PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
CITP.PRINCETON.EDU

Reducing harm from deepfakes

Sayash Kapoor and Arvind Narayanan
Princeton University
March 4, 2024

Members of the Assembly Committee on Science, Innovation, and Technology,

Thank you for the opportunity to speak to you on this important topic. I am Sayash Kapoor, a computer science Ph.D. candidate and a researcher at Princeton University's Center for Information Technology Policy. My testimony today is co-authored with Arvind Narayanan, professor of computer science and director of the Center for Information Technology Policy at Princeton University. This statement is written in our personal capacities.

We need better evidence of the harm resulting from deepfakes

The deepfake technology unit is a laudable idea. In addition to its currently outlined responsibilities, the unit should collect and publicly share aggregate data on the prevalence of various types of deepfake harms, such as non-consensual intimate imagery (NCII) and scams, and the harm resulting from them. This is essential for understanding the scope and severity of the problem.

To the best of our knowledge, the last quantitative estimate on the amount of deepfake NCII was made available in October 2020, when a report estimated that over a hundred thousand deepfake NCII images had been created and shared on private messaging app Telegram.[1] Notably, generative AI technology has progressed drastically since the report was published.

[1] Henry Ajder, Giorgio Patrini, and Francesco Cavalli, “Automating Image Abuse: Deepfake Bots on Telegram,” *Sensity*, October 2020.

While the past few months have brought to light many reports of NCII created using AI,[2] we still do not have quantitative estimates of how prevalent deepfake NCII is today. The lack of data makes it hard to advocate for more urgent action by various stakeholders. Perhaps as a result, the problem of AI-generated NCII has not been taken very seriously until recently. That changed when reports of Taylor Swift NCII made the headlines.[3] But Taylor Swift is the tip of the iceberg, and we don't know how big the iceberg is.

On the other hand, for financial scams, we do not yet know if voice-cloning technology makes these scams more effective or prevalent, compared to earlier scams that involve impersonation or social engineering and do not rely on technology.

For example, the FTC reported that 76 million dollars were lost to friends and family imposter scams in 2022. In 2023, a year when there were many news headlines about how newly available voice cloning made this scam much easier, this number was 68 million dollars—a *decrease* of 8 million dollars.[4] It's hard to know for sure what's going on, because there is no data on how often voice cloning technology was used in those scams. It is possible that we have a false impression of a scary new technology based on a few news headlines, when the scam is in fact an old one.

In short, when a problem is known only from anecdotes, it is easy to either underestimate or overestimate it. But if we can quantify it, it allows stakeholders including legislators, technologists, and academics to prioritize efforts in curbing these harms and allocate resources appropriately.

Policing AI developers is infeasible and ineffective

[2] Sayash Kapoor and Arvind Narayanan, "How to Prepare for the Deluge of Generative AI on Social Media," *Knight First Amendment Institute* 23, no. 04 (June 2023), http://knightcolumbia.org/content/how-to-prepare-for-the-deluge-of-generative-ai-on-social-media.

[3] Kate Conger and John Yoon, "Explicit Deepfake Images of Taylor Swift Elude Safeguards and Swamp Social Media," *The New York Times*, January 26, 2024, sec. Arts, https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html.

[4] Tableau Public, "Fraud Reports by the Federal Trade Commission," February 8, 2024, https://public.tableau.com/app/profile/federal.trade.commission/viz/FraudReports/SubcategoryPaymentContact.

We applaud the bills for focusing on the real problem, which is the misuse of AI for creating deepfakes. While it may initially seem more effective to try to restrict or ban the release of AI, there are a few reasons to be cautious of this approach.

First, AI models that are widely available on the internet are already good enough to create believable imagery or audio to impersonate people, so restrictions on future releases might do little to prevent image- and voice-based impersonations.

Second, AI models for creating images and audio can be run locally (e.g., on personal computing devices like laptops) instead of relying on cloud service providers (such as Google Cloud or Amazon AWS). As a result, tracking illicit uses is hard or impossible.

Third, preventing models from being released openly would prevent many legitimate uses from being realized. Often, deceptive or malicious uses of AI are context-dependent: a voice snippet used to extort someone is harmful, yet if that same snippet is used to narrate a book, it improves accessibility for visually impaired readers.

All of these reasons make proposals aimed at AI developers infeasible to implement and ineffective at curbing the real harm.[5]

Content provenance standards can help prove that an image is real

A major concern about the widespread access and low cost of deepfake technology is that media depicting real events can be falsely accused as AI generated, which casts doubt on the truthfulness of real events.[6] For example, in the context of a lawsuit or a prosecution, it may make it harder to verify that the evidence presented before a court is real.

Standards for content provenance can help validate the source of an image. Provenance is a technology that allows manufacturers of media recording devices such as cameras

[5] Sayash Kapoor and Rishi Bommasani et al., “On the Societal Impact of Open Foundation Models,” February 27, 2024, https://crfm.stanford.edu/open-fms/paper.pdf.

[6] Danielle Citron and Robert Chesney, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” *California Law Review* 107, no. 6 (December 1, 2019): 1753.

to add metadata to media.[7] This metadata is cryptographically bound to the media, so if the media is later tampered with or edited, that can be detected.

For example, a camera manufacturer could add metadata that certifies an image was captured using their device and was not edited later. If the correct metadata accompanies an image, you can be reasonably certain that it was created by a human. This helps prevent concerns about falsely accusing real images of being AI generated.

Note, however, that the absence of content provenance information does not tell us anything about the source of the image—such as whether it is human or AI-generated—because content provenance metadata can always be removed or might not have been generated in the first place.

To be clear, content provenance is not bulletproof. For example, someone might click a photograph of an AI-generated image using a camera that supports content provenance. This shows how content provenance might also be an imperfect solution. So before relying on provenance in consequential settings, we first need to test its effectiveness at scale.

Three ideas beyond the scope of the current legislative proposals

Beyond the current scope of the bills, we wanted to share three ideas for reducing the harms from deepfake technology.

First, deepfakes depicting non-consensual intimate imagery (or NCII) are often distributed on websites specifically aimed for that purpose. For example, a report from 404 Media found that AI startup CivitAI allows users to post "bounties" to purchase AI models that create NCII of specific people.[8] These platforms amplify the harms from the technology and present another potential chokepoint for curbing misuse.

[7] “Coalition for Content Provenance and Authenticity,” https://c2pa.org/.

[8] Emanuel Maiberg, “Giant AI Platform Introduces ‘Bounties’ for Deepfakes of Real People,” 404 Media, November 13, 2023, https://www.404media.co/giant-ai-platform-introduces-bounties-for-nonconsensual-images-of-real-people/.

Second, several non profits have worked on developing pathways to prevent harm from non-consensual deepfakes. StopNCII.org allows adults to share data about their non-consensual images.[9] The National Center for Missing and Exploited Children (NCMEC)'s Take It Down initiative allows children to do the same.[10] If these images are later uploaded to participating social media platforms, the non profits coordinate with the social media platform to remove the images as soon as they are uploaded. This prevents the images from being shared or seen widely.

However, there is no requirement for social media platforms to coordinate with these non profits, and as a result, many of them are not involved with these efforts. For example, YouTube and X (formerly Twitter) do not participate in StopNCII.org or NCMEC's Take It Down, leaving them without the information required to proactively remove NCII.

Third, with the widespread use of generative AI, NCMEC has seen a sharp uptick in the number of AI-generated images reported to their database.[11] Once an image is uploaded to NCMEC, the organization has two main responsibilities. It helps platforms remove these images, and it also helps start police investigations about the children depicted in the image. On this second responsibility, NCMEC faces a startling concern as deepfakes become more realistic. Resources for investigating children facing harm could instead be diverted toward identifying children who do not exist and are only present in AI-generated images.

The presence of a separate channel to report AI-generated images could help alleviate this concern.

Conclusion

We thank the members of the committee for inviting us to testify on this important concern and we are grateful for your timely action to curb harms from deepfakes.

[9] "Stop Non-Consensual Intimate Image Abuse," https://stopncii.org/.

[10] "Take It Down," https://takeitdown.ncmec.org/.

[11] Drew Harwell, "AI-Generated Child Sex Images Spawn New Nightmare for the Web," *Washington Post*, June 23, 2023, https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images.