Carl Szabo
Vice President & General Counsel, NetChoice
1401 K Street NW, Ste 502
Washington, DC 20005

NetChoice
Defending Free Speech and Free Enterprise Online

April 16, 2024

March 12, 2024: Subcommittee on Innovation hearing titled "*Addressing Real Harm Done by Deepfakes*" Questions from Rep. Nick Langworthy

Please find below the answers from Carl Szabo to the *March 12, 2024: Subcommittee on Innovation hearing titled "Addressing Real Harm Done by Deepfakes"* Questions from Rep. Nick Langworthy

Are you aware of any generative AI models that are created with guardrails in place so they can't produce C-SAM or gruesome content? All of them.

Generative AI companies are taking several key steps to prevent their systems from producing child sexual abuse material (CSAM), and to comply with existing laws to prevent child exploitation:

1. Assessing models before access – a.k.a "red teaming."

Generative AI deployers like Amazon,[1] Meta,[2] Anthropic,[3] and OpenAI[4] carefully assess their AI models for the potential to generate CSAM before allowing them to be hosted on their platforms. This process, called red teaming, involves stress testing their systems to find flaws, weaknesses, gaps, and edge cases that could allow a user to produce synthetic CSAM on their service.[5] For models found to have this potential, they are restricting access or not hosting them until mitigations are in place.[6]

2. Detecting abusive content in inputs and outputs.

Generative AI companies implement robust processes to detect abusive content in both the inputs provided to their AI models as well as the outputs generated. This includes having clear procedures for reporting and destroying any CSAM that is identified.[7]

[1] Amazon Bedrock abuse detection policy
[2] Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations | Research - AI at Meta
[3] Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned \ Anthropic
[4] OpenAI's red team: the experts hired to 'break' ChatGPT
[5] Reducing the Risk of Synthetic Content: Preventing generative AI from producing child sexual abuse material.
[6] OpenAI - Usage policies (Updated Jan. 2024); CSAM Detection and Reporting | Anthropic Help Center.
[7] *Id.*

In fact, Large Language Models like GPT-4 are increasingly used to *improve*[8] existing classifiers that identify and remove CSAM on the internet.

3. Training employees on CSAM handling.

The National Institute of Standards and Technology has released guidance, embraced by major technology companies, encouraging training employees involved in model training on the proper procedures for handling user reports and CSAM, in compliance with legal obligations.

5. Restricting model access.

For models that are assessed as high-risk for CSAM generation, companies are restricting access to hosted-generation only, rather than allowing users to fine-tune the models themselves.[9]

6. Collaborating and advocating.

Companies are collaborating with organizations like Thorn and CHILD USA to develop industry-wide standards and best practices for preventing the misuse of generative AI for CSAM.[10]

By taking these proactive, multidisciplinary steps, generative AI companies are working to minimize the possibility of their technologies being exploited by predators to further sexual harms against children.

Is it feasible to hold tech companies accountable for the content generated on their systems?

The question of holding tech companies accountable for user-generated content is complex and multifaceted. As general use tools, online platforms are utilized by billions of people worldwide, with the vast majority of content being positive and beneficial. Imposing broad legal liability on the creators of these tools for the actions of a small number of bad actors would be misguided and infeasible.

In *Sony Corp. of America v. Universal City Studios, Inc.*,[11] the US Supreme Court recognized that the creators of general use technologies should not be held liable for users' infringing activities, as such liability would stifle innovation and commerce. This principle extends to online platforms, which are akin to phones, email, or even paper and pen. We do not hold telecommunications companies responsible if criminals use a phone to plan a crime, nor do we hold paper manufacturers liable for threatening letters.

Imposing platform liability for all user content would lead to over-removal of legitimate speech, as companies would have to severely restrict what can be posted to mitigate legal risks. This would

[8] How Thorn's CSAM classifier uses artificial intelligence to build a safer internet

[9] *Id.*

[10] NIST - Reducing the Risk of Synthetic Content: Preventing generative AI from producing child sexual abuse material - THORN Comments.

[11] *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984).

undermine the free exchange of ideas online, a principle the Supreme Court has upheld as protected by the First Amendment.[12]

Even if liability was limited to illegal content, it is practically impossible for platforms to proactively police billions of posts in real-time.[13] Automated content filters are imperfect and cannot make the nuanced judgments that courts can.

While legally shielded from liability, major tech companies do invest heavily in content moderation. However, given the volume of content and complexity of human communication, no system is foolproof. Overly strict liability could stifle free speech, reduce platform diversity, and encourage invasive user monitoring.

A more balanced approach is to encourage best practices and provide regulatory guidance on specific types of harmful content. The Supreme Court has recognized this, noting that the remedy for harmful speech is more speech, not enforced silence.[14]

In conclusion, while seeking accountability is understandable, imposing broad liability on tech companies for user content would undermine free speech and innovation online. A nuanced approach, recognizing the roles of both platforms and users, is crucial.

* * *

We thank you for the opportunity to testify before your committee. As ever, we offer ourselves as a resource to discuss any of these issues with you in further detail, and we appreciate the opportunity to provide the committee with our thoughts on this important matter.

Sincerely,

Carl Szabo
Vice President & General Counsel
NetChoice

NetChoice is a trade association that works to make the internet safe for free enterprise and free expression.

[12] *Reno v. ACLU*, 521 U.S. 844 (1997).

[13] *Zeran v. America Online, Inc.,* 129 F.3d 327 (4th Cir. 1997).

[14] *United States v. Alvarez*, 567 U.S. 709 (2012).