Meta

June 11, 2024

Dear Senator Warner,

Thank you for your letter regarding Meta's efforts to advance election integrity and promote AI transparency. We appreciate the opportunity to discuss our extensive work in these areas.

No tech company does or invests more to protect elections online than Meta. We have around 40,000 people working on safety and security, with more than $20 billion invested in teams and technology in this area since 2016. While we recognize that every election brings its own challenges and complexities, we are confident that our comprehensive approach puts us in a strong position to help protect the integrity of this year's global elections on our platforms.

Meta has also been a pioneer in AI development for more than a decade. We know that progress and responsibility can and must go hand in hand. Generative AI tools offer huge opportunities, and we believe that it is both possible and necessary for these technologies to be developed in a transparent and accountable way.

We also recognize that new technology poses potential risks. That is why we are continually adapting to address new challenges, including by advancing efforts to detect and label AI-generated media. The challenges posed by AI, particularly AI-driven manipulated media, are not unique to Meta and will require a whole-of-industry approach. We have collaborated with global experts with technical, policy, media, legal, civic, and academic backgrounds to inform our policy development and improve the science of detecting manipulated media. We are also proud signatories to both the White House's voluntary commitments and the "Tech Accord to Combat Deceptive Use of AI in 2024 Elections." We look forward to continuing our work, as well as our collaboration with others in the industry to drive transparency and counter potentially harmful AI-generated content.

With that context in mind, please find answers to your specific questions below.

1. **What steps is your company taking to attach content credentials, and other relevant provenance signals, to any media created using your products? To the extent that your product is incorporated in a downstream product offered by a third-party, do license terms or other terms of use stipulate the adoption of such measures? To the extent you distribute content generated by others, does your company attach labels when you assess – based on either internal classifiers or credible third-party reports – to be machine-generated or machine-manipulated?**

As the difference between human and synthetic content gets blurred, people want to know where the boundary lies. We have committed to providing transparency and additional context to address AI-generated content. As such, it is important that we help people know when photorealistic content has been created using AI on our platforms. We also recognize that the technology landscape continues to evolve, and we are committed to continually improving our protections to address the range of risks that may emerge in the most effective way possible.

Earlier this year, we announced changes to our approach to identifying and labeling AI-generated organic content. This includes labeling a wider range of video, audio, and image content when we detect industry standard AI image indicators or when people disclose that they are uploading AI-generated content. If we determine that digitally created or altered image, video or audio content creates a particularly high risk of materially deceiving the public on a matter of importance, we may add a more prominent label. This overall approach gives people more information about the content so they can better assess it and so they will have context if they see the same content elsewhere. We will keep this content on our platforms so we can add informational labels and context, unless the content otherwise violates our policies. For example, we will remove content, regardless of whether it is created by AI or a person, if it violates our policies against voter interference, bullying and harassment, violence and incitement, or any other policy in our Community Standards.

When organic content with photorealistic images is created using Meta's AI feature, we take several steps so that people know AI is involved, including putting visible markers that you can see on the images (for example, applying "Imagined with AI" labels), and embedding both invisible watermarks and metadata within image files. Using both invisible watermarking and metadata in this way improves both the robustness of these invisible markers and helps other platforms identify them. We have been working with industry to develop common standards for identifying AI-generated content through forums like the Partnership on AI (PAI), and the invisible markers we use are in line with PAI's best practices.

At the same time, we are looking for ways to make it more difficult to remove or alter invisible watermarks. For example, Meta's AI Research lab FAIR recently shared research on an invisible watermarking technology we are developing called Stable Signature. This integrates the watermarking mechanism directly into the image generation process for some types of image generators, which could be valuable for open source models so the watermarking cannot be disabled. But, while companies are starting to include signals in their image generators, they have not started including them in AI tools that generate audio and video at the same scale. As the industry works towards this capability, we are adding a feature for people to disclose when they share AI-generated video or audio so we can add a label to it. We will require people to use this disclosure and label tool when they post organic content with a photorealistic video or realistic-sounding audio that was digitally created or altered, and we may apply penalties if they fail to do so.

In addition, advertisers also now have to disclose when they use AI or other digital techniques 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.

With respect to your question about license terms, in keeping with our commitment to Responsible AI development, Meta has undertaken a number of initiatives to discourage improper uses of its models, including Llama 2 and Llama 3. For example, we designed a bespoke license that includes a detailed and thought-out set of use restrictions that strictly prohibit a wide range of malicious uses, including the intentional deception or misleading of others, while making sure that Meta retains the ability to audit uses to ensure compliance. We have implemented numerous ways of reporting violations of this policy including through reporting issues with the model, risky content generated by the model, bugs and security concerns, or violations of the Acceptable Use Policy. We can use this information to take enforcement actions against individual licensees who violate our Acceptable Use Policy or who fail to comply with audits. Our Terms of Service for Meta AIs similarly prohibit access or use of Meta AIs in any manner that would deceive or mislead others, among other things.

2. **What specific public engagement and education initiatives have you initiated in countries holding elections this year? What has the engagement rate been thus far and what proactive steps are you undertaking to raise user awareness on the availability of new tools hosted by your platform?**

Over many years, Meta has developed a comprehensive approach for helping to protect elections on our platforms. We have also built the largest independent fact-checking network of any platform, with nearly 100 partners around the world to review and rate viral misinformation in more than 60 languages.

We remain focused on providing people reliable election information while combating misinformation across languages. That is why we continue to connect people with details about voter registration and the election from their state and local election officials through in-app notifications and our Voting Information Center. And in the United States, when people search for terms related to the 2024 elections on Facebook and Instagram they will see links to official information about how, when, and where to vote.

Since 2018, we have provided industry-leading transparency for ads about social issues, elections or politics. Advertisers who run these ads are required to complete an authorization process and include a “paid for by” disclaimer. These ads are then stored in our publicly available Ad Library for seven years. As described above, starting this year, advertisers also have to disclose when they use AI or other digital techniques to create or alter a political or social issue ad to depict a real person as saying or doing something they did not say or do or a realistic-looking person that does not exist or a realistic-looking event that did not happen, or 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.

Additionally, we label state-controlled media on Facebook, Instagram and Threads so that users know when content is from a publication that may be wholly or partially under the editorial control of a government.

In the lead-up to major elections, we also activate country-specific Elections Operations Centers, bringing together experts from across the company from our intelligence, data science, engineering, research, operations, content policy and legal teams to identify potential threats and put specific mitigations in place across our apps and technologies in real time.

While we recognize that every election brings its own challenges and complexities, we are confident that our comprehensive approach puts us in a strong position to help protect the integrity of this year's elections on our platforms.

3. **What specific resources has your company provided for independent media and civil society organizations to assist in their efforts to verify media, generate authenticated media, and educate the public?**

As noted above, we believe that addressing the challenges of AI requires a whole-of-industry approach. That is why we have collaborated with global experts with technical, policy, media, legal, civic, and academic backgrounds to inform our policy development and improve the science of detecting manipulated media. For example, Meta is a founding member of PAI, and is participating in its Framework for Collective Action on Synthetic Media, an important step in ensuring guardrails are established around AI-generated content. Meta contributed to PAI's recently published guidance on building a glossary for synthetic media transparency.

In Europe, we are also working with the European Fact-Checking Standards Network (EFCSN) to train fact-checkers across Europe on detection of AI-generated or AI-altered media misinformation and to raise the public's awareness on this issue through media literacy campaigns. This project aims to improve the skills and capabilities of the European fact-checking community in debunking and countering AI-generated misinformation, facilitate common standards in addressing and fact-checking AI content, and inform relevant stakeholders on the state of AI-generated misinformation across 30 European markets.

We have invested heavily in our third-party fact-checking program to tackle AI-generated content. 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 also rate digitally created or edited content as "Altered," which includes "manipulated or transformed audio, video, or photos" when it risks misleading people about something consequential that has no basis in fact. For example, fact-checkers may rate digitally created or edited content that makes a false claim that is separate from the digitally created or edited media, such as a watermarked-AI created image depicting a fictitious event with a caption asserting the event is real. Fact-checkers 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. 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.

As AI-generated content continues to grow, there will be debates across society about what should and should not be done to identify both synthetic and non-synthetic content. We want to help people know when photorealistic images have been created or edited using AI, so we will continue to collaborate with industry peers through forums like the PAI and remain in a dialogue with governments and civil society – and we will continue to review our approach as technology progresses.

4. **What has been your company's engagement with candidates and election officials with respect to anticipating misuse of your products, as well as the effective utilization of content credentialing or other media authentication tools for their public communications?**

In addition to the election integrity measures outlined in response to Question 2, we work with state and local elections officials to issue Voting Alerts with the latest information about registering and voting to people in their communities. We have also invested in proactive threat detection and have expanded our policies to help address harassment against election officials and poll workers.

Additionally, safety enhancements like Advanced Protection on Facebook offer security tools and additional protections for candidates and their campaigns as well as local officials. Through this program, we help accounts on Facebook that may face additional threats during an election cycle adopt stronger account security protections, like two-factor authentication. The program also provides additional security protections for people's Facebook accounts and Pages, including monitoring for potential hacking threats. This allows us to more quickly detect potentially suspicious account activity by monitoring for attempts to hack the account, such as unusual login locations or unverified devices. Should candidates or election officials have concerns about the misuse of our products or the appropriateness of labeling or handling of their communications, they may always contact us directly via email or via our Meta Support Pros.

5. **Has your company worked to develop widely-available detection tools and methods to identify, catalogue, and/or continuously track the distribution of machine-generated or machine-manipulated content?**

The approach described in Question 1 represents the cutting edge of what we believe is technically possible right now. As described, we take several steps to detect and then let people know that organic content has been developed or altered using AI, including visible labels, as well as invisible watermarks and metadata within certain image files. However, it is not yet possible to identify all AI-generated content, and there are ways that people can strip out or obfuscate invisible markers. So we are pursuing a range of options to improve our AI detection capabilities.

This work is especially important as this is likely to become an increasingly adversarial space in the years ahead. People and organizations that actively want to deceive people with AI-generated content will look for ways around safeguards that are put in place to detect it. Across our industry and society more generally, we will need to keep looking for ways to stay one step ahead.

6. **(To the extent your company offers social media or other content distribution platforms) What kinds of internal classifiers and detection measures are you developing to identify machine-generated or machine-manipulated content? To what extent to these measures depend on collaboration or contributions from generative AI vendors?**

Please see our response to Question 1.

7. **(To the extent your company offers social media or other content distribution platforms) What mechanisms has your platform implemented to enable victims of impersonation campaigns to report content that may violate your Terms of Service? Do you maintain separate reporting tools for public figures?**

We believe that reporting is an essential tool for people to stay safe and to help us respond to misleading and manipulated content. That is why we encourage our users to report content to us that they believe violates our policies using the dedicated tools we have designed for our services. This includes Pages, Groups, profiles, individual posts, and ads, among others. People with business accounts may also report content and accounts that infringe on their rights or impersonate them to our Business Help Center.

Our automated systems also flag content that may violate our policies. AI has improved to the point that it can detect violations across a wide variety of areas, often with greater accuracy than reports from users. This helps us detect misleading or violating content and prevent it from being seen by hundreds or thousands of people.

As described above, we also offer safety enhancements like Advanced Protection for candidates and their campaigns, as well as local officials. This program allows us to more quickly detect potentially suspicious account activity by monitoring for attempts to hack the account, such as unusual login locations or unverified devices.

8. **(To the extent your company offers generative AI products) What mechanisms has your platform implemented to enable victims of impersonation campaigns that may have relied on your models to report activity that may violate your Terms of Service?**

Please see our response to Question 7.

9. **(To the extent your company offers social media or other content distribution platforms) What is the current status of information sharing between platforms on detecting machine-generated or machine-manipulated content that may be used for malicious ends (such as election disinformation, non-consensual intimate imagery, online harassment, etc.)? Will your company commit to participation in a common database of violative content?**

As noted above, we believe that addressing the challenge of AI-driven manipulated media requires a whole-of-industry approach. That is why we work with industry peers to align on technologies that can make it easier for us and other platform providers to detect when someone shares content that has been AI-generated.

We are dedicated to responsible use of new technologies as well as combating the spread of deceptive AI content in elections through the Tech Accord. We have been working with other companies in our industry to develop common standards for identifying AI-generated content through forums like the PAI. Additionally, we were pleased to make voluntary commitments alongside others in the industry, including a pledge to develop robust technical mechanisms to identify AI-generated content, such as digital watermarking with respect to frontier models.

We continue to work with companies across the industry to address malicious behavior when we observe it on our platforms, including the malicious use of AI-generated content. In our most recent Adversarial Threat Report, we detail our disruption of multiple networks that utilized content generated by third-party GenAI systems. For example, we were able to work with a third-party GenAI platform to disrupt the malicious networks' use of both our platforms.

We know this work is bigger than any one company and will require a huge effort across industry, government, and civil society. We will continue to work collaboratively with others to develop common standards and guardrails.

Thank you again for the opportunity to answer your questions. We look forward to working with your offices going forward.

Sincerely,

Kevin J Martin

Kevin Martin
V.P. North America Policy