Re: ***Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence***

Background

Control AI is grateful for this opportunity to provide comments in response to this request for information.

Control AI is a non-profit research and advocacy organization focused on global security risks from advanced AI systems. Our work to date has focused on policy recommendations for the UK AI Safety Summit, potential international treaties to promote international cooperation, the EU AI Act, and the rising impacts of deepfakes. We have presented our work to relevant government bodies (such as the UK AI Safety Institute), and our work has been featured in various publications (such as Time Magazine, Bloomberg, and the Guardian). Control AI is also a member of the Campaign to Ban Deepfakes, a coalition aiming to reduce growing threats from AI-generated synthetic content.

Our team has recently been preparing a policy report focused on policies that could reduce threats from deepfakes. We focus on regulations that could be applied across the supply chain– including for AI model developers, AI service providers, and cloud compute providers. We are excited to see that NIST will be playing a leading role in reducing the risk of synthetic content. In this response, we summarize relevant sections from our report. The recommendations from our report could serve as the basis for industry standards that prevent the creation and dissemination of harmful synthetic content from AI.

We have attached a draft of our full policy report. Below, we include sections of the report that may be especially relevant to NIST. While the language of the report focuses on *regulation*, we believe our recommendations are just as relevant for attempts to create *standards*, and we look forward to following NIST's work in this area.

Threats from synthetic content

Deepfakes are a serious and growing threat to individuals, institutions, and governments. Deepfakes are commonly used to generate non-consensual pornography, including sexual content of children. 99% of the victims targeted by deepfake sexual content are women, and 99.6% of explicit images of children are of females (Home Security Heroes, 2023; Internet Watch Foundation, 2023). Victims of deepfake pornography and sexual abuse often describe feeling sexually objectified, losing control over how their bodies are portrayed, and experiencing extreme levels of distress (Law Commission, 2021). 79% of business leaders believe that deepfakes are a threat to their business, and over 1 in 3 businesses have experienced deepfake fraud (Regula, 2023). In addition to the clear dangers to individuals and businesses, many experts are concerned about the potential impacts of deepfakes on government institutions. Deepfakes can be used to

create falsified videos of elected officials, sway the results of elections, blackmail political candidates, or spread propaganda during international conflicts. For example, a deepfake of Ukrainian President Volodymyr Zelensky appeared to direct Ukrainian soldiers to surrender to Russian forces (Congressional Research Service, 2023). The Department of Homeland Security has warned that deepfakes pose major threats to national security, and the Department of Defense has described risks from fraud, phishing, and other forms of cybercrime (Department of Homeland Security, 2021; Department of Defense, 2023).

Addressing multiple parts of the supply chain

Effective regulation must address multiple parts of the deepfake supply chain. Deepfakes can be created by anyone around the world. Once the software capable of generating deepfakes is downloaded, it cannot be withdrawn. Therefore, effective and enforceable deepfake policies would need to not only punish creators of deepfakes but also hold companies liable if they fail to take reasonable steps to prevent deepfakes. This includes provisions for model developers (e.g., applying techniques to prevent models from generating deepfakes) as well as model providers and compute providers (e.g., applying techniques to identify users that are trying to create deepfakes and restricting access from malicious users). This approach is also consistent with the idea of "systemic regulation" in AI, which has highlighted the need to regulate AI as a technology as opposed to focusing solely on its downstream applications (see Arbel et al., 2023).

Standards for model developers

Model developers should be liable for negligence if they fail to take reasonable steps to prevent models from creating deepfakes. Model developers have a responsibility to design models using techniques that prevent them from creating deepfakes. Such measures could include techniques that reduce a model's ability to generate deepfake sexual material or fraudulent content, implement techniques that cause models to refuse requests to generate deepfake sexual material or fraudulent content, and show that such techniques cannot be easily circumvented. A relevant precedent can be found in hardware manufacturers: manufacturers must ensure GPS-enabling chips no longer function when used above certain speeds (to prevent their dual-use in weaponry).

Additionally, model developers should have to guarantee that the datasets they use to train their model do not contain illegal material (e.g., child sexual abuse material). Recently, a team of researchers discovered child sexual abuse material in LAION-5B, a large image dataset that was used to train many image models – including the popular Stable Diffusion model (Stanford Internet Observatory, 2023). In response, the organization responsible for LAION-5B withdrew its dataset. This decision was entirely voluntary; it ought to be incorporated into law, and model developers should be liable unless they use reasonable techniques to ensure that their models are not trained on datasets with illegal content.

Standards for model providers and compute providers

Model providers and compute providers should be liable for negligence if they fail to take reasonable steps to monitor how their resources are used and prevent users from creating deepfakes. Model providers and compute providers have a responsibility to ensure that malicious actors are detected and prevented from creating deepfakes. Such measures could include techniques to detect users who are trying to create deepfake sexual material or fraudulent content, ensuring that model access is provided via an Application Programming Interface (API) whose access can be withdrawn, requiring that users register with a verified account, and other techniques designed to detect malicious users and restrict their access. A relevant precedent can be found in banking: banks are required to monitor and prevent customers from engaging in money. To prevent money laundering, financial institutions have to follow "Know Your Customer" (KYC) regulations that require them to verify the identity of their customers, assess the risk that customers may be involved in illegal activities, monitor client transactions, and report illegal activities to relevant authorities. In AI, OctoML (a compute provider) ceased providing computational resources to CivitAI following concerns about its use for creating child sexual material (Multiplatform.ai, 2023). This decision was entirely voluntary, and it ought to be incorporated into law.

Issues with watermarking and labeling synthetic content

Watermarking. Watermarking involves embedding a statistical signal into AI-generated material, such that the material can be detected as AI-generated. Ideally, watermarking would be able to ensure that all AI-generated content can be detected, thus allowing society to clearly distinguish between AI-generated content and content that is not generated by AI. At first glance, watermarking appears to be a promising intervention. However, it suffers from two key problems. First, many deepfakes can cause harm even if they are labeled as AI-generated. Many websites that distribute deepfake sexual images make it clear that they are AI-generated, but these deepfakes still cause harm to those depicted. As an example, even if a non-consensual pornographic video of a woman starts with a disclaimer noting that the video was AI-generated, she may still experience feelings of sexual objectification, shame, and high levels of distress. Second, recent research from a team at Harvard University has shown that robust watermarking is impossible (Zhang et al., 2023a). The researchers were able to show that watermarks can be trivially eliminated, noting:

> "The bottom line is that watermarking schemes will likely not be able to resist attacks from a determined adversary. Moreover, we expect that this balance will only shift in the attacker's favor as model capabilities increase. Future regulations should be based on realistic assessments of what watermarking schemes are and are not able to achieve." – Zhang et al., 2023b

Robust watermarking is not possible, and even if it were possible, it would not address many of the harms from deepfakes. There are many technical challenges of watermarking that currently limit its viability (e.g., Saberi et al., 2023; Zhang et al., 2023a). As such, watermarking should not be considered a solution to problems caused by deepfakes. Watermarking can be incorporated

alongside other interventions to address deepfakes, but they should not be seen as a substitute for meaningful regulation.

Conclusion

We appreciate this opportunity to comment on this significant and timely topic. We look forward to continuing to contribute to discussions around how to best reduce risks from AI-generated synthetic content, and we are eager to follow NIST's work in this critical area.

Sincerely,

Andrea Miotti
Executive Director
Control AI
andrea@controlai.com

Akash Wasil
AI Policy Researcher
Control AI
akash@controlai.com