OpenAI

1960 Bryant Street
San Francisco, CA 94110

May 29, 2024

From:
Anna Makanju
VP of Global Affairs
OpenAI

To:
Senator Mark Warner
703 Hart Senate Office Building
Washington DC

Dear Senator Warner,

I am writing in response to your May 14 letter on the *Tech Accord to Combat Deceptive Use of AI in 2024 Elections* (The Accord). We welcome the opportunity to describe OpenAI's approach to the 2024 global elections, and how we are advancing election integrity worldwide through The Accord.

Developing tools for content provenance and transparency

We believe that improved provenance for audiovisual content is a key tool that can empower voters to assess the origin and history of media. All content provenance techniques have strengths and limitations, and we have been researching and developing a wide variety of techniques. The overall effectiveness of specific techniques may depend on external factors such as collaboration within the broader AI ecosystem - for example through ensuring the widespread use and availability of metadata-based techniques across platforms.

Earlier this year, we implemented the Coalition for Content Provenance and Authenticity's (C2PA) metadata approach for images created using our text-to-image generation model DALL·E 3. The C2PA standard is a digital credential that encodes details about the content's provenance using cryptography. It is an important tool in indicating the origin of AI-generated audiovisual content, and we hope to see its continued adoption within the industry. We will also be integrating C2PA metadata for Sora, our video generation model. Reflecting our commitment to this standard, we have recently joined the C2PA Steering Committee and look forward to contributing to its development.

Importantly, C2PA is not only used for AI generated images. C2PA is also being implemented by a broad range of actors, including news organizations, software companies, and camera manufacturers, as a means to track or show the origin of different types of content. As adoption of the standard increases, users will become familiar with it and may even come to expect this kind of metadata in audiovisual media, thereby filling a crucial gap in digital content authenticity practices.

OpenAI is also developing a detection image classifier - a tool that uses artificial intelligence to assess the likelihood that an image was created using DALL·E 3. Internal testing has shown promising results: the classifier has exhibited high accuracy in distinguishing between non-AI

generated images and those created with DALL·E 3. It has also proven robust to common image modifications such as compression, cropping, and saturation changes. However, there are other image modifications which may reduce performance, and the classifier is less reliable in distinguishing images generated by DALL·E 3 from those made by other AI models. Although further research in this area is needed, we predict classifiers will be an important piece of the provenance ecosystem.

Additionally, we are experimenting with tamper-resistant watermarking. This is a technique in which the generated media itself contains a subtle signal that is not apparent to the viewer or listener but can be detected with the right tool. To this end, we have incorporated audio watermarking into Voice Engine, our custom voice model, which is currently in a limited research preview.

We are committed to continuing our research on content provenance tools, focusing on their relative advantages for different content modalities and embedding them into our technologies to ensure they are as transparent and secure as possible.

Promoting Authenticity and Trust in Election Information

We believe that protecting the integrity of elections requires collaboration from every corner of the democratic process. We view it as part of our role to make sure our technology is not used in a way that could undermine trust in elections. Our work in this area includes directing people to authoritative sources of information about voting in the U.S, prohibiting the use of our tools to mislead people about candidates or institutions, and working with State and Federal officials to learn more about where we can help.

In order to promote access to accurate voting information we have partnered with the National Association of Secretaries of State (NASS). As a result of our collaboration, when a user in the United States asks certain procedural election-related questions, for example, where they should go to vote, ChatGPT will direct them to CanIVote.org, the authoritative website on US voting information. Similarly, ahead of the 2024 elections for the European Parliament, ChatGPT now directs users to the European Parliament's official source of voting information. Lessons from this work will continue to inform our approach in other countries and regions.

Additionally, we do not allow developers to create chatbots that pretend to be real people (e.g., candidates) or institutions (e.g., local

governments), and we have enabled a reporting mechanism in the GPT[1] store for potential violations of our usage policies.

We have also engaged with the Cybersecurity and Infrastructure Security Agency and other government stakeholders to ensure that we are sharing information and learning from the government to inform our approach.

Investing in societal resilience, public trust, and education

We believe it is vital to educate the public about the potential risks of AI-generated content, such as deepfakes, as well as on the available provenance tools and advancements in the field. To this end, we are building a number of public engagement and education initiatives with the aim of fostering public trust and improving societal resilience to AI-generated media.

To support AI education and literacy and promote the wider use and comprehension of provenance tools and standards, including C2PA, we launched a societal resilience fund in partnership with Microsoft. Through this initiative, the International Institute for Democracy and Electoral Assistance (International IDEA) will conduct global trainings to help foster public awareness and all-of-society resilience to AI. They will focus on equipping Electoral Management Bodies (EMBs), civil society, and the media with the necessary skills to navigate the challenges and opportunities of AI. Partnership on AI will use the fund to enhance its groundbreaking Responsible Practices for Synthetic Media, a framework on how to responsibly develop, create, and share synthetic media. And the Older Adults Technology Service (OATS) from AARP is set to create and implement a training initiative aimed at educating American adults aged 50 and above on the basic principles of AI, focusing on both in person and online training sessions to help older adults adapt to an increasingly AI-driven world.

In order to promote independent research on novel content provenance tools, we have opened applications to access our image detection classifier to a group of testers - including research labs and research-oriented journalism nonprofits. This will help us assess the classifier's effectiveness and analyze its real-world application, which will be vital to understand when it performs best and what its limitations are.

We have organized multiple roundtables bringing together civil society and academics for dialogues on elections preparation and risks. Outside the US, we have met with policymakers and election officials in regions hosting elections this year, including Europe, India, South Korea,

[1] GPTs are custom versions of ChatGPT that users can create for a specific purpose, such as support for writing code, sifting through long documents, and deciding which hike to go on over the weekend.

and Mexico, to understand election integrity risks and work together to mitigate them. We believe engagements like these are an essential component for understanding and attending to the concerns of policymakers, which helps to develop trust, build resilience, and continuously improve election integrity.

We look forward to continuing our engagement with you on these very important matters.

Sincerely,

Anna Makanju
Vice President of Global Affairs
OpenAI