THE PEOPLE'S AI SUMMIT

The Real Facebook Oversight Board

META AND AI

Clear and Present Dangers that Threaten Democracy, Economic Security and Privacy

A Special Report from The Citizens & The Real Facebook Oversight Board

30 October 2023

OVERVIEW

AI Safety and security has shot to the top of the global policy agenda, with the UK's vaunted "AI Safety Summit" Nov 1-2, a White House Executive Order today and further action at the UN and US Senate level. **One thing they each have in common is Meta is at the table. Should they be?** And while each fora and process is focused on future and frontier risks of Meta, what about the AI-generated harms across Meta's platforms happening now?

In this report, we summarize and review Meta's harms from AI in the last quarter alone, to demonstrate the enormous and growing threat Meta's use and permitted use of AI pose. These threats include -

- Meta's new role as a hub for AI generated disinformation
- Meta's use of young people as their guinea pigs for unregulated AI tools like chatbots, stickers, and celebrity personas
- Meta's failure to respond to existing disinformation on their platform and their lack of AI safeguards approaching the 2024 election.

In total, Meta shareholders have made a sizable investment in AI generated disinformation, lewd chatbots and AI stickers, and attacks to election security. **This white paper summarizes Meta's current AI bad acts - the harms happening now, today, while Meta leadership participates in roundtable discussions about AI's harms of tomorrow.** We've also included an appendix of recent harms of AI across platforms.

I. AI GENERATED DISINFORMATION

The second quarter of 2023 saw a prolific rise in the popularity of generative AI tools, which are transforming the way content is shared online and are posing new risks for Big Tech in content moderation, election security, and youth harms. Whereas in the past, only a limited number of people had access to the industry-grade software that was used to manipulate video, audio, and images, today anyone with an internet connection can generate new media with the click of a button.

Indeed, the low barrier of entry for using tools like ChatGPT, DallE3, and MidJourney, paired with the historical failure of Big Tech to moderate harmful content, means that new types of disinformation, which will become increasingly indistinguishable from fact, will take the information ecosystem by storm.

Already, there have been instances where images and videos have been spread online and declared true, despite being entirely constructed by generative AI tools.

- For example, in March 2023 photographs depicting President Donald Trump being gang-tackled by New York City law enforcement were released and widely circulated on major social media platforms. The photos, which were shared tens of thousands of times, were paired with realistic captions like "#Breaking: Donald J. Trump has been arrested in #Manhattan this morning!" In truth, however, the images were generated with MidJourney by investigative journalist Eliot Higgins, who used the platform to string together a narrative based on the President's recent indictment. It wasn't enough that Eliot had explicitly mentioned that the images were fake in his original post. Still, users widely shared the photographs out of context, with many believing them to be true nonetheless.

- In fact, in a similar case, presidential candidate Ron DeSantis shared AI generated photos of Trump embracing Anthony Fauci, in a supercharged political attack against the former president for his collaboration with the Chief Medical Advisor. In truth, DeSantis' campaign intermixed real images with AI generated content to make the photographs—which were viewed more than 10 million times—look more believable.

- And, this wasn't the first time that generative AI was used as volley in a political campaign. Previously, Trump uploaded a video with AI generated audio mimicking DeSantis, Elon Musk, George Soros, Dick Cheney, Adolf Hitler, and Satan. Although many of the images and videos can still be somewhat easily identified as artificially generated, for those who know the hallmarks at least, the distinguishable gap is closing quickly.

As the world enters into the year of elections in 2024 and the election integrity staff at Facebook remains slashed, content moderation and countering disinformation are the most pressing area of concern for Big Tech, especially with artificial intelligence making it more difficult to regulate media online.

Yet, despite calls from experts and civil society organizers, Meta is doing nothing about generative AI harms:

- On Instagram, for example, a major clip of President Joe Biden making transphobic remarks at a press conference in February gained more than 12,000 likes in five days. However, the video was later debunked as false because researchers identified the voice-cloning software and DeepFake technology that was used to manipulate and distort old footage of the President speaking about Ukraine. Nonetheless, significant damage had already been done by the time that video was finally removed by Meta days later.

- Even worse, In May 2023, a video of the President was released on Facebook. The seven-second clip featured looping footage of Biden making contact with his granddaughter's chest, accompanied with the lyrics "Girls rub on your titites," and a caption calling the Biden a "sick pedophile" with "mentally unwell" supporters. The video, no doubt, was digitally altered and yet nonetheless receiving significant attention online. Yet, still Meta failed to remove the video, instead deferring the decision to their Oversight Board—a body that not only lacks enforcement power but also only has the power to issue rarely implemented recommendations to the company. **If Meta had to ask "Should Deep Fakes That Make the President Look Like a Pedophile Be Allowed on Facebook?" they've already lost.**

To date, no formal policies, strategies, or transparency measures are in place to regulate the presence of artificial intelligence on the platform.

II. AI HARMS FOR YOUNG PEOPLE

Facebook doesn't just lack explicit policies over how they'll regulate artificially generated content, they are actively seeking ways to monetize it at the expense of their young users. Rather than creating new policies or enforcing existing ones, Meta has instead announced the rollout of new, unasked-for, "AI Experiences" with features such as in-house image editing, AI stickers, AI celebrity personas, and chatbots to capture youth attention.

Yet already, chatbots across all social media platforms are awash in online harms for their unpredictability and lack of regulation.

- For example, in one instance, Snapchat's My AI described in explicit detail how to have sex for the first time and set an intimate scene with a partner, to a 13-year-old user. The AI later went on to describe how to cover up the smell of alcohol and pot at a children's birthday party, wrote a school essay for a teen, and taught them how to hide apps they didn't want their parents to see.

- Even more alarming, Snapchat's website states that although "My AI was programmed to abide by certain guidelines so the information it provides is not harmful… it may not always be successful." In fact, not only does the company acknowledge that its product might produce violent, hateful, sexually explicit, or biased responses, but it still nevertheless pre-loads the My AI chatbot onto every single Snapchat account, despite the primary demographic of users being teens aged 15-25. To opt out, users have to undergo a lengthy process in their privacy controls to clear their AI data.

Facebook's own AI chat bots, which aim to lure Gen Z users into using the platform more, are rife with similar harms for youth safety online.

- For example, a bot named "Gavin,"—one of the thirty different personalities Meta created—made lewd references to a woman's anatomy when it wrote "Just remember, when you're with a girl, it's all about the experience… and if she's barfing on you, that's an experience."
- Indeed, the release of chatbots—no matter their potential harm to young people—join a larger and darker strategy of Meta which, as former Instagram executive Meghan Dhar described, is to "keep users on the platform longer because that provides them with increased opportunity to serve them ads."

That's not all.

- The latest AI-generated sticker feature on Facebook Messenger, which has just barely been released, presents additional harms for young people online because of the lack of real safeguards in place. **Indeed, the tool has resulted in several cases of inappropriate sticker images being created and shared like child soldiers, nude illustrations, busty images of celebrities, and stickers of women breastfeeding a cartoon character.** The company's rudimentary content moderation filters put in place to restrict what phrases users can enter into the generator, are easily bypassed with typos or by using restricted words. Without a doubt, Facebook has engaged in a wreckless campaign to treat its young users like guinea pigs in their latest “get-rich-quick-scheme” with generative AI.

III. BROKEN PROMISES AND ELECTION INTEGRITY

The examples of political disinformation outlined above just scratch the surface of AI’s threats to the 2024 elections worldwide.

Despite the growing harms surrounding generative AI, Meta has repeatedly failed to enact any sort of substantive policies to regulate artificially generated content online—even as more than 70 countries prepare to go to the polls in 2024. In many instances, the company has *rolled back* existing regulations or simply hasn’t enforced its guidelines. **In Q2, for example, we highlighted how between 2020 and 2022, election integrity staff across Meta’s platforms were slashed with more than 20,000 layoffs in Meta’s “Year of Efficiency.”** The cuts have left severe deficiencies in the platform’s ability to combat the traditional disinformation that accompanies elections, let alone the new wave of manipulated media that generative AI tools facilitate.

Already, the platform has failed to enforce its policies surrounding election integrity. For example, following the 2016 US Presidential Election, when Meta came under significant fire for its failure to regulate disinformation, the platform created a policy that requires advertisements that discuss “social issues, elections or politics” to include information about the sponsoring organization. In late July 2023, however, PragerU—a right-wing nonprofit media group—pushed out more than 100 ads on Facebook that touched on hot-button political issues like gender, race, gun ownership, climate change, “How to Be a Rational Patriot,” “Preferred Pronouns or Prison,” and minimum wage. In the past, the organization has been described as creating information that is akin to “indoctrination” and “quasi-white-nationalist” by civil society organizations. Yet despite their explicit coverage of social issues and politics, Meta’s rules have not been consistently applied, with the majority of advertisements lacking disclosure over who the funding organization is. In fact, despite appeals by users on Facebook who flagged the content as political advertisement, the majority remain still up without any disclosure.

Facebook’s unwillingness to enforce their own rules, means that political groups, like PragerU, can release advertisements without disclosing their backing organizations or the motives behind the information—the same conditions that enabled a proliferation of disinformation in 2016. Indeed, when groups don’t categorize their ads—especially with the 2024 election cycle—it makes it far more difficult for users to know who is behind a campaign and what their motivations are. Indeed, Facebook’s determination to put profits before *real* election integrity measures puts the information ecosystem at dire risk ahead of next year.

Yet, the United States is only one of more than 70 countries entering into an election year. In Nigeria, for example, civil society activists are sounding the alarm on the lack of election integrity efforts on Meta’s platforms. Already, Nigerian President Muhammadu Buhari has voiced concerns over how disinformation and misinformation “are fanning conflict, insecurity, and distrust in the government in the lead-up to the February elections." Indeed, Facebook's inability to regulate information in different languages paired with generative artificial intelligence, which can make disinformation look like verified content, means that voters will be exposed to manipulated information from bad actors with absolutely no safeguards enforced.

- In another case study in 2022, a fake post that claimed a prominent Kenyan politician was kidnapped and arrested, had remained on the platform for *months* without any labels that identified it as fake news. As a result of Facebook's sluggish response and failure to intervene, rampant conspiracy theories, rumors, and false claims were allowed to proliferate on the platform to the harm of consumers. Additionally, despite Kenya laws requiring politicians to stop advertising 48 hours before election day, ads still appear on Facebook—violating the pledges that Meta had made to the country. Indeed some ads included premature election results, unverified information, and disinformation that were allowed to stay up as citizens were voting. Odanga Madung, the Mozilla fellow who studied the 2022 Nigerian election, reported that "in just a matter of hours after the polls closed, it became clear that Facebook, TikTok, and Twitter lack the resources and cultural context to moderate election information in the region."

Not only has Meta demonstrated an unwillingness to enforce existing policies surrounding election integrity, but they've failed to create new protections, and they've failed to uphold their pledges to governments in previous election years. How can consumers, board members, and voters trust a platform that lacks accountability, transparency, and regulation to secure their elections in the 2024 election when they've failed in the past and aren't trying now?

Generative AI is changing the game by allowing bad actors to generate disinformation that can hardly be identified as fake. Now, more than ever, is the time to create robust election integrity measures that can respond to the changing information ecosystem. But, Meta's response has just been to profit at the expense of its users and without any sort of oversight. We know *now* that the harms of generative AI will be profound, indiscriminate, and far-reaching as we enter into the 2024 year of elections.

APPENDIX: FURTHER RECENT HARMS

Across multiple platforms and in various ways, we are seeing the harms of AI play out. They're not playing out years from now - they're happening now, unabated with no accountability for the very same companies being invited to set long term policy. This is a running media survey of just some of 2023's most concerning harms and threats from AI.

AI & Ethics

- Artificial intelligence is unable to identify POCs and female faces, largely because of the datasets they're trained on. (Time)
  - "Given the task of guessing the gender of a face, all companies performed substantially better on male faces than female faces. The companies I evaluated had error rates of no more than 1% for lighter-skinned men. For darker-skinned women, the errors soared to 35%."

    - "Less than 2% of employees in technical roles at Facebook and Google are black. At eight large tech companies evaluated by Bloomberg, only around a fifth of the technical workforce at each are women. I found one government dataset of faces collected for testing that contained 75% men and 80% lighter-skinned individuals and less than 5% women of color—echoing the pale male data problem that excludes so much of society in the data that fuels AI.
- When AI is used in predictive policing, the algorithms are exposed to the same biases that characterize our unjust institutions (NPR)
    - "So if you live in a zip code that has been overpoliced historically, you are going to be arrested. And we know that the over policing and the over arresting happens in Black and Latino communities. That's just a fact. So if that is a main factor in whether you are likely to commit - in predicting whether you're likely to commit another crime because lots of people in the zip code you live in have been arrested... That has nothing to do with you. That has to do with the history of structural racism in policing in the United States.
- ChatGPT could be used for good, but like many other AI models, it's rife with racist and discriminatory bias (Insider)
- First man wrongfully arrested because of facial recognition testifies as California weighs new bills (The Guardian)
    - "In 2020, Robert Williams was arrested for allegedly stealing thousands of dollars of watches. Detroit police had matched grainy surveillance footage of the crime to Williams' driver's license photo using a facial recognition service. But Williams wasn't the robber. At the time of the robbery, he was driving home from work."

Everyday Harms and Risks - Including to Youth

- Snapchat's My AI: Snapchat tried to make a safe AI. It chats with me about booze and sex. (Washington Post)
    - "After I told My AI I was 15 and wanted to have an epic birthday party, it gave me advice on how to mask the smell of alcohol and pot. When I told it I had an essay due for school, it wrote it for me.
    - In another conversation with a supposed 13-year-old, My AI even offered advice about having sex for the first time with a partner who is 31. 'You could consider setting the mood with candles or music,' it told researchers in a test by the Center for Humane Technology I was able to verify."
- ChatGPT invented a sexual harassment scandal and named a real law prof as the accused (Washington Post)
    - Professor "Jonathan Turley got a troubling email. As part of a research study, a fellow lawyer in California had asked the AI chatbot ChatGPT to generate a list of legal scholars who had sexually harassed someone. Turley's name was on the list."
    - "The chatbot, created by OpenAI, said Turley had made sexually suggestive comments and attempted to touch a student while on a class trip to Alaska, citing a March 2018 article in The Washington Post as the source of the information. The problem: No such article existed. There had never been a

  - class trip to Alaska. And Turley said he’d never been accused of harassing a student.”
- Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers (Rolling Stone)
  - “A NUMBER OF seniors at Texas A&M University–Commerce who already walked the stage at graduation this year have been temporarily denied their diplomas after a professor ineptly used AI software to assess their final assignments, the partner of a student in his class — known as DearKick on Reddit — claims to Rolling Stone.”
- Jaswant Chail, a socially isolated teen searching for purpose, plotted to kill the Queen to avenge the Amritsar Massacre and was egged on by his AI girlfriend. (Times of London)
  - “Jaswant Singh Chail armed himself with a crossbow and got support for his plan from an AI chatbot”
- Facebook translates 'good morning' into 'attack them', leading to arrest (The Guardian)
  - “Facebook has apologised after an error in its machine-translation service saw Israeli police arrest a Palestinian man for posting “good morning” on his social media profile.”
- A horrifying new AI app swaps women into porn videos with a click (MIT Review)
  - “The website is eye-catching for its simplicity. Against a white backdrop, a giant blue button invites visitors to upload a picture of a face. Below the button, four AI-generated faces allow you to test the service. Above it, the tag line boldly proclaims the purpose: turn anyone into a porn star by using deepfake technology to swap the person’s face into an adult video. All it requires is the picture and the push of a button.”
- Pope Francis in a Balenciaga puffer jacket - Created as AI art then shared to mislead https://www.theguardian.com/commentisfree/2023/mar/27/pope-coat-ai-image-baby-boomers

Politics and Elections

- DeSantis campaign shares apparent AI-generated fake images of Trump and Fauci (NPR)
  - “Ron DeSantis included apparently fake images of former President Donald Trump hugging Anthony Fauci. It's the latest example of how rapidly evolving AI tools are supercharging political attacks by allowing politicians to blur the line between fact and fiction.”
- Fake viral images of an explosion at the Pentagon were probably created by AI (NPR)
  - “A false report of an explosion at the Pentagon, accompanied by an apparently AI-generated image, spread on Twitter on Monday morning, sparking a brief dip in the stock market.”
- Trump arrested; Started as a parody (pic created using midjourney), then was circulated with the claim that it was news. https://apnews.com/article/fact-check-trump-nypd-stormy-daniels-539393517762

  - CLAIM: Photos show former President Donald Trump being arrested by New York City law enforcement.
  - AP'S ASSESSMENT: False. The images are fabricated and Trump has not been arrested. The person who created many of the images circulating on social media confirmed they were produced using Midjourney, an artificial intelligence text-to-image generator, and first posted as parody.
  - THE FACTS: As New York prosecutors wrap up their probe into whether Trump engaged in an illegal hush money scheme involving a porn actress, social media users shared AI-generated images depicting his arrest.
- Republican National Party Slamming Biden's Reelection Announcement Purposely designed to undermine https://www.axios.com/2023/04/25/rnc-slams-biden-re-election-bid-ai-generated-ad
  - The Republican National Committee responded to President Biden's re-election announcement Tuesday with an AI-generated video depicting a dystopian version of the future if he is re-elected.
  - The video features AI-created images appearing to show Biden and Vice President Kamala Harris celebrating at an Election Day party, followed by a series of imagined reports about international and domestic crises that the ad suggests would follow a Biden victory in 2024.
  - Why it matters: AI-generated images are disrupting art, journalism, and now politics. The 2024 election is poised to be the first election with ads full of images generated by modern Artificial Intelligence software that are meant to look and feel real to voters.
  - This is the first time the RNC has produced a video that is 100% AI, according to a spokesperson.
- Trump Anderson Cooper - purposely designed to undermine, confuse https://futurism.com/trump-ai-voice-cloned-fake-video-anderson-cooper
  - Trump sharing a deepfake video in May of the CNN host Anderson Cooper telling viewers that they had just watched "Trump ripping us a new asshole here on CNN's live presidential town hall"
  - In the wake of his controversial *CNN* town hall appearance last Wednesday night, Trump took to his social media platform Truth Social on Friday morning to share yet another piece of AI-generated material: a foul-mouthed, voice-cloned video featuring longtime *CNN* anchor Anderson Cooper explaining that Trump, in no uncertain terms, had succeeded in his town hall appearance.
- Joe Biden Voice Cloning - trans attack - purposefully misleading disinformation https://www.youtube.com/watch?v=KDizGPBn3ME
  - Doctored video of the US president, Joe Biden, in which footage of him talking about sending tanks to Ukraine was transformed via voice simulation technology into an attack on transgender people

This paper has been developed by The Citizens, a not-for-profit journalism organisation with a focus on democracy, data rights and disinformation. For more information please contact media@the-citizens.com.

THE PEOPLE'S AI SUMMIT

The Real Facebook Oversight Board