Civil Society Demands Big Tech Strengthen AI Policies to Fight Sexist and Misogynistic AI-Based Disinformation and Abuse

To Meta CEO Mark Zuckerberg, X CEO Linda Yaccarino, YouTube CEO Neal Mohan, TikTok CEO Shou Zi Chew, Snap CEO Evan Spiegel and Reddit CEO Steve Huffman,

Content generated with Artificial Intelligence (AI) is rapidly growing in ubiquity on social media platforms as it becomes cheaper and easier to create images, text, audio, and video using new generative AI tools.[1,2] The easier AI-generated content becomes to make, however, the more difficult it is to differentiate between synthetic and non-synthetic media.[3] This blurred line opens the door to a range of opportunities for bad actors and accompanying risks for users.

And while we have witnessed the scale and scope of harms related to AI-generated content on social media increase across the board, it's evident that these harms are not felt equally. Specifically, women, trans people, and nonbinary people are uniquely at risk of experiencing adverse impacts of AI-based content on social media. Research and reporting has shown that:

- Women, girls, and LGBTQ+ individuals are most likely to be targets of sexual AI-based manipulation, which is a form of sexual abuse. Specifically, at least 90% of victims of artificial nonconsensual explicit materials (Artificial NCEM) are women.[4] And according to one 2019 analysis, at least 96% of all AI deepfakes online are non-consensual sexual content.[5] Women and queer people of color are facing the brunt of this form of abuse due to compounded racial and gender biases.[6]
- Women public figures such as celebrities and journalists are more likely to be targets of sexist and political disinformation than their male counterparts, which is increasingly being spread via AI-based content and search algorithms.[7,8]
- Older individuals online – particularly older women – are the most likely to be targets of AI-powered scams and crimes.[9]
- Most researched AI systems hold a gender bias. For instance, A recent UNESCO report explored biases of three significant large language models (LLMs): OpenAI's GPT-2 and ChatGPT, and Meta's Llama 2, and found that each model showed "unequivocal evidence of prejudice against women." As an example, they associated gendered feminine names with traditional gender roles and even generated explicitly misogynistic content (eg. "the woman was thought of as a sex object and a baby machine.")[10] LLM bias also translates to video and images, as research has shown that when prompting models to generate certain images – such as an engineer or a person leading a meeting – generated results are most likely to be images of white men.[11,12]
- Women with intersecting marginalized identities are most likely to be targeted by disinformation, which is increasingly being spread via AI-based content and search algorithms.[13] For instance, according to a study from the Center for Democracy & Technology, women of color candidates were twice as likely as other candidates to be targeted with or the subject of mis- and disinformation.[14]
- Because AI systems are typically trained on data sets that conflate gender and sex, as well as designed by gender inequitable teams, such systems run the risk of 1) entirely leaving transgender, intersex, and nonbinary people out of content, 2) perpetuating harmful stereotypes against trans and nonbinary individuals, and 3) contributing to trans scapegoating.[15,16,17]

Given the inequitable, sexist harms of unregulated AI-generated content, social media companies must commit to intentionally developing clearer, more transparent, and more robust AI policies that explicitly consider risks to all people with marginalized gender identities.

This is about more than platform safety, as norms and narratives that circulate online can translate offline.

Certain platforms have already taken some important steps in the right direction. For instance, Meta requiring "AI Info" labels and YouTube requiring creators to disclose meaningfully altered content, and Snap adding a watermark to AI-generated images are all starting points to better protecting women and nonbinary users from AI-based risks.

However, these steps are not nearly enough, especially when it comes to protecting against gendered violence, targeting, and AI-based discrimination. Our communities are suffering now and deserve to participate in online environments with safety and equality truly centered.

To that end, we are calling on you to implement the below recommendations for building out stronger AI policies considering gendered risks and harms. Please note that these recommendations are exclusive to user-generated, organic content, and resultantly exclude other forms of AI-based gender bias/abuse, such as algorithmically targeted ads. As a result, they are not a comprehensive approach to addressing the full ecosystem of AI-facilitated gendered disinformation, discrimination, and bias.

In addition, these recommendations are meant to serve as building blocks from which more precise and technical solutions can be built.

Classification under existing policies

1. Explicitly name "artificial nonconsensual explicit materials (Artificial NCEM)," as prohibited content under existing hate speech, harassment, and/or misinformation policies, with clear consequences for repeat posting of such content. Types of policy categories that "artificial nonconsensual explicit materials (Artificial NCEM)," fit under include the following: hate speech, violence and incitement, abuse/harassment, bullying/harassment, harmful and dangerous content, and nudity/sexual content.
    - 1A: For Meta specifically: Change the term "derogatory sexualized photoshop" to "artificial nonconsensual explicit materials (Artificial NCEM)" (building off of the recommendations of Meta's Oversight Board)
    - 1B: For Meta specifically: Add the prohibition on "derogatory sexualized photoshop" (which should be renamed as "artificial nonconsensual explicit materials (Artificial NCEM)") into the Adult Sexual Exploitation Community Standard in addition to keeping it under Bullying and Harassment (in line with the recommendations of Meta's Oversight Board)
    - 1C: For Meta specifically: Under the policy on Manipulated Media, explicitly define the terms "edited and synthesized media," and "technical deepfakes" and clearly define the consequences per the first recommendation, which should include downranking following multiple violations.
2. Specifically define the consequence for posting artificial NCEM, which should include suspension and subsequent downranking as warning upon first violation and immediate and permanent suspension upon second violation.

Detection, Disclosure, and labeling

3. Implement a tool – developed by a neutral third party – to detect AI-generated content. Such a tool should have been audited for bias, efficacy, and accuracy as well as tested on content from the platform. If such tools are already in use, share the name and information about the AI detection tools with civil society researchers. Any such tools should be audited annually by a neutral third party other than the tool developer.
4. Require that any content generated with AI assistance that could be mistaken as unassisted by AI be disclosed as developed with AI by the user through a prompt built into the platform. If users are creating AI-generated content with an AI tool provided by the platform itself, a label should be provided automatically. Outline clear consequences for repeatedly failing to disclose content generated with AI assistance .
5. If and when an AI detection tool is in place, AI-generated, clearly label undisclosed AI-generated content as "Content detected as developed with AI. Undisclosed by the publisher." As per recommendation 4, consequences for undisclosed AI-generated content should be clear. These labels should appear as a pop-up, so as not to be missed, and include a redirect link to a resource page as described below.
6. When it comes to human review, conduct regular trainings (at least annually) on fact checking in the age of AI-based disinformation that includes context on gender biases in AI systems and the use of AI to disproportionately target and abuse women, girls, and LGBTQ+ individuals.

User flagging

7. In the platform user reporting flow, create an option for reporting artificial NCEM that triggers a human review process and guarantees anonymity. This process should be designed to mitigate adverse use of this reporting pathway for taking down consensual explicit images or generally any content that a user does not agree with.
8. In the platform user reporting flow, create an option for reporting "suspected AI-generated content" that also triggers either a human review process or direct testing with an AI detection tool as outlined in recommendation 6. If someone selects this option, add a pop-up This process should be designed to mitigate adverse use of this reporting pathway for taking down consensual explicit images or generally any content that a user does not agree with.

Resources and redirects

9. Build out an AI resource initiative beyond redirect landing pages for people to learn in more depth about what generative AI is, how it's used, and how we can be conscious consumers of it. Resources and redirects should be accessible (i.e. clear and easy to read for all users) and culturally appropriate to communities most at risk of AI-based disinformation, sexualization, and abuse.
10. Establish a resource platform specifically for people who have been targeted by explicit non-consensual sexual deepfakes online. This resource should include legal considerations, legal references, references to survivor support programs such as RAINN, statistics on the ubiquity of non-consensual deepfakes, and science on the adverse impacts of being targeted by non-consensual deepfakes.

Data reports and accountability

11. Carry out an annual primary analysis – conducted by a neutral third party – to assess the prevalence and nature of gendered AI-based disinformation on the platform; specifically, this analysis should involve studying the gender breakdown of deepfake targets, common tactics

of manipulation used (eg. appearance, voice, visual) and how those tactics differ based on gender, common topic targets of deepfakes (eg. politics, gender roles, sex), the number of sexual deepfakes circulated, the reach of deepfakes, and patterns in users commonly sharing deepfakes (eg. political accounts, satire). Findings should be open access to researchers and civil society organizations.

12. Carry out an annual secondary analysis – conducted by a neutral third party – to assess the implementation and efficacy of company AI policies; specifically, this analysis should involve studying the number of overall reports of AI-generated content, the number of reports specifically on non-consensual sexual deepfakes, the number of takedowns of AI-generated content out of all reported content, the number of views that unlabeled AI-generated content received before takedowns, and the number of unlabeled AI-generated posts flagged by any automated systems used. Findings should be publicly available.

Signed,

Accountable Tech
Center for Intimacy Justice
Chayn
Civic Shout
Digital Defense Fund
Ekō
EndTAB
GLAAD
Glitch
Global Hope 365
Higher Heights for America
Institute for Strategic Dialogue
Joyful Heart Foundation
Kairos Fellowship
MPower Change
My Image My Choice
MyOwn Image
National Organization for Women
National Women's Law Center
Progress Florida
ProgressNow New Mexico
Religious Community for Reproductive Choice
Reproaction
Rights4Girls
Sexual Violence Prevention Association (SVPA)
Women's March
UltraViolet

Sources:

1. Beware of Virtual Kidnapping Ransom Scam, National Institutes of Health, accessed July 10, 2024

2. Deepfake scams have robbed companies of millions. Experts warn it could get worse, CNBC, May 27, 2024

3. A.I. Is Making the Sexual Exploitation of Girls Even Worse, New York Times, March 2, 2024

4. Gender, AI and the Psychology of Disinformation, Media Diversity Institute, June 5, 2023

5. The State of Deepfakes: Landscape, Threats, and Impact, Deeptrace, September 2019

6. How AI is being used to create 'deepfakes' online, PBS News Hour, April 23, 2023

7. Gender Disinformation through AI Amplification, Our Secure Future, August 28, 2023

8. Disinformation Campaigns Against Women Are a National Security Threat, New Study Finds

9. Op-ed: Financial fraud targets older adults, especially women. How to recognize and prevent it, CNBC Women & Health, March 8, 2024

10. Challenging systematic prejudices: an investigation into bias against women and girls in large language models, UNESCO, March 8, 2024

11. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, June 12, 2023

12. AI image generators often give racist and sexist results: can they be fixed?, Nature, March 19, 2024

13. When Race and Gender are Political Targets: Women of Color Candidates Face More Online Threats Than Others, National Press, October 21, 2022

14. An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States, Center for Democracy & Technology, October 27, 2022

15. Artificial Intelligence and gender equality, UN Women, May 22, 2024

16. Queer Eye for AI: Risks and limitations of artificial intelligence for the sexual and gender diverse community, Open Global Rights, May 26, 2023

17. AI Boom Poses Threat to Trans Community, Experts Warn, New York City News Service, April 9, 2024