ISD | Powering solutions to extremism, hate and disinformation

The Institute for Strategic Dialogue's Recommendations and Comments on Oversight Board Case 2023-029-FB-UA

Thank you to the Oversight Board for the opportunity to comment on case 2023-029-FB-UA**,** regarding an altered video of President Biden. With the rapid development of deepfake technology and artificial intelligence (AI), it is crucial for social media platforms to have policies and safeguards in place to protect not only public figures but private individuals from malicious uses of new technology – especially for the many critical upcoming elections in 2024.

The Institute for Strategic Dialogue (ISD) is an independent, non-profit organization dedicated to safeguarding human rights and reversing the rising tide of polarization, extremism, and disinformation worldwide. Our work includes in-depth research and analysis identifying and tracking online manipulation, mis- and disinformation, hate, and extremism in real time. We also formulate, advocate and deliver evidence-based policy approaches and programming.

Artificial intelligence technology can produce synthetic images and audio of candidates, generating wholly artificial scenarios to further a political agenda. Deepfakes are now so realistic that it is becoming nearly impossible for the average voter to discern fact from fiction. More concerningly, they are already being deployed by candidates and political parties in the US and beyond, demonstrating the need for regulation and clear social media policies. As shown by recent events across the world -- from the January 6 insurrection in Washington, DC to the Brazilian Congress attack in Brasília -- elections are already vulnerable to online disinformation and could become even more so with disinformation driven by AI. With some already referring to the upcoming 2024 US presidential election as "the most online election ever," social media platforms are in a unique position to be at the forefront of preventing harmful online narratives that will likely rapidly emerge and proliferate.

Our submission seeks to address the Oversight Board's request for comments on research into online trends of using altered or manipulated video content and Meta's policies and enforcement practices regarding manipulated media:

1. ***Research into online trends of using altered or manipulated video content to influence the perception of political figures, especially in the United States.***

Previous research published by ISD has found that social media product features are already amplifying election disinformation, harming candidates, and assisting the organizing efforts of those disseminating false and harmful claims. It is now easier than ever for users and networks (such as foreign influence operations) to generate altered or manipulated content containing disinformation or misleading narratives and use social media product features to spread it more widely and efficiently. Past research has shown that women are targeted most often by deepfake videos, particularly with non-consensual intimate images – a trend that will undoubtedly affect women candidates and politicians, who are already targeted by online gendered and sexualized disinformation campaigns.

With the rise of AI-generated deepfakes that anybody can now create, the Federal Election Commission (FEC) is already considering regulating AI-generated deepfakes in political ads ahead of the 2024 election. Google is also implementing a new policy that political ads using artificial intelligence must be accompanied by a prominent disclosure. Civil society organizations have published reports with policy suggestions on protecting democracy and warnings on how AI could affect voters in 2024. The time is ripe for Meta to participate in these discussions and policy work to ensure its platforms are ready for such a pivotal year.

Recommendations:

- Meta's policy teams need to be responsive to these kinds of online trends, especially when it comes to emerging technology, and adapt policies and enforcement quickly and accordingly.
- Meta should be aware of how generative AI affects different users of different backgrounds. For example, women are likely to be more targeted with deepfake non-consensual intimate images than men.
- Many generative AI technologies are also capable of producing content in languages other than English. Meta already allocates fewer resources towards moderating non-English language content. Meta should ensure that adequate resources are being put towards understanding and responding to generative AI content in non-English languages.

2. ***The suitability of Meta's misinformation policies, including on manipulated media, to respond to present and future challenges in this area, particularly in the context of elections.***

While Meta made several strides in combatting election disinformation in past years, the recent backsliding of misinformation and other policies does not bode well for the upcoming 2024 election. For example, in late August, Meta reportedly started allowing users to opt out of its fact-checking program, which gave users access to reliable information – particularly related to elections. ISD found that the criteria laid out in Meta's Manipulated Media Policy do not do enough to address emerging online trends with not only AI-generated deepfakes, but also regular deceptively edited videos, as proven in case 2023-29-FB-UA. Even though the manipulation was apparent and recognizable in this case, it still pushed a narrative containing misinformation about President Biden.

The current Manipulated Media Policy, which is also repeated in the Misinformation Policy, bans videos that have been edited or synthesized in ways that "are not apparent to an average person" and would mislead them to believe that: 1) a subject of the video said words they did not say; *and* 2) the content is a product of artificial intelligence or machine learning. The addition of "and" automatically does not include content that was deceptively edited by a user, which is likely part of the reason why the video in this case was not removed under this policy.

Although Meta did not consider this post for removal under their Hate Speech Policy or Bullying and Harassment Policy, it is important to note that the tactic of calling a public figure (or private individual) a pedophile, whether on the basis of a protected characteristic or not, has been increasingly weaponized by violent and conspiratorial movements and is the sort of language that is more likely to lead to actual violence. ISD's recent report on online and offline anti-drag mobilization highlights how this narrative is used to justify hate and violence against drag performers and members of the LGBTQ+ community. The accusation made in the video against Biden is clearly not a credible allegation and does not provide any sort of evidence or personal testimony, and the furthering of this narrative could lead to harmful online and offline threats against Biden.

Recommendations:

- Meta must recognize that content does not have to be the product of artificial intelligence or machine learning to be manipulated media. It also does not have to be unapparent or particularly well-edited.
- If Meta acknowledges and recognizes that a video is altered or purposefully misleading, it should remove the video from its platforms, no matter how many views it receives. As the Manipulated Media Policy states, this should not extend to content that is parody or satire, or is edited to omit words that were said or change the order of words that were said.
- Meta should include a clause in its Manipulated Media Policy about manipulated actions, not just manipulated speech. In this case, there was no speech involved, just manipulated actions.
- Meta should better protect public figures in its policies and ban disproven false narratives about public figures (in this case, President Biden being a pedophile) from being posted on its platforms.

3. ***Meta's human rights responsibilities when it comes to video content that has been altered to create a misleading impression of a public figure, and how they should be understood with developments in generative artificial intelligence in mind.***

In the US, every individual has a right to commercialize aspects of themselves – including personalities. These rights are protected by the right of publicity (ROP), but in the age of generative AI, the "theft of personality" is a real threat. When looking at the issue from a commercial lens, the misuse of generative AI could damage a public figure's earning potential and brand. But non-consensual deepfakes or content that has been altered to create a misleading impression of a public figure could have reputational damages, too. These damages might not be enough to be clearly labelled as defamation or privacy violations and could therefore fall under the right of publicity, which protects a person's "likeness." Meta has a responsibility to all the public figures on its platforms to protect their personality rights, or ROP.