AI INSIGHT FORUM: ELECTIONS & DEMOCRACY

November 8, 2023
10:30AM to 12:30PM EST

Statement of Neil Potts
Vice President of Public Policy within Trust and Safety
Meta Platforms, Inc.

Introduction

Senator Schumer, and distinguished members of the AI Insight Forum, thank you for the opportunity to participate in this important discussion today regarding AI and election integrity. My name is Neil Potts, and I am the Vice President of Public Policy focused on Trust and Safety at Meta.

Election Integrity Efforts

Meta has developed a comprehensive approach to do its part to help safeguard how elections play out across our services, including by reducing harmful misinformation that can interfere with voting and by combatting covert influence operations. One of our top priorities is taking steps to help protect the upcoming elections in the US and across the globe. With each election we incorporate the lessons we've learned to help stay ahead of emerging threats, and we are intensely focused on addressing industry-wide challenges. As part of this, we have activated an Elections Operations Center dozens of times around the globe since 2018, including for US elections, to bring together subject matter experts across the company for real-time monitoring of emerging trends and content to be reviewed.

Meta has dedicated significant resources to detecting content on our platform, including AI-generated content, that violates our policies, including those regarding misinformation and election-related information. Our investments have allowed us to build technologies to proactively identify content, prioritize the most critical content to be reviewed, and act on content that violates our Community Standards.

We remove content we identify that violates our policies, including attempts to interfere with voting, whether it is AI-generated or not. AI-generated content is also subject to our Community Standards and eligible for fact-checking. For example, if such content is spread in a coordinated manner by fake accounts, this would violate our policies against inauthentic behavior; in such cases, the content posted by such accounts would also be removed. We actively monitor new trends in content, including AI-generated content, and update our policies as needed to account for potential associated harms. We also recognize that the increasing accessibility of generative AI tools may change the ways people share this type of content, and have been discussing the issue with experts.

We enforce our policies through a combination of people and technology that work to identify violations of our Community Standards across the billions of pieces of content that are posted to our platform every day. For example, our systems flag content that may violate our policies, people who use our apps report content to us they believe is questionable, and our own teams review content. We have also built a parallel content review system to flag posts that may be going viral—no matter what type of content it is—as an additional safety net. This helps us catch content that our traditional systems may not pick up. We use this tool to detect and review Facebook and Instagram posts that were likely to go viral and take action if that content violated our policies.

We also reduce the distribution of false information that has been debunked by our independent third-party fact-checking partners. We partner with nearly 100 fact-checking organizations around the world who rate content in more than 60 languages. Many of our third-party fact-checking partners have expertise evaluating photos and videos and are trained in visual verification techniques, such as reverse image searching and analyzing the image metadata that indicates when and where the photo or video was taken. Fact-checkers are able to rate a photo or video by combining these skills with other journalistic practices, including by using research from technical experts, academics, or government agencies.

Our fact-checking partners can rate content as altered, which includes "faked, manipulated or transformed audio, video, or photos." They do not need to identify the creation mechanism to rate the content if they can otherwise debunk it. Once a fact-checker rates a piece of content as altered, or we detect it as near identical, it appears lower in Feed on Facebook. We also dramatically reduce the content's distribution. On Instagram, altered content gets filtered out of Explore and is featured less prominently in feed and stories. This significantly reduces the number of people who see it.

We apply our warning labels for content rated false or altered. Critically, people who see this content, try to share it, or have already shared it, will see warnings alerting them that it is false or altered. Additionally, any person can report potentially violating content—including AI-generated content and ads—through our reporting tools. Our automated systems and human review teams work in more than 70 languages to review user-generated reports. We also do not allow advertisers to run ads that contain content that has been debunked by third-party fact-checkers.

We are also announcing a new policy to help people understand when a social issue, election, and political ad on Facebook or Instagram has been digitally created or altered, including through the use of AI. Starting in the new year, advertisers will have to disclose whenever a social issue, elections, or political ad contains a photorealistic image or video, or realistic sounding audio, that was digitally created or altered to (1) depict a real person as saying or doing something they did not say or do; (2) depict a realistic-looking person that does not exist or a realistic-looking event that did not happen, or alter footage of a real event that happened; or (3) depict a realistic event that allegedly occurred, but that is not a true image, video, or audio recording of the event. When an advertiser discloses that the content is digitally created or altered, Meta will add information on the ad and include it in the Ad Library. If we determine that an advertiser doesn't disclose as required, we will reject the ad and repeated failure to disclose may result in penalties against the advertiser.

One challenge platforms face in enforcing against manipulated media is that it is not possible to detect automatically whether photorealistic content that people share is AI-generated, particularly when that content is generated by sophisticated threat actors. This is why the measures discussed above – which apply regardless of whether content is detected as AI-generated – are so critical. But we also are working with industry partners through Partnership on AI on additional ways to address this problem, including working on visible and invisible watermarks, and reading metadata and watermarks shared by other AI developers. Our fact-checking partners also continue to provide insight and feedback on trends they are seeing about AI-generated misinformation trends, and we will use that to inform our approach in this space going forward. By releasing some of our models for research and commercial use, we are empowering the developer community to quickly develop tools that use generative AI to detect misinfo trends and we are excited to learn from their advancements. Additionally, we're following industry best practices so it's harder for people to spread misinformation with our tools. Images created or edited by Meta AI, restyle, and backdrop will have visible markers so people know the content was created by AI. We're also developing additional techniques to include information within image files that were created by Meta AI, and we intend to expand this to other experiences as the technology improves.

Currently, there aren't any common standards for identifying and labeling AI-generated content across the industry. We think there should be, so we are working with other companies through forums, including the Partnership on AI, in the hope of developing them.

We also signed on to the Partnership's Principles for Synthetic Media, in order to further our collaboration with the industry in this field. We also partnered with Reuters, the world's largest multimedia news provider, to help newsrooms worldwide to identify deepfakes and manipulated media through a free online training course. Our teams are constantly working to improve our systems to detect misinformation and disinformation, and to stop harmful content from spreading.

Using AI to Combat Election-Related Misinformation and Disinformation

AI is also a key part of how we tackle misinformation and other harmful content. Since 2016, we have built an advanced system combining people and technology to review the billions of pieces of content that are posted to our platform every day. Our AI systems flag content that may violate our policies, users report content to us they believe is questionable, and our own teams review content. AI is also integral to the viral content review system mentioned above. We have used this tool throughout elections, and in countries around the world, to detect and review Facebook and Instagram posts that were likely to go viral and take action if that content violated our policies.

While Generative AI enables easier content creation, it also shows huge promise for combatting harmful content. We have started testing large-language models (LLM) by training them on our Community Standards to help determine whether a piece of content violates or not. We are already seeing promising signs the LLMs can perform better than existing machine learning models. We are hopeful the use of generative AI can help us take down more violating content faster and more accurately than existing AI tools.

LLMs can also enable defenders to quickly identify and accurately enforce against violating content across multiple languages. For example, in August we released SeamlessM4T, an LLM that can perform rapid translation for up to 100 languages. Tools like this have incredible potential for communication around the world, and will also be powerful for integrity teams looking to protect elections, and counter any type of fast-moving threat, across many languages.

We believe generative AI could also be particularly useful in enforcing our policies in response to moments of heightened risk, including elections when information is changing rapidly. We are preparing for upcoming elections by talking with experts about potential risks and opportunities, and are working through the feasibility and practicality of technical improvements like what transparency for AI-generated content could look like.

White House Commitments and Legislative Efforts

Important questions about generative AI's impact are not unique to Meta, and therefore are important to tackle as an industry. It is why Meta voluntarily joined the White House's commitments for frontier AI models. These commitments are a critical step in ensuring responsible guardrails are established and they create a model for other governments to follow. We joined these efforts because they represent an emerging industry-wide consensus around the things that we have been building into our products for years. We believe they strike a reasonable balance of addressing today's concerns and convening industry to address the potential risks of the future. They enable the tremendous potential for AI while focusing on the greatest risks. The White House's recent Executive Order builds meaningfully on the White House Commitments.

Continued US leadership by the White House and Congress is important in ensuring that there is a considered, collaborative approach to the regulation of AI. US-led frameworks for approaching these issues would help drive toward a global consensus that doesn't yet exist, and provide alternatives to approaches that are designed to curtail American innovation. The fact that Congress is engaging on these issues encourages us that guardrails will be put in place so that society can benefit from innovation in AI while striking the right balance with protecting rights and freedoms, preserving national security interests, and mitigating risks where necessary.

Conclusion

Thank you for your attention to these important issues. We will continue investing in and improving our processes and tools so we can do our part to protect the integrity of future elections. We look forward to finding new ways to drive innovation and progress, in a manner that is safe and secure and works to protect our elections.