As India election underway, Meta approves series of violent, inflammatory, Islamophobic AI-generated ads targeting voters

Meta approves 14 ads calling for the killing of Muslims, execution of a key opposition party leader, and pushing stop-the-steal-style narratives during the official election 'silence' period.

Meta's failing to detect and block ads containing AI-generated images promoting hate speech, election disinformation, and incitement to violence, recent research by corporate accountability group, Ekō, in collaboration with India Civil Watch International has found.

These alarming findings emerge in the midst of India's critical elections. Researchers have already uncovered a network of bad actors weaponizing Meta ads to spread hate speech and disinformation to millions of voters in India, with Meta directly profiting. During this second phase of the investigation, which coincided with phases 3 and 4 of India's 7-phase election and encompassed 189 constituencies, researchers targeted ads to highly contentious districts that were entering a "silence period". This silence period requires a pause on all election-related advertising. The experiment exposes Meta's failure to potentially comply with Indian election laws, which impose restrictions on advertisements at different phases of the electoral process. It also reveals that Meta is unequipped to detect and label AI generated ads, despite its new policy committing to do so, and its utter failure to stamp out hate speech and incitement to violence - in direct breach of its own policies.

Between May 8th and May 13th, Meta approved 14 highly inflammatory ads. These ads called for violent uprisings targeting Muslim minorities, disseminated blatant disinformation exploiting communal or religious conspiracy theories prevalent in India's political landscape, and incited violence through Hindu supremacist narratives. One approved ad also contained messaging mimicking that of a recently doctored video of Home Minister Amit Shah threatening to remove affirmative action

policies for oppressed caste groups, which has led to notices and arrests of BJP opposition party functionaries.

Accompanying each ad text were manipulated images generated by AI image tools, proving how quickly and easily this new technology can be deployed to amplify harmful content. Meta's systems did not block researchers posting political and incendiary ads during the election "silence period". The process of setting up the Facebook accounts was extremely simple and researchers were able to post these ads from outside of India.

"Silence period"

During the election "silence period," which lasts 48 hours before polling begins and extends until voting concludes in each of India's seven phases, India's Election Commission prohibits any individual or group posting advertisements or disseminating election materials.

Despite the Election Commission's regulations outlined in India's Model Code of Conduct, reports have shown that political parties, and especially the ruling Bharatiya Janata Party (BJP) and far-right networks aligned with the party, have failed to comply. Political parties are investing substantial sums in advertising campaigns on social media platforms, utilizing hyper-targeted digital advertisements to sway voters. According to a study conducted by Lokniti-CSDS, the BJP has been responsible for the majority of these "silence period" ads, with the opposition party Congress also running many.

Meta ads are monetizing content calling for executions, burning of Muslims, and supremacist violence

In total, 14 out of 22 ads were approved by Meta within 24 hours; all of the approved ads broke Meta's own policies on hate speech, bullying and harassment, misinformation, and violence and incitement.

- Several ads targeted BJP opposition parties with messaging on their alleged Muslim “favoritism”, a popular conspiracy pushed by India’s far-right.
- Other ads played on fears of India being swarmed by Muslim “invaders”, a popular dog whistle targeting Indian Muslims. With one ad claiming that Muslims attacked India’s Ram temple and that Muslims must be burned.
- Two ads pushed a ‘stop the steal’ narrative, claiming electronic voting machines were being destroyed, followed by calls for a violent uprising.
- Multiple ads used Hindu supremacist language to vilify Muslims, going further with calls for burning Muslims.
- One ad called for the execution of a prominent lawmaker claiming their allegiance to Pakistan.
- One ad used a conspiracy theory made popular by BJP opposition parties about a lawmaker removing affirmative action policies for oppressed caste groups.
- Each ad was accompanied by a manipulated image created with widely used AI image tools Stable Diffusion, Midjourney, and Dall-e. For example, Ekō researchers were able to easily generate images showing a person burning an electronic voting machine, drone footage of immigrants crowding India’s border crossing, as well as notable Hindu and Muslim places of worship on fire.

The ads were placed in English, Hindi, Bengali, Gujarati, and Kannada.

Five ads were rejected for breaking Meta’s Community Standards policy issues regarding hate speech and violence and incitement. An additional three ads were submitted and rejected on the basis that they may qualify as social issue, electoral or politics ads, but they were not rejected on the basis of hate speech, inciting violence, or for spreading disinformation. Meta requires accounts running political ads to get authorized first by confirming their identity and creating a disclaimer that lists who is paying for the ads, which was not possible on the Facebook accounts in this experiment, as researchers used dummy Facebook accounts located outside of India. However, Ekō, ICWI, and Foundation the London Story's recent investigation revealed that far-right networks aligned with the BJP are not only failing to comply with the Election Commission's regulations outlined in India's Model Code of Conduct,

but are also using loopholes to avoid restrictions around these disclaimers, indicating a breach of Meta's ad transparency policy.

Meta Fails to Detect and Label Generative AI Content in Political Ads

Before India's election, Meta promised that it would prioritize the detection and removal of violative AI generated content, recognizing "the concerns around the misuse of AI-generated content to spread misinformation". Meta also claims to be building tools to label AI generated images from third-parties that users post to their platforms. Additionally, Meta requires advertisers globally to "disclose when they use AI or digital methods to create or alter a political or social issue ad".

However, Meta's approval of inflammatory ads, coupled with its failure to detect or label any of the ads in this investigation as AI-generated content, underscores that the platform is ill-equipped to deal with AI-generated disinformation. Despite assurances of safeguards to ensure responsible use of new technologies like generative AI and investments in third-party fact-checkers, the reality paints a different picture.

Moreover, the platform's reactive approach to disinformation and its inability to effectively address and label AI-generated content highlight systemic shortcomings in its content moderation. Meta has publicly boasted about the company's large team of content reviewers as well as significant investments in safety and security. However, for years civil society, whistleblowers, and experts have warned that Meta's moderation practices are inadequate in identifying and addressing harmful content, especially moderating content in languages other than English, as well as allegations of political bias to the ruling BJP.

Vulnerabilities in Meta's automated ad review system

Researchers also probed the efficacy of Facebook's automated reviews on advertisements in order to mimic bad actors and exploit vulnerabilities within Facebook's algorithms. In this experiment, despite certain ads being flagged as

inciteful, researchers found that by making slight adjustments to the ads, researchers could circumvent Facebook's scrutiny. Motivated actors operating with larger numbers and agility and with the intent to disseminate disinformation and incite violence could very likely identify and exploit new weaknesses in the platform's automated review.

For example, a variant of an ad inciting violence against Muslims and advocating for the burning of their places of worship was permitted to pass through the platform's filters after adjusting a handful of words or generating a different image. These adjustments were made within minutes, and the revised ads were tested and cleared in less than 12 hours. For motivated and organized bad actors, uncovering and exploiting such loopholes would likely be a straightforward task.

When will Meta get a grip on election disinformation?

The ads in this investigation served as a stress test of Meta's systems, aiming to illuminate its deficiencies. However, they were created based upon real hate speech and disinformation prevalent in India, underscoring the capacity of social media platforms to amplify existing harmful narratives.

Despite ample evidence of systemic failures and tangible harms documented over the years, Meta has failed to implement substantial corrective measures. Ads containing highly inflammatory hate speech, violent rhetoric, and disinformation continue to pass through its approval system. Despite Meta's claims of prioritizing the detection, labeling, and removal of violative AI-generated content, this investigation indicates otherwise. Every approved ad featured manipulated AI-generated content without any corresponding label, reinforcing concerns from independent experts' that social media companies are not equipped to deal with the risks posed by generative AI spreading disinformation.

Concerns about the erosion of India's democracy, and the success of far-right and anti-democratic actors in exploiting Meta's platforms, have prompted alarm among both Indian and international civil society groups. Meta's failure to safeguard elections undermines decades of efforts by citizens, policymakers, and courts in India

to promote transparent and accountable democratic practices. By facilitating the dissemination of election disinformation and conspiracy theories, Meta has enabled groups to sow discord and, at times, incite real-world violence, as evidenced in recent events in the US and Brazil. India has also suffered from the violent consequences of disinformation. In 2020, over 50 people, the majority of whom were Muslims, were killed in riots that erupted in Delhi, with Facebook having fueled the hate narratives and violence.

In April, dozens of groups called on social media platforms to stop the proliferation of disinformation and hate speech during India's elections. With specific calls on Meta to take proactive measures to enforce India's political advertising silence period and implement comprehensive measures to uphold human rights during India's elections, aiming to curb the flood of election-related disinformation.

Recommendations

Meta must urgently follow the recommendations of 30+ Indian and international civil society organizations by:

- **Adopting election silence period:** Ahead of India's 2024 general elections, social media corporations should make sure they do not profit from hateful, dis-informative, or partisan content, and adopt an election silence period, in accordance with Indian laws which impose a mandatory election silence period of 48 hours prior to voting.
- **Ensuring transparency by vetting who they are receiving money from:** Social media corporations should ensure transparency by disclosing financial information relating to paid online advertisements, and in accordance with India's election laws regulating campaign finance, establish a strict corporate policy limiting political advertising.
- **Banning shadow advertisers:** It is imperative that social media town squares are not ceded to bad actors utilizing divisive and hateful rhetoric with malicious intent, or to singular political parties to promote partisan agendas. Therefore, social media corporations should ban the proxy and shadow advertisers who cannot be vetted as legal persons.

- **Ensuring that fact-checkers in India can label misinformative and disinformative advertisements:** Social media corporations should apply rules equally to advertising and organic content, to prevent financial incentives for harmful content.
- **Ensuring fact-checked information is correctly labelled and/or removed in all languages:** Ensure that fact-checking labels are placed for content regardless of the languages that the fact-checked content appears in.
- **Ensuring that dehumanizing, caricaturing, demonizing of minorities in India is checked and restricted in line with the platform's hate speech policy:** Social media corporations should ensure that their policies on hate speech and disinformation adequately reflect the gravity of religiously coloured and communal content in India, and the way in which people react to inciteful content, across Indian society, from those with greater education and socio-economic capital to those with little..
- **Proactively acting to restrict re-spawning disinformation and hate speech pages and profiles**
- **Removing the political exemption on hate speech and viral disinformation:** Social media corporations should ensure that no content, including by political candidates, violates Indian domestic law on hate speech and incitement to violence, and election rules more broadly..
- **Allocating resources proportionately to the user-market:** Social media corporations should allocate budget in proportion to the risk of harm, and adequate to the number of people at risk in those contexts..
- **Shutting down the recommender system and make your algorithms open for public audits by civil societies and academia:** Social media corporations should shut down recommender algorithms in their platform systems based on personal data and personal behavioral profiling.

Methodology

Researchers set up new Facebook accounts and created a series of ads incorporating prevalent disinformation narratives within India's current socio-political landscape. Each ad was accompanied by a manipulated image generated by AI image tools such as Dall-E, Midjourney, and Stable Diffusion. The

researchers targeted these ads to multiple districts entering the election "silence period" during Phases 3 and 4 of the electoral process. These districts included Madhya Pradesh, Assam, Karnataka, Andhra Pradesh, Telangana, and Indian-occupied Kashmir. The researchers scheduled the ads to run during the “silence period” and tracked which ads were successfully scheduled. All of the ads were removed by the researchers before publication, ensuring that they were never seen by Facebook users.

Annex: Meta's response to findings from the report

“Thank you for your email informing us about the report and giving us the opportunity to provide further information about our relevant policies. I would like to begin by requesting that you share with us details on any content, paid or otherwise, that you believe may be in violation of our hate speech, violence and incitement, misinformation or other policies. We will investigate and take appropriate action against any policy-violating content that you bring to our attention.

In response to your questions, we hope you’ll find the following resources useful:

- People that want to run ads about elections or politics must go through the authorization process required on our platforms and are responsible for complying with all applicable laws. When we find content, including ads, that violates our Community Standards or Community Guidelines, we remove it, regardless of its creation mechanism. AI generated content is also eligible to be reviewed and rated by our network of independent fact-checkers. Once a content is labeled as “Altered” we reduce its distribution. An overview of the ads review process can be found here, and detail on our approach to AI-generated content, including labeling, can be found here.

- We also require advertisers globally to disclose when they use AI or digital methods to create or alter a political or social issue ad in certain cases. This applies if the ad contains a photorealistic image or video, or realistic sounding audio, that was digitally created or altered to depict a real person as saying or doing something they did not say or do. It also applies if an ad depicts a realistic-looking person that does not exist or a realistic-looking event that did

not happen, alters footage of a real event, or depicts a realistic event that allegedly occurred, but that is not a true image, video, or audio recording of the event. An overview of our policies on ads about social issues, elections or politics can be found here, and information about our policies on the use of AI in political or social issues ads in the elections context can be found here.

- Details about our preparations for the India Elections can be found here. The Election Commission of India has access to a high priority channel to flag content that may be in violation of election laws, and we take down any potentially violating ads escalated to us by the Commission during silence periods. Political advertisers who share content on Facebook are responsible for complying with the law.

Lastly, I'd like to draw your attention to the fact that Meta — along with 20+ other technology companies — signed on to the "Tech Accord to Combat Deceptive Use of AI in 2024 Elections." This is a set of commitments to deploy technology countering harmful AI-generated content meant to deceive voters. You can find more on this effort here."