelligence Discourse'. Religions, 11(5), 253. https://doi.org/10.3390/rel11050253

[30] Chang,H., et al. (2024). 'Generative Memesis: AI Mediates Political Memes in the 2024 USA Presidential Election'.

[31] Milner, R. (2013). 'FCJ-156 Hacking the Social: Internet Memes, Identity Antagonism, and the Logic of Lulz.' The Fibreculture Journal, 22.

[32] Phillips, W. (2015). 'This Is Why We Can't Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture'. The MIT Press. http://www.jstor.org/stable/j.ctt17kk8k7

Legal Provisions Pertaining to Online Harassment and AI-Generated Content

With technological advancements, crimes, too, take new shape. Generally, reliance is placed upon existing legal provisions for recourse, unless sufficient cause moves lawmakers to enact new laws to tackle certain technology-enabled offences. At present, while laws punishing online harassment exist, there are no specific provisions to tackle offences relating to AI-generated content in India. One may seek recourse by relying on existing umbrella laws, namely: the Bharatiya Nyay Sanhita, 2023 (BNS), which replaced the Indian Penal Code; the Information Technology Act, 2000 (IT Act); and the Protection of Children from Sexual Offences Act, 2012 (POCSO). Broader provisions punishing the transmission of obscene material,[33] or material containing sexually explicit acts in electronic form,[34] may apply to image-based sexual abuse cases. Additionally, under the BNS, perpetrators may be punished for criminal intimidation, defamation, insulting the modesty of a woman or sexual harassment, as the case may be. For content that constitutes CSAM, the law is clear: Section 67B of the IT Act punishes the creation, possession, transmission, and distribution of any CSAM material, and POCSO punishes the creation and usage of any material of children for pornographic purposes, and this may include instances of image-based abuse where the content has been generated by AI.

Online harassment and image-based abuse often occurs on social media and messaging platforms. Platforms are obligated to take moderation decisions with respect to the content they host under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Content that does not adhere to platforms' community guidelines must be taken down upon receipt of a court order in 36 hours, and also provide information to authorized government agencies for the purpose of investigating offences under any law. Subsequent to a memorandum given by the Directorate General of Police, police officers have been given instructions on handling obscenity and NCII cases. More recently, in a case before the Madras High Court where the petitioner's former partner leaked her private videos online[35], the court directed MEITY to submit an affidavit detailing the steps they had initiated and to provide a 'prototype' of recourse available to girls who were victims of such offences. The National Cyber Crime Reporting Portal also allows

[33] Section 67, IT Act, 2000.

[34] Section 67A, IT Act, 2000.

[35] X v. Union of India, 2025. Mad HC. https://www.mhc.tn.gov.in/judis/madras-do/index.php/casestatus/viewpdf/WP_25017_2025_XXX_0_0_15072025_177.pdf

complaints to be filed: there is an anonymous option to file sensitive complaints (for matters relating to NCII and CSAM). In India, even as the government adopts a 'light-touch' approach to AI regulation, advisories have been released to platforms for offences such as creating NCII and deepfakes.[36]

Internationally, different ways to regulate tech-facilitated gender-based violence (TFGBV) offences are being considered, some of which cover AI-generated content:

1. The US enacted the TAKE IT DOWN Act in 2025 to tackle the issue of non-consensual intimate imagery (NCII). It requires platforms, when notified by the subject or someone acting on their behalf, to reasonably identify and remove the content within forty-eight hours. However there have been concerns with regard to its constitutionality and impact on free speech.[37]

2. The UK Online Safety Act contains takedown provisions for illegal pornographic content, including NCII and deepfakes. Ofcom, the regulator for online communications, has taken cognisance of online harms against women and girls, and TFGBV as well.

3. The EU AI Act has disclosure requirements: anyone creating deepfake content must disclose that the content has been artificially generated or face heavy non-compliance penalties.

4. In Mexico, a set of reforms dubbed the 'Olimpia' law recognize and punish digital violence against women, including NCII. Argentina, too, adopted the Olimpia law in recognition of gender-based violence against women.

Platform Reporting

Social media companies, in accordance with intermediary laws around the world, are required to implement policies governing the nature of content that can be hosted and disseminated on their platforms. Accordingly, platforms institute terms of use policies, adapted to what are commonly referred to as 'community guidelines'. YouTube, Meta, X, and all other such social media platforms have community guidelines to lay out

[36] Press Information Bureau. (2023, December 23). MeitY issues advisory to all intermediaries to comply with existing IT rules [Press Release]. https://www.pib.gov.in/PressReleaseIframePage.aspx?PRID=1990542

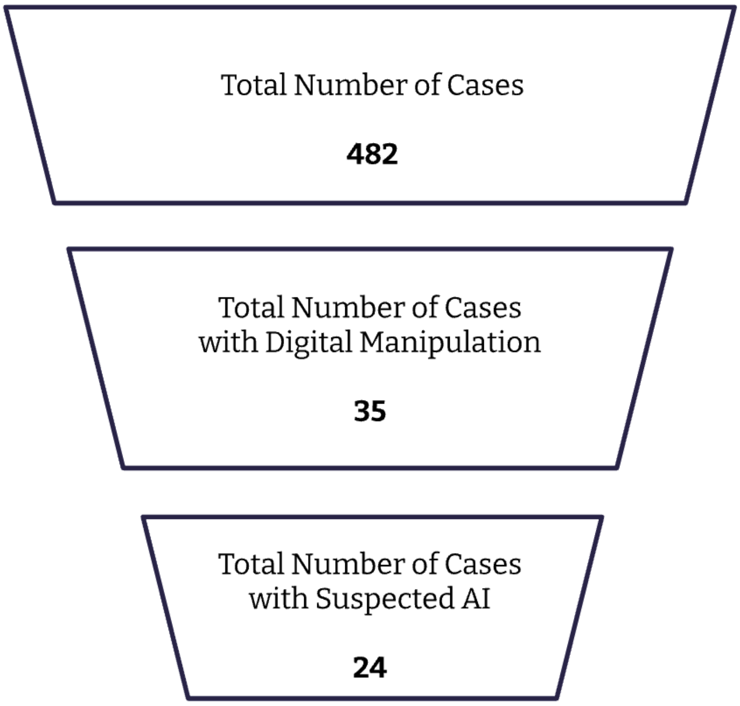
[37] Williams, K. (2025, February 21). 'Free speech advocates express concerns as take it down act passes US senate'. Tech Policy Press. https://www.techpolicy.press/free-speech-advocates-express-concerns-as-take-it-down-act-passes-us-senate/

how users should behave with one another in online spaces, and they also specify content that is prohibited on their platforms. Private messaging such as WhatsApp and Telegram specify prohibited behaviour in their terms of use policies. Online harassment is prohibited in all platform policies, and more recently, platforms have begun to institute guidelines regarding prohibited uses of AI-generated content as well. The table below describes the broad policies that are relevant to AIGC and online harassment.

Platform	Relevant Policies
YouTube	Harassment and Cyberbullying policy Child Safety Nudity and Sexual Content Policy Copyright Policy
Meta	Child Sexual Exploitation Adult Nudity and Sexual Activity NCII and Sextortion IPR Policy Bullying and Harassment Oversight Board Decision on Non Consensual Deepfakes
X	Adult Content Policy Non-Consensual Nudity Policy Child Sexual Exploitation Abuse and Harassment Authenticity Copyright Policy

Evidence from Meri Trustline

Since 2022, over 482 survivors have sought support from Meri Trustline. The majority of cases involve some digital content. This could be sharing of NCII on social media or in messaging groups, or recording a person viewing sexual content for extortion.[38] 35 cases reported have involved digitally manipulated content. The spike in digitally manipulated content is correlated with advancements in AI. Most of the content (24 out of 35) reported as digitally manipulated is suspected to involve some AI-based manipulation.


Total Number of Cases
482
Total Number of Cases with Digital Manipulation
35
Total Number of Cases with Suspected AI
24

In this section, we describe four of the 24 case studies. These four were selected on the basis of the clarity of evidence demonstrating AI manipulation, the severity of harm caused to the victims, the complexities of the abuse and the diversity of contexts they represent.

[38] Also called 'sextortion'

We also describe two cases reported to the Trustline that did not involve the use of AI, to provide a reference for comparative analysis in the discussion section. For each case that comes to the Trustline, the team maintains detailed case histories beginning from the first call till the case closure. A case is formally closed during a case management review once it is determined that the client's primary expectations have been met, a sufficient level of safety has been established and no further intervention is deemed necessary. Case closure is approximately three months after the interventions have ceased. Throughout this process, only the information that is necessary for managing each case is gathered. This data collection is conducted through a consensual method, depending entirely on the client's willingness to provide the information. The Trustline relies on these case histories for the description of each case. While the case histories are rich and document the journey from the first call to a victim affirming that their case was resolved, in this report we have only listed the details relevant to the scope of the report. After documenting the facts of the incident, we analysed each case on the following dimensions:

Form of Tech/AI Use: This describes the technical manipulation that the Trustline team identified in the content.

Sphere of Attack: Was the attack carried out in public or private digital or physical space?

Relationship between the Perpetrator and Victim: How, if at all, were the perpetrator and victim known to each other?

Platforms to Which the Content was Reported

Clauses Under Which the Content was Reported to the Platform

Form of Support the Victim Sought. The Trustline provides the following kinds of support:

1. Technical Support,
   such as how to make profile private, blocking perpetrators
2. Takedown of content
3. Mental health counseling
4. Legal advice
5. Social/family support

While not a perfect indicator, the level of support sought provides some indication of the seriousness of harm perceived by the victim. Technical support and content takedown are often the first line of action. The Trustline has observed that in the most serious cases there is always a need for mental health counseling. Requests for mental health counseling often signal not only the severity of psychological impact but also the need for long-term intervention. Similarly, recourse to social or family support frequently reflects the degree of isolation or lack of informal safety nets in a victim's immediate environment. Patterns in technical support requests can serve as indirect markers of a victim's digital literacy, their fluency with the platform where harm is occurring or the technical complexity of the abuse itself. In contrast, seeking legal advice suggests both a willingness to engage formal systems and the perceived necessity of institutional redress.

These categories are not discrete but rather interdependent, reflecting overlapping vulnerabilities, needs and inclinations that shape victims' pathways to support.

Other details: Here we document other details from the case history that are relevant to understand the interaction of AI with online harassment.

Note on Case Histories:

Case files are maintained in line with best practices in social work and counselling. It centers the survivor's testimony and expectations. Details are recorded primarily to facilitate effective survivor support. Precisely categorizing digital manipulation may not always be essential to understanding the violation or providing counseling and support. While counselors strive to document instances or mentions of digital manipulation and AI-generated content, occasional inaccuracies or omissions can occur. For example, when deepfake cases first began emerging, there was no dedicated category to record them. They were categorized as "images showcasing non-penetrative explicit content". Furthermore, if a caller reports that certain content is AI-generated, counselors do not strive to verify or assess this claim's accuracy if such confirmation is not necessary to determine the nature of the violation. Detecting AI-generated content is beyond counselors' primary responsibilities, and such cases may be labeled accordingly based solely on the client's statements.

In addition to the cases reported by survivors to the Trustline, we also describe two social media trends involving audio content that were discovered by the Trustline incidentally during routine scrolling. While no case involving audio has been reported to the Trustline by survivors, we think it is worthwhile to include these. Synthetic and AI-generated audio is a growing cause of concern since audio manipulations are especially difficult to detect.

In the discussion section, we build on the case studies as comparative data points to understand the impact of AI on online harassment.

Online Harassment Suspected to Contain AI

Case Study 01:
Deepfake Sextortion via Loan App Scam

Age: 31, **Gender:** Female, **Location:** Assam, India
Timeline: April 2025 – Case Closed in July 2025.

In April 2025, a 31-year-old woman from Assam contacted Meri Trustline after facing sextortion, doxxing and unsolicited harassment. The abuse began after she downloaded a loan app called ScoreClimb, where she uploaded her PAN card and photograph. Without her requesting a loan, ₹1800 was deposited into her account. Soon after, the app's operators began demanding repayment.

The survivor stated that she repaid the original loan amount along with significant interest, but the payment demands continued. When she refused to continue with the payments, her uploaded photograph was digitally altered using a nudify app and placed on pornographic imagery. Her phone number was also embedded on the image. The offenders threatened to leak this morphed content across social media platforms and to her personal contacts.

When the survivor refused to yield to the pressure, the image was circulated via WhatsApp. This resulted in a barrage of sexually explicit calls and messages from unknown individuals. Close contacts also received this message.

Prior to reaching out to the Trustline, the survivor had reported the case to the cybercrime helpline and website.[1] The Trustline advised her to block and report the offending numbers and escalated the case via WhatsApp's community reporting channels.

On 21st April, WhatsApp confirmed that action had been taken. A follow-up check revealed that the survivor had not received further threats. However, she disengaged from communication with the Trustline despite multiple follow-ups.

[1] Cybercrime Helpline: 1930; reporting portal: www.cybercrime.gov.in.

Analysis: Case Study 01

Form of AI Use: Nudify app used to generate non-consensual sexually explicit synthetic imagery.

Sphere of Attack: The abuse began in a private setting (via the loan app) but escalated with threats of exposure and ultimately circulation in the public domain. The harassment culminated in targeted dissemination to her close contacts as well as wider spread through WhatsApp groups that also contained persons unknown to the survivor.

Relationship Between Perpetrator and Victim:
Stranger or unknown entity operating through a predatory app.

Platforms to Which Content Was Reported: WhatsApp

Clauses Under Which Content Was Reported:

○ Harm to WhatsApp or our users

○ Legal and acceptable use

Support Sought by the Victim:

○ For the harassment to stop

○ For the offending accounts to be actioned

Other Salient Details:

○ Example of loan app sextortion

○ Contact list of the survivor exploited for targeted harassment and pressure.

○ After the immediate threat was resolved, the survivor disengaged from further contact with the Trustline. This is a noted pattern in sextortion cases where survivors often step back once the harassment subsides. In this instance, the Trustline had already provided available interventions and its scope of support had been exhausted.

○ WhatsApp took action, but it was insufficient since the content had already spread.

○ The survivor shared that even though some people may have suspected that the image was fake, she still felt deeply shamed and socially marked, as though she had been "involved in something dirty" and felt guilty that she was involved in this.

Case Study 02:
Deepfake Threat via Snapchat

Gender: Female. **Age**: 15. **Location**: Purnia, Bihar, India
Timeline: November 2023 – Case closed on 5th April 2023.

In November 2023, a 15-year-old girl reached out to Meri Trustline after being referred by a concerned adult. She had added a stranger to her Snapchat account. The conversation initially appeared casual but quickly turned coercive. The stranger began demanding nude images. When she refused, he threatened to use software to morph her existing photos and create synthetic nude images. He further threatened to upload these doctored images on pornographic websites.

The girl was highly distressed by the threat, especially the fear of being framed in explicit content she had never created. She expressed concern about the impact such an image could have on her life and reputation. Although counseling support was offered, she declined at the time. The Trustline team filed a cybercrime complaint on her behalf and also reported the incident to Snapchat. The account was actioned.

A follow-up was conducted on November 15, 2023. She informed the team that the perpetrator had stopped messaging her and that she was feeling better, hence requested that the case be closed.

Analysis: Case Study 02

Spectrum of AI Use: Threat of using nudify or deepfake software to morph regular images into nude photographs. AI was not actually used but the threat of it was central to the coercion.

Sphere of Attack: Private digital space (Snapchat chat).
Threats involved public exposure.

Relationship Between Perpetrator and Victim: Stranger met online.

Platforms to Which Content Was Reported: Snapchat

Clause Under Which Content Was Reported: Snapchat community guidelines: sexual content and threats, violence and harms

Support Sought by the Victim: To remove the offending account

Other Salient Details from Counselor Notes:

- Represents an emerging trend where the threat of deepfake abuse alone creates coercive power, even without actual images.
- The child victim was particularly constrained by her family environment, which limited how openly she could speak about the abuse.
- The Trustline enabled the victim to file a complaint on the government cybercrime reporting website. The risk was mitigated, and the victim received counselling in time

Case Study 03:
Up-Down Troll

Gender: Female. **Age**: 16. Region: Haveri, Karnataka

Timeline: Up-Down Troll Case: October 2023 – January 2024.

Reel that Denigrated her: Accounts were active from 24 June 2024 – 15 July 2024

Impersonation: Accounts were active from 15 November 2024 – 27 November 2024

Case closed on 20 January 2025

On October 27, 2023, a 16-year-old woman from Haveri, Karnataka, contacted Meri Trustline after discovering that two of her Instagram reels had been edited and re-uploaded by a troll account named "uk_belagavi_trooll_". The videos were manipulated using AI to depict her as nude. On closer inspection, she realized this was not an isolated case. Many similar reels featuring young women and girls, including the survivors' friends, had been altered and shared by a network of troll accounts.

These accounts followed a common visual pattern, marked by a specific logo and watermark, and used a distinct format known as "Up-Down Trolls". The manipulation involved inserting an AI-generated nude image into the reel: a dynamic logo would move across the screen and overlap with a static one, at which point a brief flash, typically five to ten seconds in, would show a morphed nude image of the creator.

Instagram initially responded to the complaint by saying the content did not violate its community standards. However, persistent follow-ups and escalations by the Trustline led to the eventual removal of the reels and takedown of the original account. In early November of 2024, the survivor reported similar AI-morphed content being reposted by other troll accounts bearing near-identical usernames and branding, indicating a coordinated ecosystem. These, too, were tracked and reported by the Trustline team.

In July 2024, a second form of harm was reported by the client. A troll account posted another reel targeting her and a friend. This time, the abuse took the form of bullying and harassment. The girls' reel was again manipulated and turned into a meme. The original reel contained an assertion of the women's autonomy over choosing her romantic partner. It was spliced with footage of a boy who in a very crude manner asks her to "Shut up and go home to wash vessels". After sustained effort, the account was taken down. Here again, the initial complaint was considered non-violative but after escalating and giving context that the survivor was a previous victim of online harassment, the content was taken down.

In November 2024, a third incident emerged. The survivor discovered an impersonation account that had been created using her name and photos. This fake profile contacted her followers and asked them for money, pretending to be in crisis. She was supported in submitting a complaint via Instagram's Grievance Officer form. After some delay, the impersonating account was successfully taken down.

Analysis: Case Study 03

Spectrum of AI Use: Clothoff app used to generate synthetic nudity inserted into the victim's original Instagram reels. These were designed using a stylized "Up-Down Troll" format, where logos and motion graphics masked a brief flash of a morphed nude image.

Sphere of Attack: All three attacks were carried out in a public digital space (Instagram).

Relationship Between Perpetrator and Victim: There was no direct personal relationship. The perpetrators were anonymous troll account operators, likely part of a coordinated network targeting local-language women influencers.

Platforms to Which Content Was Reported: Instagram

Clause Under Which Content Was Reported: "We have zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others" as well as "We remove ... content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages." "Bullying and Harassment" for the reels using abusive overlays. "Impersonation and Fraud" for the fake account soliciting money from followers.

Support Sought by the Victim:

○ Removal of Accounts

○ Legal Advice for Up Down Troll Case

Other Salient Points:

○ The survivor linked this series of coordinated harms to her growing visibility as a local-language influencer. With rising reach came increasing vulnerability to tech-facilitated violence.

○ The "Up-Down Troll" format's stylized covert delivery of synthetic nudity and made detection and moderation harder.

○ The abuse appeared systematic and targeted, suggesting a regional troll ecosystem focused on Kannada-speaking women.

○ Initial platform response was delayed and inadequate, it also did not consider the full context of the survivors' circumstances and the format in which the reels were manipulated. Takedowns only took place after sustained escalation.

- ○ The Trustline enabled the survivor to file a complaint on the government cybercrime reporting website.
- ○ The Trustline escalated intelligence from this attack to the concerned police departments and platforms.

Case Study 04:
Deepfake Bullying of Boy in an Online Group

Gender: Male. **Age:** 17. **Region:** Katihar, Bihar

Timeline: November 2023 – Case Closed on 22 February 2024.

The survivor, a 17-year-old student from Katihar, Bihar, was referred to the Trustline by a peer. He had been a member of a loosely organized Instagram group titled Tharkis, comprising random members. After a minor disagreement with the group's administrator, which involved the use of abusive language, the conflict escalated dramatically.

Following the dispute, the group admin used AI deepfake tools to create manipulated nudes from the survivors' publicly available Instagram photos. These edited images, which showed the survivor's face on a bikini-clad woman's body, were then circulated within the same Instagram group as a form of revenge and humiliation.

The survivor was unaware of the identity of the perpetrator beyond the Instagram handle.

Analysis: Case Study 04

Spectrum of AI Use: Very rudimentary AI Deepfake tool used to generate non-consensual nude images from regular profile photos.

Sphere of Attack: Began in a closed peer group (Instagram DM), with the threat of reputation damage through further circulation.

Relationship to Perpetrator: Known online peer within the group- the admin of the group escalated a personal argument into targeted digital abuse.

Platforms to Which Content Was Reported: Instagram

Reported Under: "We have zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others. We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages."

Support Sought by the Victim: Removal of offending account

Other Salient Points:

○ Abuse was retaliatory, peer-based and involved male-on-male AI sexual violence.

○ The harm lay not in how realistic the image was, but in how deliberately ugly it was made to appear. The deepfake wasn't designed to convince. It was designed to shame. It used distortion to render the survivor mockable.

○ The Trustline enabled the survivor to file a complaint on the government cybercrime reporting website.

Cases of Non-AI-based Online Harassment

Case Study 1:
Pause Challenge

Age: 24, **Gender**: Female, **Location**: Chengalpattu, Tamil Nadu
Timeline: July 2024 to August 2024 Case Closed on 24 February 2025

On 4 July 2024, a 24-year-old woman from Chengalpattu, Tamil Nadu, contacted Meri Trustline after finding the helpline on StopNCII.org's partner list. She reported that her nude photographs had been leaked on Instagram as part of the 'Pause Challenge', a trend where reels feature a non-explicit cover image, but flash nude content timed to the beat drop in the accompanying audio. If paused at the right frame, the nude image becomes visible. A similar incident had occurred in January 2023 on X, formerly known as Twitter. The survivor thus feared that despite a takedown on Instagram, the content could be reuploaded or resurface elsewhere.

The survivor suspected that the source of the leak was her Google Photos account. Her phone had been stolen in the past, but she also recalled logging into her Google account on friends' devices. This raised the possibility that someone in her known circle may have accessed and leaked her files. She considered the latter scenario more likely.

The Trustline found her photos to be hosted on two pornographic mirror sites: dropmms.net and dropmms.com. Both sites also featured links to cloud storage folders containing zipped archives of her private images.

Analysis: Case Study 01

Spectrum of Tech Use: Real private images were non-consensually extracted and re-edited into Pause Challenge reels using frame-level manipulation. Content was disseminated through pornographic platforms (DropMMS) and third-party cloud storage links.

Sphere of Attack: Public exposure on Instagram and X, followed by persistent reappearances on indexed pornographic sites and search engines. The survivor faced secondary trauma through visibility on Google.

Relationship to Perpetrator: Likely known circle with access to synced Google Photos. No perpetrator was conclusively identified but suspicion remained on individuals the survivor had previously trusted.

Platforms to Which Content Was Reported: Google (Search), DropMMS, Instagram

Reported Under:

Google: Personal Content & Product Policy

DropMMS: DMCA

Instagram: This goes against Instagram's Community Guidelines that state:

- ○ "We have zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others."
- ○ "We remove ... content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages."
- ○ "We don't allow nudity on Instagram. This includes photos, videos and some digitally created content that show sexual intercourse, genitals, and close-ups of fully nude buttocks. It also includes some photos of female nipples."
- ○ Potential CSAM-pattern reporting due to nature of site and circulation method.

Support Sought by the Victim:

- ○ To remove the content from the platforms reported.
- ○ To search for other copies of the content online and ensure they were removed as well.

Other Salient Points:

○ Pause Challenge abuse hinges not on creating new content but on inserting stolen intimate material into fleeting frames, making detection harder.

○ The real images have a potential for going viral in the porn/NCII online ecosystem/networks.

○ Search engine visibility was a major vector of harm. The Google takedown was pivotal to restoring a sense of safety.

○ The Trustline escalated intelligence from this attack to the concerned police departments and platforms.

Case Study 2: **Use of NCII in Interpersonal Violence**

Age: 21, **Gender**: Female, **Location**: Unnao, Uttar Pradesh

Timeline: July 2024 – June 2025 ; Case Closed on 18 July 2025

On 2nd July 2024, a 21-year-old woman from Uttar Pradesh contacted Meri Trustline after discovering that her intimate videos and call recordings had been uploaded to Instagram without her consent. The content featured her nude during a private video call, with her face clearly visible. These recordings were allegedly captured without her knowledge by a former partner, a young man from a nearby village whom she had known since school.

The Instagram account also added sexually explicit captions, tagged her by name and falsely linked her to other users. The perpetrator obscured his own face using emojis and stickers but kept her identity fully visible. He later messaged her on WhatsApp, claiming to have more such videos and threatening to leak them further unless she complied with his demands.

Over the next three months, more than 23 impersonation accounts were created on Instagram, repeatedly uploading the same content. Some of these accounts included her phone number in the bio or captions, leading to a flood of calls and harassment from strangers. At one point, her own previously used Instagram account was hacked and repurposed to post more material.

Over time, the perpetrator began uploading slightly blurred versions of the content, likely an attempt to evade Instagram's detection systems. By connecting this content to the previous uploads, the Trustline was able to have it removed.

Analysis: Case Study 02

Spectrum of Tech Use:

- Intimate content recorded without consent during private calls, shared with overlays to obscure perpetrator identity.
- Account hacking and impersonation.

Sphere of Attack:

- Began as private abuse (secret recording during relationship) but eventual public exposure via Instagram, with intimate content and personal information shared on the platform.
- Private intimidation via WhatsApp.

Relationship to Perpetrator: Known person. A former partner with access to private content during the relationship.

Reported Under:

- "We have zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others."
- "We remove ... content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages."
- "We don't allow nudity on Instagram. This includes photos, videos, and some digitally created content that show sexual intercourse, genitals, and close-ups of fully nude buttocks. It also includes some photos of female nipples."

Support Sought by the Victim:

- Account takedown.
- Legal advice on how to file a police report.
- Deletion of the images at source.
- Prevention of reupload.

Other Salient Feature:

- Despite repeated reports and platform actions, content reappeared frequently.
- Shows gaps in platform moderation when abusers adapt content to avoid takedown (for e.g., using blurring techniques).
- Survivor's resilience and family support were key to navigating long-term redressal.
- Case also flagged the need for protocols on handling impersonator calls to helplines.

Content Discovered by Rati through Social Media Scanning

Case Study 1:

Revenge-Based Audio Denigration and Doxxing via Instagram Reels

Gender: Mixed victims (primarily women). **Location**: Unknown, India
Timeline: 13 August 2024 – 18 September 2024

In mid-2024, the Trustline identified a cluster of Instagram accounts using locally produced audio tracks to target women through sexually denigrating content. The audios, often presented as poetry with a simple rhyme scheme, carried misogynistic narratives about betrayal, women being "gold diggers" or ex-girlfriends labeled as sexually promiscuous and compared to sex workers.

These audio clips were overlaid on women's images, in some cases accompanied by identifiable details such as names or contact information. The voiceovers were vulgar, with explicit references to sexual acts and body parts, designed to induce social humiliation and reputational harm.

The Trustline escalated the content to Instagram, resulting in the removal of nineteen reels from multiple accounts on September 16, 2024 for violating community guidelines. However, the music producers responsible for creating the abusive audio tracks remain active on the platform.

Analysis: Case Study 01

Spectrum of Tech Use: Non-AI audio-based abuse relying on Instagram's reel and music features.

Sphere of Attack: Entirely public facing, designed for mass exposure and social humiliation.

Relationship to Perpetrator: Survivors had no known personal relationship with the producer of the audio tracks.However, some of the reels where photographs were layered with songs indicate patterns common in relationship abuse or cyberstalking.

Reported Under:

○ Bullying and Harassment

○ Adult Sexual Exploitation

Impact on Victims: Unknown

Other Salient Points:

○ Highlights how cultural audio forms such as poetry can be co-opted into sexualized harassment.

○ A network of influencers created audio that could easily be weaponized and made into targeted abuse

○ Misogynistic, revenge-themed tracks weaponized against women's images and identities.

○ Incorporated doxxing in several cases, combining reputational harm with public shaming.

○ The Trustline filed a complaint on the government cybercrime reporting website.

○ The Trustline escalated intelligence from this attack to the concerned platforms.

Case 2:
Audio-Based Denigration and Sexualized Manipulation on Instagram

Gender: Mixed victims. **Location**: Unknown, India
Case Active: Since 4 June 2025 (ongoing)

In June 2025, the Trustline identified a growing cluster of Instagram reels circulating sexually explicit and denigrating content using localized audio tracks. The reels featured voiceovers and song clips containing explicit references to sexual acts and private body parts, delivered in vulgar language and regional dialects. These audios were being used across multiple accounts, many of which had amassed significant reach.

The central content source was traced back to a main Instagram account with nearly 300,000 followers, which also maintained a YouTube channel that distributed similar material. The reels were often designed as song visualizers with AI-generated thumbnails. The voices used in the tracks resembled those of popular Bollywood singers, which were assessed by the Trustline counselors as likely AI-generated or synthetically altered.

Among the reels, one included Child Sexual Abuse Material (CSAM)—a video showing a child with visible genitals overlaid with one of these abusive sounds. Another reel manipulated a couple's private video to falsely present them as siblings involved in sexual acts. In several others, women's photos and videos were paired with sexually denigrating audio, with captions assigning fake prices like "₹300" or "₹500", implying prostitution. These were not isolated incidents but part of a larger network of sexualized harassment, built around a shared sound library and repeated patterns of abuse.

The Trustline escalated these reels to Instagram through its reporting process, flagged the CSAM content to the cybercrime portal and filed broader alerts to initiate takedown of the audio-based harassment ecosystem. This led to 21 links being actioned, including the producer of the audio files. For the producer of the audio files, Instagram cited "Child Sexual Exploitation" for the action.

Spectrum of Tech Use: AI-generated visuals and likely AI-cloned voices used in sexualized song reels.Vulgar tracks were layered over manipulated or decontextualized imagery.

Sphere of Attack:

○ Entirely public: the reels were hosted on public Instagram accounts.

○ Mass circulation enabled by viral format and platform-native features (reels, music library).

Relationship to Perpetrator: Survivors had no known personal relationship with the producer of the audio tracks. However, some of the reels where photographs were layered with songs indicate patterns common in relationship abuse or cyberstalking.

Reported Under:

○ Child Sexual Exploitation

○ Adult sexual imagery

○ Bullying and harassment

Impact on Victims: Unknown

Other Salient Points:

○ Illustrates how AI audio tools and generative visual design are being used not to mimic the victim, but to build scalable, sexually explicit abuse ecosystems.

○ Shows the difficulty of platform moderation to detect audio-led IBSA, even when CSAM is involved.

○ The Trustline filed a complaint on the government cybercrime reporting website.

○ The Trustline escalated intelligence from this attack to the concerned platforms.

Analysis

Contrasting Harms

To understand how AI is intervening in an already complex terrain of online harassment, we attempted to analyse how the harm differs between incidents involving AI-generated content and those without it. To do this, we used five of the nine dimensions from Scheurmann, et al's framework[39] on online harm that are relevant for a survivor-centered analysis. Unlike the original paper, we didn't attempt to rank the harm from each dimension on a scale. Instead, we provided a qualitative description of how the factors played out in the cases involving AI versus others.[40] We added a final dimension of analysis on the perceived harm by the victim, as understood from the support they asked from the Trustline.

Perceived Intent to Harm the Victim:

Not AI-Generated: Most cases of non-AI-based online harassment are of NCII. The victim and perpetrator often occupy consensual intimate spaces. The intent to harm is targeted towards the specific individual, often as an expression of revenge.

AI-Generated: In the majority of the cases involving AI-generated content the perpetrator and target were not close in the physical world. In some, but not all, cases the perpetrators and victims became acquainted online before the harassment started. Many of the cases involving AI-based sexual abuse, appear to be a part of a larger online sexual or financial harassment racket. While the intent is to harm an individual, it isn't contingent or complete in harming just that specific individual.

[39] Scheuerman et al.'s (2021)

[40] The dimension of perspective compares the harm perceived by the victim and the perpetrator, with the former perceiving it as higher and the latter perceiving it as lower. This remains true, regardless of the form of act. Analysing this is not meaningful for contrasting AI and non-AI based IBSA.

◆ Agency of Victim in the Cause or Resolution of Harassment:

Not AI-Generated: In most cases, the victim had at some point consensually shared the content. Their own role in the incident is perceived to be higher by the victim. Consequently, they approach support systems with shame and self-blame, perceiving their own consent to share intimate content as a factor. In some cases, since the victim and perpetrator are known to each other, the victims also negotiated with the perpetrator to take down the content.

AI-Generated: In nearly all cases, the AI-generated content relies on photographs the victim has posted on their public profiles—the only 'agency' the victims had in causing the attack. This is to say that the victims had low to no agency in the content that contributed to their harassment.

◆ Urgency:

Not AI-Generated: The victims expressed urgency and asked for immediate removal of the content.

AI-Generated: The victims expressed urgency in asking for immediate removal of the content. They were more likely to express anger and confusion, emphasizing their lack of involvement and demanding immediate action. The Trustline doesn't perceive the urgency to be any different.

◆ Vulnerability:

The Trusline only collects gender and age as the demographic information on the callers. This is done to preserve the privacy of the callers. Thus, it is hard to comment on broader socio-economic vulnerabilities. On gender, the breakdown is as follows:

Not AI-Generated: Roughly, 75% of the cases targeted victims who identified as female.

AI-Generated: 23 out of the 25 incidents (92%) of the cases involving AI targeted victims who identified as female.

◆ Medium:

Not AI-Generated: These involve videos, images and audio.

AI-Generated: The cases reported to the Trustline so far, have only involved photographs. One of the cases discovered by social media scanning involved the use of AI generated audio.

◆ Sphere:

Not AI-Generated: The not-AI-generated cases lie on a spectrum of extremely private to more public forms of harassment. Some cases are strongly private where NCII is shared with the victim and/or their family members. In some cases, the harassment is on and through public social media platforms. Often, the private forms of harassment gave way to the public forms.

AI-Generated: Since none of the perpetrators and victims were closely acquainted before the incident, the access to the victim's personal contact was restricted. The one exception was the loan app scam from Assam, where contact details were collected as part of the sign-up process. The AI-generated content was circulated or threatened to be circulated in groups or on public social media accounts. Compared to other IBSA, attacks involving AI-generated content appear to be more public.

◆ Support Sought by the Victims:

Across cases, technical assistance and content removal consistently emerge as the first line of action, underscoring victims' immediate imperative to reassert control over their compromised digital environments.

Not AI-Generated: Victims whose images were shared non-consensually reported a heightened sense of betrayal because the abuse emerged from a relationship that once carried trust or emotional connection. These prior dynamics shaped not only the perceived severity of danger but also the type of redressal sought. They worried that disclosure would lead to significant disruptions and inviting scrutiny into all aspects of their life. In the long term, survivors of NCII have requested psycho-social support.

AI-Generated: Victims frequently described a sense of disorientation and confusion, both about the synthetic nature of the images and the impersonal manner of the attack. Even when the manipulated imagery appears obviously artificial, victims expressed fear around disclosure. They worried that explaining the circumstances of the image creation may invite scrutiny of their online behaviour.

None of the victims in cases involving AI asked for long-term support such as psycho-social or legal counseling.

Reporting AIGC and Online Harassment to Platforms:

There are several pathways for reporting AI-generated and image-based abusive content online to platforms. Prominent social media platforms often establish trusted partner channels, where certain organisations are given the privilege to request escalated removal of certain egregious and harmful content, and this content is generally actioned upon within a short duration of time. For example, on Meta (Instagram, Facebook, Threads), cases are escalated through the Global Trusted Partner Program. On YouTube, cases can be reported via Trusted Flagger access. At Aylo as well, content can be escalated through established flagging partner channels. As per Indian law, platforms such as Snapchat, Telegram, WhatsApp and X are required to have a Grievance Officer, whose details must be publicly posted on their website, and escalations can be sent to the Officer. Snapchat also provides an escalation channel beyond the standard reporting form. The terms on X, as implemented by their moderation teams, favour the uploader of content, and ignore the right of the individual targeted. On non-significant intermediaries abuse reporting is done via the available in-platform reporting forms, often with limited or no response.

◆ Use of DMCA:

In cases where the victim's likeness has been used, reporting the content to the DMCA for copyright infringement has proven to be more effective than framing and reporting the abuse under the category of gender-based harm. For instance, when dealing with pornographic sites or resistant platforms, Rati Foundation found that framing NCII as a copyright violation under DMCA often yielded faster results.

Further, the DMCA has been used where platforms do not have functional abuse reporting mechanisms, which often includes pornographic websites. Thus far, it has not been used for reporting AI-generated content.

◆ Seeking Legal Recourse:

No FIR has yet been filed in any of the cases involving AI-generated content. Reports are filed largely on platforms and the cybercrime portal for content takedown. Authorities have recommended that reports be filed on platforms and the portal before approaching a police station. For the cases outlined above, Rati has not found any deficiency of avenues for legal recourse in law. However, procedural challenges have been encountered when legal action was pursued, some of which include:

- Lack of clear protocol at police stations for handling digital evidence in a manner that preserves its integrity and is sensitive to the victim's experience.

- Insufficient guidance to victims on handling and preserving evidence relevant to the incident.

- Confiscation of digital devices as evidence, often without proper explanation or consent.

- Delays at the Forensic Science Laboratory (FSL), which can hold -down investigations.

- The technical nature of FSL reports, which are difficult for non-technical experts (for e.g., police personnel, judges and prosecutors) to interpret and to provide no clarity on how to read or report findings once they are received.

- Challenges in presenting and validating digital evidence in court, especially as the Investigating Officer (IO) is expected to explain the entire chain of custody and collection process.

Even as international regulations come into force, the evidence from the cases cited suggests that, at this point, India needs support in capacity-building and ease of enforcement more than new regulations to tackle specific online offences facilitated by AI.

Discussion

The cases handled by the Trustline convey a couple of different things about the use of AI-generated content for harassment of private individuals. First, synthetic content is used primarily when real intimate content is unavailable. The Trustline has, so far, not encountered a case where AI-generated content was used in intimate partner violence. The support sought by victims also indicates that the impact of AI-enabled harassment is perceived to be comparatively less severe as opposed to content involving non-consensual intimate imagery. Yet, as seen in the case of the teenager from Bihar who was blackmailed on Snapchat, the threat of generating a deepfake is emerging as a common tactic for harassment. The analysis also foregrounds how the threat of deepfake creation is increasingly leveraged, not only by online stalkers and former partners but also by actors such as predatory loan apps to affect victims' sense of vulnerability.

This begets the question of why the threat of generating AI content succeeds at all? In India, there is intense shame associated with nudity and sexual acts. Harassment has relied on the believability of the survivor being involved in sexual acts portrayed in the content. This was also the attempt with older technologies for digital manipulation. But AI makes the creation of realistic-looking content much easier. The authenticity of the content, however, is not a consideration for the Trustline in providing support to the survivors or for reporting the content to platforms and law enforcement.[41] More recently, victims have begun to report cases to the Trustline claiming that the content being circulated online is AI-generated, even when it is not. Claiming that media is AI-generated to refute real content has been described as the Liar's Dividend. But in contexts of restrictive social norms, this dividend can be advantageous as it also enables victims to seek support for cases of non-consensual intimate imagery. While in some cases, the counselors at the Trustline can gauge if the content is real and not manipulated, this is incidental to their primary task of supporting the victims.

[41] In cases of public interest, such as the R.G. Kar Medical College and Hospital rape-murder case, the Trustline team debunked a video circulating that showed AI-generated likeness of the victim. Some depicted fabricated crime scenes, while others portrayed the girl's life using her face and showing her in settings such as wearing a doctor's coat.

In many cases of harassment, such as the Up-Down trolling in Karnataka; the bullying of the boy from Bihar in an online group; or the AI-generated audio tracks, AI is not deployed to generate realistic content but rather sexualized satirical content. AI is the latest trick in meme cultures where affect rather than factuality is the goal. When targeted towards private individuals, the perpetrators may piggyback on an ongoing meme trend to create content about the victim. They may even aim to generate a meme trend by creating similar content about a number of private individuals. Often, the victims are not tagged, and virality emerges from the aesthetic and implied humour rather than viewers' association with the individuals in the content, as might happen with AIGC targeted towards public figures.

While the degree of harm may vary with AI deployed for realistic versus satirical content, the individuals targeted experience equal fear and anxiety. It is this perspective from which the content needs to be primarily addressed. The Trustline reports content on the basis of violations to an individual's safety, not the status of its authenticity. In reporting AI-generated content to platforms, the Trustline team has found that the greatest challenge lies in the overall reporting architecture provided by a platform. On platforms where reporting all content is difficult, such as X, reporting AI-generated content is also difficult. On the other hand, platforms with more expansive definitions of harmful content are more likely to address AI-generated content. Meme-like content, which has been a grey zone in platform policies, remains a grey zone even with AI-generated content. The Trustline is a trusted flagger for many platforms and is able to make the case for the removal of content that is on the boundaries of platform policies. But victims are not accorded the same privilege. Many victims approach the Trustline after having tried to, and failed at, reporting the content themselves. Similarly, audio content, which is less addressed in platform policies, is also a gap in addressing AI-generated sexualized content.

In cases where there is no identified victim—as is the case with content that the Trustline discovers through social media scanning—it becomes harder to report content. This is also true of synthetic images used for creating sexualizing or misogynistic narratives. One of the abiding characteristics of AI-generated abuse is its tendency to multiply. It is created easily, shared widely and tends to resurface repeatedly. This pattern of 'content recidivism', where similar or identical content reappears across accounts and over time, is often only visible at the platform level. Understanding these distribution dynamics, and the broader societal-level impact on AIGC, will require far greater transparency and data access from platforms themselves.

Technical guardrails to AIGC such as watermarking of AI outputs, limitations on sexualized prompts on AI bots and improved detection of synthetic media, while promising, are concentrated on a handful of "significant" platforms where visibility, media pressure and risk of reputational damage are high. There is a wider production and distribution network of lesser-known apps that get by without having or enforcing any community guidelines or terms of use protecting their users. The fragmented response mirrors the historical trajectory of online sexual abuse. When takedown of content on the bigger platforms becomes strict, circulation of harmful content often migrates to under-regulated spaces.

With law enforcement and judiciary, the challenge is not the lack of legal provisions but rather of capacity. Judicial decisions may be taken without an adequate understanding of online and digital interactions, increasing the burden on the victims. Perceptions of what constitutes a "serious" offence significantly influence the decision to report the incident to the police. Many victims express fear, confusion and hesitation when engaging with the police. On the ground, the absence of standardized protocols for handling online harm presents further challenges. If the perpetrators' identity is unclear or they cannot be apprehended immediately, police often record the complaint without registering a First Information Report (FIR), unless the survivor is persistent or escalates the matter to senior officers. In cases involving online threats or sextortion linked to financial fraud, police may redirect victims to the cybercrime reporting portal, a response that many find inadequate. Despite the availability of online reporting mechanisms, survivors are still required to visit police stations in person to initiate any tangible legal action. This requirement not only delays intervention but also imposes additional emotional and logistical burdens, particularly on already vulnerable individuals. There is also a lack of consistent procedures for collecting and safeguarding evidence. Victim-sensitive protocols are frequently absent, leaving survivors uncertain about how their data and content will be handled.

In some instances, especially where abusive content has been circulated to the victim's personal contacts, police seize the devices of all parties involved. This not only disrupts everyday life but also imposes significant financial strain as devices are often returned only after long delays and in unusable condition. In cases involving both physical and digital abuse, police tend to prioritize charges under physical assault provisions, possibly due to greater familiarity with those legal frameworks. Courts frequently lack the technical capacity and procedural clarity required to assess and admit electronic evidence. Delays in forensic examination further obstruct access to timely justice.

As discussed above, there are currently no specific legal provisions for AI-generated abuse in India, though such harm is comprehensively addressed under existing laws. The real challenges lie in implementation. Any future regulatory approach should begin by identifying where the gaps lie and how proposed measures will meaningfully serve the victim. Simply introducing a new law may risk straining existing capacities.

Key barriers likely to be exacerbated by the rise of AI-generated content include difficulties in establishing harm, limited forensic preparedness and a lack of procedural clarity. There is also a need to build the capacity of personnel across justice and enforcement systems to recognize and respond to manipulated content, in ways that are scientific, sensitive and clear of victim-blaming narratives.

With AIGC, as with other forms of online harassment, the harm perceived by the victim is correlated with the power disparity between the perpetrator and the target. Celebrity-focused material often contains elements of sexual fantasy and places public figures in scenarios they would never consent to or produce themselves. While this reflects entrenched misogyny and objectification, celebrities often retain a relatively favorable power equation with both the producers and consumers of such content. Their public status affords a degree of resilience and mediated control over reputation. In contrast, when private individuals are targeted, AI-generated sexual content functions primarily as a tool of denigration and erasure. Its purpose or its consequence is to strip individuals of agency, deplatform them from their social and professional spaces and frame their identities through coercive digital narratives.

With this backdrop, it is important that the safety messaging around online safety not be around restricting or limiting online expression. Even if the victims show the resilience to overcome their experience of harassment and continue with their digital lives, they may be pressured by parents and peers to do so. Many of the younger respondents who approached the Trustline were hesitant to tell their families about their experience of harassment out of fear of being asked to reduce or stop any online engagement.

The long-term battle remains to destigmatize victims of NCII and shift norms so that people support rather than blame the victims of online harassment. In absence of this long-term change, addressing specific technological shifts will remain piecemeal attempts and limited in their effectiveness to support survivors.

Are You a Survivor of Sexual Abuse or Online Harassment?

Guidance on Reporting AI-Generated and Other Forms of Online Sexual Abuse:

1. Reportable Incidents

Individuals may report a wide range of online harms, including digitally manipulated content such as AI-generated sexual imagery (deepfakes) or other forms of altered images, impersonation, threats, harassment and the non-consensual circulation of intimate content. Intimate imagery need not be explicit to cause significant harm. Notably, even the threat of such content without actual dissemination may warrant redress.

2. Documentation and Evidence

Before taking any action such as blocking, reporting or deleting content, it is crucial to collect and preserve:

- Hyperlinks to the offending material
- Usernames or account identifiers
- Timestamps and dates
- Screenshots of messages, images or posts

Some form of verifiable evidence is essential for escalation. Lack of documentation remains a major barrier to effective platform or legal action.

3. Available Reporting Mechanisms

- In-platform reporting tools are a first step, though they often yield inconsistent outcomes.
- Under India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, major platforms are required to appoint a Grievance Officer – a designated contact responsible for addressing user complaints related to harmful or unlawful content. Platforms must acknowledge complaints within 24 hours and resolve them within 15 days of receipt.

- This applies to widely used services such as Instagram, Facebook, Telegram, Snapchat, Tinder, Reddit, WhatsApp, among others.
- Some Grievance Officer Contact Links:
  - Instagram: https://help.instagram.com/contact/779201836048501
  - Facebook: https://www.facebook.com/help/contact/278770247037228
  - YouTube: https://support.google.com/youtube/answer/10728153
  - X: https://help.twitter.com/en/forms/report-to-grievance-officer-india
  - Snapchat: grievance-officer-in@snap.com
  - WhatsApp: https://www.whatsapp.com/contact/forms/1534459096974129/
  - Telegram: grievance-in@telegram.org
- If the Grievance Officer fails to respond within the stipulated timeframe or provides an unsatisfactory response, users may escalate the matter to the Grievance Appellate Committee (GAC) at gac.gov.in, which is expected to address appeals within 30 days.
- Users may also report incidents to the police through the national cybercrime reporting portal at cybercrime.gov.in or call the helpline at 1930.
- Alternatively, individuals may approach their local police station or cybercrime unit, depending on state jurisdiction.
- For platforms without a visible grievance mechanism, users are advised to search for a support form, contact form or abuse reporting portal or email, typically found in the Help or Terms of Service sections.

4. Additional Information for Escalation

When escalating a complaint, you may be asked to provide contextual details such as the identities of the accounts involved, the nature and timeline of the offence and relevant cultural or social context that may inform the interpretation of harm.

Complainants may also be required to submit a government-issued ID or verification details to authenticate the claim.

5. Seeking Support

For individuals who find it difficult to navigate these mechanisms independently, Meri Trustline offers free and confidential support, including:

○ Escalation to platforms or authorities for content takedown

○ Legal guidance

○ Mental health counselling

○ Social support to address familial or relational impacts of online harm

○ Technical support with online safety tools and systems

Meri Trustline Helpline: +91 6363176363

meriTrustline@ratifoundation.org

Monday to Friday, 9:00 AM to 5:00 PM

Make It Real: Mapping AI-Facilitated Gendered Harm

Tattle RATI

Date
November 3, 2025

Authors
Siddharth Pillai, Tarunima Prabhakar and Kaustubha Kalidindi

Editor
Tanima Saha

Designer
Yatharth

Acknowledgements
This report builds on the expertise of the entire teams at Rati Foundation and Tattle Civic Tech. In particular, we are grateful to Salman Mujawar, Sameer Parmar, Ritu Jain and Ruta Sawant, counselors at Rati Foundation, whose frontline work with survivors forms the basis of much of the evidence presented here. We would also like to thank Cheshta Arora & Baarish for reviewing an earlier draft of the report.