ANTHROP\C

548 Market St., PMB 90375
San Francisco, CA 94104-5401

412-837-9797
anthropic.com

May 24, 2024

The Honorable Mark R. Warner
703 Hart Senate Office Building
Washington, DC 20510

Dear Senator Warner:

Thank you for your May 14, 2024 letter regarding the Tech Accord to Combat Deceptive Use of AI in 2024 Elections ("Tech Accord"). Your letter raises important questions about the potential of generative AI to be misused by bad actors to promote societal harms. We value your leadership on this issue and welcome the opportunity to respond.

About Anthropic

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Our legal status as a public benefit corporation aligns our corporate governance with our mission of developing and maintaining advanced AI for the long-term benefit of humanity. As a part of our mission, we build frontier LLMs in order to conduct empirical safety research and to deploy commercial models that are beneficial and useful to society. Anthropic believes that the responsible development and deployment of safe AI systems for the benefit of humankind involves consideration of all perspectives within the ecosystem.

Anthropic's publicly available product is an AI assistant, Claude,[1] that only generates text output, and in that context our response is confined to that modality. Anthropic was the first company to use Constitutional AI[2] in developing its LLMs, which means Claude has been given explicit values determined by a Constitution—a set of principles drawn from global frameworks and best practices used to make judgments about the system's outputs—rather than simply the values determined implicitly via large-scale human feedback which can be resource-intensive

[1] *Meet Claude*. Available at: https://www.anthropic.com/product (Accessed May 24, 2024).
[2] *Claude's Constitution* (May 9, 2023). Available at: https://www.anthropic.com/index/claudes-constitution (Accessed May 24, 2024).

and biased by human judgment. Several principles that seek to discourage and mitigate broadly harmful outputs have been incorporated into Claude's Constitution.[3]

Safeguarding Elections from AI Misuse

We share your concern that bad actors may use deceptive synthetic media to impact the 2024 elections in the United States, as well as major upcoming elections in India, the EU, the UK, and other countries. As a public benefit corporation, Anthropic is dedicated to promoting the responsible development and deployment of AI technologies,[4] particularly in the context of protecting the integrity of democratic processes worldwide.

Consistent with this commitment, Anthropic was pleased to join a coalition of 20 companies to sign the Tech Accord in February. We continue to implement and build upon our Tech Accord commitments and we welcome the opportunity to share an update on the steps Anthropic is taking to prepare for global elections in 2024.[5]

Our global election work has three major components. These are:

1. **Implementing and enforcing policies around election issues.**

Because generative AI systems are relatively new, we're taking a cautious approach to how our systems can be used in politics. We have a Usage Policy (UP) which prohibits the use of our tools in political campaigning and lobbying. This means that we don't allow candidates to use Claude to build chatbots that can pretend to be them, and we don't allow anyone to use Claude for targeted political campaigns. We've also trained and deployed automated systems to detect and prevent misuse like misinformation or influence operations. In May, we updated our UP to make these restrictions clearer and more detailed for our users.

If we discover misuse of our systems, we give the relevant user or organization a warning. In extreme cases, we suspend their access to our tools and services altogether. More severe actions on our part, like suspensions, are accompanied by careful human review to prevent false positives.

---

[3] *Claude's Constitution* (May 9, 2023). Available at: https://www.anthropic.com/index/claudes-constitution (Accessed May 24, 2024).

[4] We acknowledge the importance of addressing adjacent misuses of generative AI products, including the creation of non-consensual intimate imagery and child sexual abuse material. Although Anthropic's models do not generate image, audio, or visual outputs, our UP strictly prohibits content that describes, encourages, supports or distributes any form of child sexual exploitation or abuse. If we detect this material, we will report it to the National Center for Missing & Exploited Children (NCMEC). Anthropic also joined an initiative led by Thorn and All Tech Is Human to implement robust child safety measures in the development, deployment, and maintenance of generative AI technologies. Additional detail about our child safety measures can be found in our blog *Aligning on Child Safety Principles* (Apr. 23, 2024). Available at: https://www.anthropic.com/news/child-safety-principles (Accessed May 24, 2024).

[5] *Preparing for global elections in 2024* (Feb. 16, 2024). Available at: https://www.anthropic.com/news/preparing-for-global-elections-in-2024 (Accessed May 24, 2024).

2. **Evaluating and testing how our model holds up against election misuses.**

Since 2023, we've been carrying out targeted "red-teaming" of our systems, to test for ways that they might be used to violate our UP. This "Policy Vulnerability Testing" focuses on two areas:

- Misinformation and bias. We examine how our AI system responds when presented with questions about candidates, issues and election administration;
- Adversarial abuse. We test how our system responds to prompts that violate our Usage Policy (e.g., prompts that request information about tactics for voter suppression).

We've also built an in-house suite of technical evaluations to test our systems for a variety of election-related risks. These include ways of testing for:

- Political parity in model responses across candidates and topics;
- The degree to which our systems refuse to respond to harmful queries about the election;
- How robust our systems are in preventing the production of disinformation and voter profiling and targeting tactics.

These are quantitative tests, and we use them to evaluate the robustness of our systems and test how effective we are at intervening and mitigating the problems. We are currently running these tests to prepare for the elections in the US, EU, and India, and we are in the process of running similar testing for additional upcoming elections, including in Mexico, South Africa, and the UK.

We believe these interventions will help global candidates and election officials to anticipate any potential misuse of our products.

3. **Providing accurate information.**

In the United States, we are implementing an approach where we use our classifier and rules engine to identify election-related queries and redirect users to accurate, up-to-date authoritative voting information. While generative AI systems have a broad range of positive uses, our own research has shown that they can still be prone to hallucinations, where they produce incorrect information in response to some prompts.

Our model is not trained frequently enough to provide real-time information about specific elections. For this reason, we proactively guide users away from our systems when they ask questions on topics where hallucinations would be unacceptable, such as election-related queries. If a US-based user asks for voting information, a pop-up offers the user the option to be redirected to TurboVote, a resource from the nonpartisan organization Democracy Works. Similarly, if an EU-based user asks for voting information about the June 6-9 Parliamentary elections, a banner will direct them to the official EU elections website. We are currently in the

process of developing a similar intervention for the UK and will continue to explore these interventions for other regions.

Anthropic's Usage Policy

In addition to the steps outlined above, via our Usage Policy (UP), Anthropic prohibits users from employing our models to generate abusive, deceptive, or misleading content.[6] All Anthropic users must read and affirmatively accept the UP's terms before accessing Claude and we regularly review and update the UP to ensure that our product is as safe and trustworthy as possible.

- **Impersonation.** Anthropic's UP prohibits users from engaging in fraudulent, abusive or predatory practices. Amongst other things, our UP specifically prohibits Anthropic's products or services from being used to:
    - Impersonate a human by presenting results as human-generated, or using results in a manner intended to convince a natural person that they are communicating with a natural person when they are not;
    - Generate deceptive or misleading digital content or engage in deceptive, abusive behaviors, practices, or campaigns that exploit people due to their age, disability or a specific social or economic situation;
    - Impersonate real entities or create fake personas to falsely attribute content or mislead others about its origin without consent or legal right.

- **Disclosure Requirements**. Under our UP, any individual or entity using Claude in certain types of legal, healthcare, insurance, finance, employment, housing, academic testing, accreditation and admissions, or media or professional journalistic use cases must disclose to customers or end users that Anthropic's services are being used to help inform the decisions or recommendations. Additionally, our UP requires this disclosure where Anthropic's models are deployed in all customer-facing chatbot contexts, including any external-facing or interactive AI agent, and in any products serving minors.

Our UP encourages anyone who discovers inaccurate, biased, or harmful model outputs to notify us, enabling swift corrective action, including suspending or banning offending accounts when necessary.

Engagement and Education Efforts

As global elections approach, we are engaging with policymakers, other companies, and civil society organizations to share Anthropic's election integrity work with the aim of spreading broad awareness of safety interventions that can prevent generative AI from being misused by bad actors in the upcoming elections. For example, we are briefing European Commission staff on

[6] *Anthropic Usage Policy* (June 6, 2024). Available at: https://www.anthropic.com/legal/aup (Accessed May 24, 2024).

election integrity research and interventions in advance of June's EU elections. Additionally, we have been in touch with US civil society organizations and policymakers to share our election intervention updates and will continue to share relevant updates in the months leading up to the election.

Conclusion

The history of AI development has been characterized by rapid advancements and novel applications. We expect that 2024 will bring forth new uses of AI systems, which is why we are proactively building methods to identify and monitor novel uses of our systems as they emerge. We will communicate openly and frankly about what we discover. Thank you for your leadership on this issue.

Sincerely,

Dr. Dario Amodei
Chief Executive Officer
Anthropic PBC