OpenAI

1455 3rd Street,
San Francisco, CA 94158

December 20, 2024

From:
Chan Park
Head of U.S. and Canada
Policy and Partnerships,
OpenAI

To:
The Honorable Jeanne Shaheen
United States Senate
Washington, DC 20510

The Honorable Rick Scott
United States Senate
Washington, DC 20510

CC:
Senator Joe Manchin III
Senator Mazie K. Hirono
Senator Margaret Wood Hassan
Senator Catherine Cortez Masto
Senator Ron Wyden
Senator Angus S. King, Jr.
Senator Richard Blumenthal
Senator Marsha Blackburn
Senator Robert P. Casey, Jr.
Senator Marco Rubio

Dear Senator Shaheen and Senator Scott,

Thank you for your letter concerning non-consensual intimate imagery (NCII) and for highlighting the importance of preventing its creation and dissemination. OpenAI appreciates your commitment to safeguarding individuals from the devastating impacts of NCII.

At OpenAI, we're building artificial intelligence that helps people solve hard problems. By helping with the hard problems, AI can benefit the most people possible – through better healthcare and education, more scientific discoveries, better public policies and services, improved productivity, and new tools for creativity. Our work begins with research, we develop products to make benefits real for people, and we maintain a focus on safety throughout.

Our Approach to CSAM and NCII

OpenAI prioritizes the responsible development and deployment of AI technologies and we have implemented robust safeguards to combat abuse, including the creation of child sexual abuse material (CSAM) and NCII.

We have made significant efforts to minimize the potential for our models to generate content that harms children and actively engage with the National Center for Missing and Exploited Children (NCMEC), the Tech Coalition, and other government and industry stakeholders on child protection issues and enhancements to reporting mechanisms.

As you know, OpenAI is not a social media platform and the tools for addressing abusive content are unique to generative AI developers. To this end, we filter and remove unwanted content, including sexually explicit material, during model training and continue these efforts post-deployment. We partnered with Thorn's Safer to detect, review and report CSAM to NCMEC if users attempt to upload it to our image tools. Between January to June 2024, we reported 3,252 pieces of content and 947 CyberTipline Reports to NCMEC and we publish child safety reports on our website. In March, during testimony at a House Oversight and Accountability Subcommittee hearing, John Shehan, Vice President of NCMEC's Exploited Children Division, acknowledged that we engage in meaningful efforts to detect, report, and prevent child sexual exploitation.

We are proud to have signed onto a number of commitments, including Thorn's Safety by Design Principles and the White House Voluntary Commitments to Combat Image-Based Sexual Abuse. Together, these commitments encourage a proactive and coordinated effort to counteract the misuse of generative AI and foster a world in which generative AI tools can be trusted. The frameworks emphasize the importance of ongoing collaboration among private, public and civil society to create safer digital environments, and address many of the risks in the development and deployment of AI technologies, especially on women and girls.

We also make our moderation application programming interface publicly available and free. This is a tool anyone can use to preemptively filter user inputs and to scrutinize LLM-generated responses for appropriateness. As a leader in this space, we believe that making this tool available for wide consumption can help educate and empower developers with robust tools to prevent all types of abuse, including in CSAM and NCII.

We have implemented NCMEC's Take It Down program. We apply Take It Down's hash filter to all image uploads across all products unless customers meet stringent requirements permitting them to be exempt from scanning. Additionally, we continue to maintain safeguards intended to stop the generation of erotic, sexual, or otherwise adult images, including but not limited to NCII. For example, our models are trained to refuse a wide variety of requests that violate our policies, including a policy that prohibits the generation of erotic content across all of our tools in our usage policy and outlined in the Model Spec.

Sora and Advanced Safeguards

Our previously mentioned efforts are applicable to all of our technology, including Sora, our video generation model. Sora was first previewed earlier this year as we shared our research progress. On December 9, we made Sora available to a wider creative community. Because Sora, and video generation technology generally, presents novel challenges, we have a multi-tiered safety stack to prevent the generation of explicit content, including some mitigations specifically built for Sora.

We're taking an iterative approach to enabling users to generate videos using real people's uploaded photos or videos due to potential misuse. While early feedback from artists highlights its creative value, we're limiting access initially to a small group of users, and the ability to upload images or videos of people will only be made available to a subset of users to start. We will continue to actively monitor this limitation to adjust our approach to safety. In addition, our classifiers, which detect disallowed content, are particularly conservative for image/video uploads and we have dedicated red teamers to specifically test the ability to create sexual deepfakes. We have active, in-depth monitoring in place to understand the value of it to the Sora community and to adjust our approach to safety as we learn.

We have additional protections for both text-to-video and image/video-to-video content that involves minors. We developed a classifier to analyze text and images to predict whether an individual under the age of 18 is depicted or if the accompanying caption references a minor. Uploads containing images of minors will not be permitted during our test of the image/video-to-video feature. If the subject of a text-to-video Sora video is determined to be under 18, we enforce much stricter thresholds for moderation related to sexual, violent or self-harm content. We are also embedding Sora-generated content with C2PA metadata to enable social media platforms to label Sora videos as AI generated content. We will continue to monitor and enforce our policies that prohibit the use of someone's likeness without their permission, including for NCII.

Along with the launch of Sora, we announced a publicly accessible online form to report potential policy violations or illegal content, and are working to integrate this reporting into our tools.

We recognize that there is still significant work to be done to ensure AI is developed and deployed responsibly. We look forward to remaining in touch with you about NCII and CSAM, and we stand ready to assist in efforts to prevent abusive AI generated content.

Sincerely,

Chan Park

Chan Park
Head of U.S. and Canada Policy and Partnerships
OpenAI