CDT Response to the Oversight Board's call for public comments: "Explicit AI Images of Female Public Figures"

Authors: Dhanaraj Thakur and Asha Allen

April 2024

Introduction

The Center for Democracy & Technology (CDT) submits these comments in response to the Oversight Board's request for public comments on "Explicit AI Images of Female Public Figures." CDT's work includes assessing the impacts of online abuse on digital platforms and advocating for solutions that protect free expression, privacy and security, and other fundamental rights.

Deepfakes are synthetic manipulations of identities and expressions in the form of video, images, or audio which make it appear as if someone says or does something they never did.[1] As one well known study noted, the vast majority of deepfake videos are pornographic, and almost all of those are targeted at women.[2] Other researchers have for some time highlighted concerns that women journalists and politicians are often targeted by deepfakes.[3]

The problem of false and sexualized information about women is not new. For example, researchers and journalists have identified cheapfakes or shallowfakes as the manipulation of media in less sophisticated ways compared to deepfakes, including crudely editing, mislabeling, or misrepresenting the original context of an image or video.[4] In fact, researchers note that as far back as the 19th century there were documented examples of women in the U.S. who were warned that photographers could combine a photo of their face with that of another woman's body in a sexualized way.[5] A key difference stemming from the advent of AI today is the increasing ease with which machine learning tools are becoming accessible and affordable through a network of websites and apps that allow users to produce and share deepfakes very quickly and regularly.[6]

The phenomena of deepfakes, cheapfakes, and their antecedents are not random observations but part of a systemic problem - patriarchy. Patriarchy exists where positions of power in political,

---

[1] Shenkman, C., Thakur, D., & Llansó, E. (2021). Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis. Center for Democracy & Technology. https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/

[2] Ajder, H., et . al..(2019). *The State of Deepfakes:Landscape, Threats, and Impact.* Deeptrace Labs.

[3] Di Meco, L. (2019). #ShePersisted Women, Politics, & Power in the New Media World (pp. 1–58). The Wilson Center; Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1).; Chesney, R., & Citron, D. (2018, December 11). Deepfakes and the New Disinformation War. *Foreign Affairs*, *98*(1).

[4] Paris, B., & Donovan, J. (2019). Deepfakes and Cheap Fakes. Data & Society.

[5] Erickson, S. (2023, April 3). Deepfake Technology Poses a Threat to Reality. *Journal of Gender, Race & Justice - The University of Iowa*. https://jgrj.law.uiowa.edu/news/2023/04/deepfake-technology-poses-threat-reality

[6] Koltai, K. (2024, February 23). *Behind a Secretive Global Network of Non-Consensual Deepfake Pornography*. Bellingcat. https://www.bellingcat.com/news/2024/02/23/behind-a-secretive-global-network-of-non-consensual-deepfake-pornography/

social, and economic structures and organizations in a country are dominated by men. When control of these systems of power are perceived to be under threat, as is the case with the increase in the number of women running for political office in the U.S., we observe a disproportionate amount of online harassment and abuse targeted at those women.[7] Such harassment and abuse based on one's gender expression (or online gender-based violence - GBV) can take a range of forms, including non-consensual image/video sharing, and more specifically creating and sharing fake images/video without consent.

However, deepfakes are not only an expression of violence, as they also are created to spread false information about persons or groups based on their gender identity (i.e., gendered disinformation).[8] Such disinformation campaigns often include deepfakes as an attempt to undermine a woman's ability to participate in representative politics by harming them, their staff, and their political candidacy in potentially severe ways. The use of deepfakes targeted at women in politics in particular is a form of online GBV and gendered disinformation that is meant to challenge, control, and attack their presence in spaces of public authority.

The Impacts of Deepfakes on Women in Public Life.[9]

Deepfakes can impact politically engaged women, including candidates, journalists, advocates, and civic leaders, in a variety of ways, not only on the individual level, but on women as a group. For those experiencing these videos and images firsthand, they can prove to be persistent forms of distraction: by trying to regularly refute such attacks and falsehoods, women candidates will have less time to focus on substantive issues, and the wider discussion about them will follow that pattern as well.[10] Further, such experiences can cause personal harm, such as distress, as well as a chilling effect on political or other speech. More broadly, the severity of some deepfake videos as part of a larger campaign of online harassment and disinformation targeted at a woman politician can make other women who are interested in politics more likely to reconsider their ambitions,[11] which in turn harms efforts to build and sustain inclusive democracies. Similarly, the prospect of harassment and other kinds of abuse that can follow from deepfakes can actively discourage women and gender nonconforming individuals in political roles from expressing themselves online in a way that might draw public attention and scrutiny.

Women who are the subject of these campaigns can also face significant long-term effects as, given their severe nature, some of these attacks can yield physical and psychological damage that

---

[7] Thakur, D., Hankerson, D. L., Luria, M., Savage, S., Rodriguez, M., & Valdovinos, M. G. (2022). *An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States*. Center for Democracy & Technology. https://cdt.org/insights/an-unrepresentative-democracy-how-disinformation-and-online-abuse-hinder-women-of-color-political-candidates-in-the-united-states/

[8] Sessa, M. G. (2020). *Misogyny and Misinformation: An analysis of gendered disinformation tactics during the COVID-19 pandemic*. EU DisinfoLab. .

[9] Thakur, D., & Allen, A. (2022). *The Impacts of Online Gender-Based Violence and Disinformation on Women Politicians in Representative Democracies* (UN Women Expert Group Meeting 'Innovation and Technological Change, and Education in the Digital Age for Achieving Gender Equality and the Empowerment of All Women and Girls' 10-13 October 2022). UN Women. https://www.unwomen.org/sites/default/files/2022-12/EP.12_Dhanaraj%20Thakur%20and%20Asha%20Allen.pdf

[10] Oates, S., Gurevich, O., Walker, C., & Di Meco, L. (2019). Running While Female: Using AI to Track how Twitter Commentary Disadvantages Women in the 2020 U.S. Primaries (SSRN Scholarly Paper ID 3444200). Social Science Research Network. https://doi.org/10.2139/ssrn.3444200.

[11] Di Meco, L. (2019).

requires longer recovery times, with implications for their political careers.[12] These harms are equally experienced by women journalists who, as essential civic space actors, are often confronted with similar campaigns aimed to discredit their journalistic efforts and which, in some cases, lead to threats against their physical safety.

Intersectionality and other harms

Gender only represents one type of identity and is perhaps only a starting point for trying to understand the various impacts of deepfakes. In reality, individuals traverse multiple identities all the time, and disinformation can also operate across race, gender, and other aspects of identity simultaneously. Recognizing the reality of intersectional identities challenges researchers and policymakers to understand both how a person may have to contend with multiple sources of oppression at the same time, and the unique impact from this multifaceted oppression.[13] Among other problems, this means that the unique experiences and needs of people who are, for example, neither white nor male can go unexamined and unaddressed in research and policymaking. Limiting our analysis and measures to address deepfakes to the population at large may in turn undermine our ability to effectively counter the harm that such disinformation campaigns and online GBV have on democracy and attempts to advance gender equality.[14]

When we think about deepfakes we should therefore recognize that they will be used to exploit existing forms of discrimination not only based on gender, but also a range of other identities — such as disability status, LGBTQIA+ communities, age, religious background and immigration status.

Although there is limited research using intersectionality to examine deepfakes, we do have some related evidence from other forms of online GBV and gendered disinformation targeted at women politicians that may be instructive. From a study[15] of posts on Twitter/X that targeted a representative sample of all candidates that ran for Congress in the 2020 U.S. election we found that:

1. Women of color candidates were twice as likely as other candidates to be targeted with or the subject of mis- and disinformation, which often included cheapfakes or manipulated images and photos.
2. Women of color candidates were the most likely to be the target of particular forms of online abuse, including sexist abuse (as compared to white women), racist abuse (as compared to men of color), and violent abuse (four times more than white candidates and two times more than men of color.)
3. Women of color candidates were also most likely to be targeted with or the subject of posts that combined mis- and disinformation *and* abuse.

---

[12] Wilfore, Kristina. (2020, September 25) Disinformation and Women's Leadership. Presentation at the Center for Democracy & Technology Research Workshop on Disinformation: Understanding the Impacts in Terms of Race and Gender. September 2020
[13] Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review, 43(6), 1241–1299.
[14] Allen, A. (2023, March 14). CDT Europe's Asha Allen Gives Remarks Before OSCE on Role of Media in Achieving Gender Justice. *Center for Democracy and Technology*. https://cdt.org/insights/cdt-europes-asha-allen-gives-remarks-before-osce-on-role-of-media-in-achieving-gender-justice/
[15] Thakur, D., et. al. (2022.

When we interviewed women of color that ran in those elections, they reported feeling diminished, questioning their worth, and other negative effects. In other words, they perceived the purpose was for them to drop out of politics and to accept the oppression they faced.[16]

The use of automated solutions to address these harms

While many online platforms use various forms of automated technologies to detect and analyze user generated content, some actors are developing new techniques to evade such systems. Deepfakes emerged as one such circumvention effort. There may be some legitimate use cases of the technologies underlying deepfakes, in fields like movie production, game design, or improving quality of real-time video streams.[17]

That said, deepfake detection is presently a major industry priority and challenge. Using AI based tools to detect deepfakes can introduce additional challenges, however, depending on how they are incorporated within existing trust and safety systems. For example, such detection systems could introduce bias depending on the models and data used; they may lack explainability, which could be important particularly during an appeals process with end-users; they often cannot assess context as well as humans; and it's difficult to assess their performance because of a lack of useful metrics, particularly those that can be easily understood by non-experts.[18] Another potential challenge could be the design of detection systems that focus on deepfakes in general. However, as noted earlier, most deepfakes are pornographic and target women specifically, which should have implications for how such systems are designed in the first place.

These limitations point to the importance of complementary mechanisms such as user reporting of deepfakes. In fact, in the cases outlined by the Oversight Board in this call for public comments, one includes user-reporting. Given the focus on female public figures and that (as noted earlier) women of color political candidates in the U.S. are more likely to be targeted with online abuse and disinformation, it is also important to understand their perspectives on reporting mechanisms. Most candidates we spoke to in our study described using social media platform reporting mechanisms at least once. Of these reports, only one respondent successfully petitioned the platform to remove content that was false or abusive. According to these women of color candidates, the platform least responsive to user reporting was Facebook.[19] In another study which included women journalists, users felt that "reporting mechanisms on social media platforms are often profoundly confusing, time-consuming, frustrating, and disappointing."[20]

---

[16] Thakur, et. al. (2022).
[17] Vincent, J. (2020, October 5). *Nvidia says its AI can fix some of the biggest problems in video calls—The Verge*. https://www.theverge.com/2020/10/5/21502003/nvidia-ai-videoconferencing-maxine-platform-face-gaze-alignment-gans-compression-resolution
[18] Shenkman, C., Thakur, D., & Llansó, E. (2021).
[19] Thakur, et. al. (2022).
[20] Vilk, V., & Lo, K. (2023). *Shouting into the Void: Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It—PEN America*. https://pen.org/report/shouting-into-the-void/

A Harmonised Approach to tackling deepfakes in the context of Online GBV

The two cases being reviewed by the Oversight Board are of particular relevance to legislative advancements in the European Union. The EU co-legislative bodies have just adopted the final text of the Directive to combat violence against women and domestic violence, which aligns with the international standards already established by the Council of Europe Istanbul Convention and its first General Recommendation. Once transposed into national law, this means that across the EU, the production, *manipulation, or altering* of an image, video, or similar content which make it falsely appear as though a person is engaged in sexually explicit activities, without that person's consent, and subsequently making that content publicly accessible, will be a criminal offense punishable up to at least one year of imprisonment. The decision of the Oversight Board in these cases therefore may need to take into due regard these obligations and the necessity of ensuring coherence at a global level.

General Recommendations for Meta to address Deepfakes targeted at Women in Public Life

- Meta should clearly articulate policies that prohibit content such as deepfakes that harasses or abuses someone on the basis of gender or race. These policies, and the moderation processes that enforce them, should adopt an intersectional approach that considers the unique ways in which abuse can manifest against women with multiple identities.
- With regard to women politicians, Meta should ideally provide transparency reports around election mis/disinformation before, during, and after an election. These reports could provide a holistic view into content moderation and integrity operations by the service during the period around a specific election, and should include a focus on online GBV, gendered disinformation, and deepfakes that target political candidates, broken down by demographics.
- Meta should make data available to independent researchers that enables them to study the nature and impact of deepfakes, gendered mis- and disinformation, and online GBV on political candidates. This includes annual risk assessments performed in the context of Article 34 of the DSA, which expressly requires mitigation of risks related to the spread of disinformation and GBV.
- Meta should take additional steps to protect and prevent abuse, particularly explicit AI images and other sexualized deepfake abuse targeting women political candidates, journalists, and other public figures. They should:
    - Offer tools that allow users to report content that violates the companies' policies against abuse or mis- and disinformation including additional tooling (e.g., granular levels of control) for verified accounts including women public figures, to quickly escalate abuse reports to specially trained moderators.
    - Ensure that content moderation systems, including human moderators and algorithmic systems, are attuned to the needs of and the threats faced by women public figures, and women public figures whose identities maybe particularly targeted in a given society (e.g., women of color in the U.S. and women of caste oppressed and religious minority communities in India).