A Tech Accord to Combat Deceptive Use of AI in 2024 Elections

[Proposed for public signature and announcement by
technology companies at the Munich Security Conference on February 16, 2024]

2024 will bring more elections to more people than any year in history, with more than 40 countries and more than four billion people choosing their leaders and representatives through the right to vote. At the same time, the rapid development of artificial intelligence, or AI, is creating new opportunities as well as challenges for the democratic process. All of society will have to lean into the opportunities afforded by AI and to take new steps together to protect elections and the electoral process during this exceptional year.

The intentional and undisclosed generation and distribution of Deceptive AI Election content can deceive the public in ways that jeopardize the integrity of electoral processes. For the purpose of this accord, **Deceptive AI Election Content** consists of convincing AI-generated audio, video, and images that deceptively fake or alter the appearance, voice, or actions of political candidates, election officials, and other key stakeholders in a democratic election, or that provide false information to voters about when, where, and how they can lawfully vote.

This accord seeks to set expectations for how signatories will manage the risks arising from Deceptive AI Election Content created through their publicly accessible, large-scale platforms or open foundational models, or distributed on their large-scale social or publishing platforms, in line with their own policies and practices as relevant to the commitments in the accord. Models or demos intended for research purposes or for primarily business-to-business (or "enterprise") uses, which present different risks and opportunities to mitigate harm, are not covered by this accord.

While the novelty of Deceptive AI Election Content warrants action during this exceptional year, this accord acknowledges that Deceptive AI Election Content is not alone in posing these risks, in that traditional manipulations ("cheapfakes") can be used for similar purposes. It also acknowledges it is essential for Signatories to continue investing in addressing other key election risks, such as cyber security threats to campaigns and elections writ large–as those present an ongoing threat to the safety and integrity of democratic elections around the world.

AI also offers important opportunities for defenders looking to counter bad actors. It can support rapid detection of deceptive campaigns, enable teams to operate consistently across a wide range of languages, and help scale defenses to stay ahead of the volume that attackers can muster. AI tools can also significantly lower the cost of defense overall, empowering smaller institutions to implement more robust protections. These benefits can help counter adversaries leveraging AI technology.

As leaders and representatives of organizations that value and uphold democracy, we recognize the need for a whole-of-society response to these developments throughout the year. We are committed to doing our part as technology companies, while acknowledging that the deceptive use of AI is not only a technical challenge, but a political, social, and ethical issue and hope others will similarly commit to action across society.

We affirm that the protection of electoral integrity and public trust is a shared responsibility and a common good that transcends partisan interests and national borders.

We appreciate that the effective protection of our elections and electoral processes will require government leadership, trustworthy technology practices, responsible campaign practices and reporting, and active educational efforts to support an informed citizenry.

We will continue to build upon efforts we have collectively and individually deployed over the years to counter risks from the creation and dissemination of Deceptive AI Election Content and its dissemination, including developing technologies, standards, open- source tools, user information features, and more.

We acknowledge the importance of pursuing this work in a manner that respects and upholds human rights, including freedom of expression and privacy, and that fosters innovation and promotes accountability.

We acknowledge the importance of pursuing these issues with transparency about our work, without partisan interests or favoritism towards individual candidates, parties, or ideologies, and through inclusive opportunities to listen to views across civil society, academia, the private sector, and all political parties.

We recognize that no individual solution or combination of solutions, including those described below such as metadata, watermarking, classifiers, or other forms of provenance or detection techniques, can fully mitigate risks related to Deceptive AI Election Content, and that accordingly it behooves all parts of society to help educate the public on these challenges.

We sign this accord as a voluntary framework of principles and actions to advance seven principal goals:

1. Prevention: Researching, investing in, and/or deploying reasonable precautions to limit risks of deliberately Deceptive AI Election Content being generated.

2. Provenance: Attaching provenance signals to identify the origin of content where appropriate and technically feasible.

3. Detection: Attempting to detect Deceptive AI Election Content or authenticated content, including with methods such as reading provenance signals across platforms.

4. Responsive Protection: Providing swift and proportionate responses to incidents involving the creation and dissemination of Deceptive AI Election Content.

5. Evaluation: Undertaking collective efforts to evaluate and learn from the experiences and outcomes of dealing with Deceptive AI Election Content.

6. Public Awareness: Engaging in shared efforts to educate the public about media literacy best practices, in particular regarding Deceptive AI Election Content, and ways citizens can protect themselves from being manipulated or deceived by this content.

7. Resilience: Supporting efforts to develop and make available defensive tools and resources, such as AI literacy and other public programs, AI-based solutions (including open-source tools where appropriate), or contextual features, to help protect public debate, defend the integrity of the democratic process, and build whole-of-society resilience against the use of Deceptive AI Election Content.

In pursuit of these goals, we commit to the following steps through 2024:

1. **Developing and implementing technology to mitigate risks related to Deceptive AI Election content** by:

    a. Supporting the development of technological innovations to mitigate risks arising from Deceptive AI Election Content by identifying realistic AI-generated images and/or certifying the authenticity of content and its origin, with the understanding that all such solutions have limitations. This work could include but is not limited to developing classifiers or robust provenance methods like watermarking or signed metadata (e.g. the standard developed by C2PA or SynthID watermarking).

    b. Continuing to invest in advancing new provenance technology innovations for audio video, and images.

    c. Working toward attaching machine-readable information, as appropriate, to realistic AI-generated audio, video, and image content that is generated by users with models in scope of this accord.

2. **Assessing models in scope of this accord to understand the risks they may present regarding Deceptive AI Election Content** so we may better understand vectors for abuse in furtherance of improving our controls against this abuse.

3. **Seeking to detect the distribution of Deceptive AI election content** hosted on our online distribution platforms where such content is intended for public distribution and could be mistaken as real. This might include using detection technology, ingesting open standards-based identifiers created by AI-producing companies or using content moderation services, enabling creators to disclose their use of AI when they upload content, and/or providing pathways for the public to report suspected Deceptive AI Election Content.

4. **Seeking to appropriately address Deceptive AI Election Content we detect** that is hosted on our online distribution platforms and intended for public distribution, in a manner consistent with principles of free expression and safety. This may include—but is not limited to—adopting and publishing policies and working to provide contextual information on realistic AI-generated audio, video, or image content where we can detect it so it is clear it has been AI-generated or manipulated.

   In considering actions, operators of online platform services will pay attention to context and in particular to safeguarding educational, documentary, artistic, satirical, and political expression.

5. **Fostering cross-industry resilience to Deceptive AI Election Content** by sharing best practices and exploring pathways to share best-in-class tools and/or technical signals about Deceptive AI Election Content in response to incidents.

6. **Providing transparency to the public** regarding how we address Deceptive AI Election Content—for instance, by publishing the policies that explain how we will address such content, providing updates on provenance research, or informing the public about other actions taken in line with these commitments.

7. **Continuing to engage with a diverse set of global civil society organizations, academics,** and other relevant subject matter experts through established channels or events, in order to inform the companies' understanding of the global risk landscape as part of the independent development of their technologies, tools, and initiatives described in this accord.

8. **Supporting efforts to foster public awareness and all-of-society resilience** regarding Deceptive AI Election Content—for instance by means of education campaigns regarding the risks created for the public and ways citizens can learn about these risks to better protect themselves from being manipulated or deceived by this content; via tools, interfaces, or procedures that can provide users useful context about the content than see online; by developing and releasing open source tools to support others who try to mitigate these risks; or by otherwise supporting the work of organizations and communities engaging in responding to these risks.