stability.ai

May 24, 2024

stability.ai

The Honorable Mark R. Warner
United States Senator
703 Hart Senate Office Building
Washington, D.C. 20510

Dear Senator Warner,

Thank you for your letter dated May 14. Stability AI is committed to the safe development and safe deployment of AI, and we were pleased to support the Tech Accord to Combat Deceptive Use of AI in 2024 Elections (Tech Accord). The Tech Accord recognizes that there are no silver bullets to prevent the misuse of AI, and acknowledges that “the deceptive use of AI is not only a technical challenge, but a political, social, and ethical” one too. However, there are layers of mitigation across the supply chain – from model developers to system deployers to content distributors – that can help to mitigate risks such as election disinformation or manipulation.

Stability AI is a global company working to amplify human intelligence by making foundational AI technology accessible to all. Our small team consists of passionate researchers and developers producing AI models and AI tools across a range of modalities. These include image, language, audio, video, and 3D. Since the Tech Accord, we have released a variety of models and services, including compact language models for conversational or coding tasks (e.g. Stable LM 2 and Stable Code), applications to generate high-quality soundtracks (e.g. Stable Audio 2.0), models for rapid 3D rendering (e.g. Stable Video 3D), and multimodal chatbots (e.g. Stable Assistant).

All actors have a role to play in promoting the responsible use of AI tools. We continue to explore mitigations for emerging risks in our models and services, consistent with the Tech Accord, and we welcome the opportunity to outline our progress below. We share your commitment to safety and security in AI, and we would be pleased to continue discussing these efforts with your team.

Sincerely,

Shan Shan Wong
Interim Co-Chief Executive Officer
Stability AI

Christian LaForte
Interim Co-Chief Executive Officer
Stability AI

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?*

Stability AI is proactively implementing a range of features to mitigate the spread of unintentional misinformation and intentional disinformation. For example, we have implemented content credentials to help users and content platforms better identify AI-generated content. Images generated through our application programming interface (API) are tagged with metadata to indicate the content was produced with an AI tool. In partnership with the Content Authenticity Initiative (CAI) led by Adobe, we are adopting the Coalition for Content Provenance and Authenticity (C2PA) standard for metadata.[1] This metadata includes the model name and version number used to generate the content. Once the metadata is generated, it is digitally sealed with a cryptographic Stability AI certificate and stored in the file.

We are studying possible ways to apply imperceptible watermarks to images and video generated through our API, and we are considering a range of options for future best practices.

We continue to explore new techniques to improve the robustness of these features, and we continue to engage with other actors in the supply chain, including content distributors such as Google and Meta. Content distributors – such as social media, search, or streaming platforms – play an outsized role in the dissemination of harmful content, regardless of whether it is generated with or without AI tools. We encourage these platforms to use metadata, watermarks, classifier scores, and other signals to assess the provenance of content before amplifying it through their network.

Since the Tech Accord, we have implemented new features in our Stable Assistant and Stable Artisan applications to help prevent the generation of credible election disinformation, such as prompt rewriting. For example, high-risk prompts featuring known candidates can be intercepted and modified to prevent the system generating a photorealistic or recognizable image of that person. These measures can help to prevent bad actors from misusing AI services to produce disinformation at scale.

In addition, we have conducted a comprehensive review of the licenses that apply to our technology. Our earlier image models, including Stable Diffusion 2 and SDXL, remain subject to ethical use licenses that prohibit misuse for a variety of unlawful or exploitative purposes.[2] Our latest image and video models – including Stable Video Diffusion and Stable Diffusion 3 – and our applications and APIs are subject to our updated Acceptable Use Policy (AUP).[3] The AUP elaborates on categories of prohibited use, including: generating or promoting disinformation,

[1] CAI, 'C2PA', available here.

[2] See e.g. the Open Responsible AI License (OpenRAIL) for Stable Diffusion 2 and SDXL, prohibiting a range of unlawful or misleading uses, available here.

[3] See the Stability AI Acceptable Use Policy, available here (updated March 1, 2024).

unlawful impersonation, the production of defamatory content, generating political advertisements, or misrepresenting AI outputs as human-generated.

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?*

Stability AI is a small team of researchers and developers, and there are limitations on our ability to conduct broad public campaigns. However, we continue to deepen public understanding of AI technology and risks through our partnerships with the US AI Safety Institute Consortium (AISIC), CAI, and the AI Alliance, as well as our bilateral engagement with other actors in the AI supply chain. We continue to engage widely with authorities in the US, European Union, United Kingdom and elsewhere about AI developments, risks, and oversight (see question 4 below).

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?*

Stability AI has contributed resources and expertise to the AISIC, and we have partnered with CAI. These forums are spearheading the development of guidance, standards, tools, and evaluation frameworks for content provenance and detection. We expect these efforts will support a range of organizations, including media and nonprofit observatories.

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?*

Stability AI has responded proactively to public requests for comment from agencies such as the Federal Election Commission,[4] and continues to engage widely with Congress and the Administration. Stability AI has testified in multiple Senate hearings on emerging AI risks, where we publicly welcomed clear legal guardrails governing the misuse of AI, including for abuse, fraud, or election manipulation.[5] Stability AI has engaged widely across other jurisdictions to discuss the risks of AI generated content and the role of content transparency, including the European Commission, the European Parliament, the European Council, the UK Government, the UK Parliament, the Singapore Infocomm Media and Development Authority, and the Australian eSafety Commissioner.

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?*

Stability AI continues to monitor developments in content detection.

6. *(To the extent your company offers social media or other content distribution platforms) What*

[4] Stability AI, Comment on Petition for Rulemaking on AI, October 2023, available here.

[5] See e..g. Stability AI, Senate Judiciary Committee, Subcommittee on Intellectual Property, July 2023, available here; Stability AI, Senate AI Insight Forum, November 2023, available here.

kinds of internal classifiers and detection measures are you developing to identify machine-generated or machine-manipulated content? To what extent do these measures depend on collaboration or contributions from generative AI vendors?

Not applicable.

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?*

Not applicable.

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?*

Stability AI maintains a safety hotline at safety@stability.ai for reports of misuse. We investigate and respond to these reports promptly.

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?*

In addition to our partnerships with the AISIC and CAI, Stability AI has established relationships with the National Center for Missing and Exploited Children, Thorn, and the Internet Watch Foundation (UK) to support the detection and reporting of unlawful content. With Thorn and other industry partners, Stability AI has developed a set of Safety By Design principles that define possible mitigations across the supply chain to help prevent the production and distribution of unlawful child sexual abuse material.[6] In addition, Stability AI has engaged with agencies such as GCHQ in the UK to improve information sharing on emerging threats, pursuant to our commitments under the UK Government's *Joint Statement on Tackling Child Sexual Abuse in the Age of AI*.

Stability AI is open to exploring a common database of violative content, across multiple harm categories, in collaboration with other firms and relevant agencies.

[6] Thorn, 'Thorn and All Tech Is Human Forge Generative AI Principles with AI Leaders to Enact Strong Child Safety Commitments', April 2024, available here.