DOCKET #NIST-2024-0001
NIST AI 100-4
National Institute of Standards and Technology
Herbert C. Hoover Building, 1401 Constitution Ave. NW
Washington, DC 20230

May 31, 2024

Re: Comment on NIST AI 100-4

The Center for Democracy & Technology (CDT) respectfully submits these comments in response to NIST's request for comment on its draft report, "Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency" ("Draft Report").[1] CDT is a nonprofit 501(c)(3) organization that works to advance civil rights and civil liberties in the digital age. Among our priorities, CDT advocates for the responsible and equitable design, deployment, and use of new technologies such as artificial intelligence, and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems.

Generative AI creates new opportunities for creativity, scientific advancement, and efficiency across modalities, industries, and platforms. It also comes with significant risks, however, and as its adoption increases, our efforts to mitigate those risks must as well. Synthetic content labeling and detection mechanisms may be quite useful for identifying the authenticity of content and could aid in mitigating some of the harms that flow from new generative AI technologies that can produce realistic content across modalities. These tools require cautious implementation, however, to ensure they are maximally effective and human rights, including free expression and privacy, are protected.

The Draft Report provides a nuanced and thorough overview of the risks and benefits of existing and developing direct and indirect content-labeling techniques, as well as synthetic content detection methods. The Draft Report also correctly emphasizes that each approach discussed in the report has limitations and that "none of these techniques can be considered as comprehensive solutions[.]"[2] Overall, CDT believes the contents of this report will assist government, industry, and civil society organizations as they develop strategies for risk mitigation going forward.

---

[1] National Institute of Standards & Technology, Draft Report: Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency (Apr. 2024) https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf. [hereinafter "Draft Report"]
[2] *Id*. at 45.

CDT submits this comment to make three overarching recommendations for the final report. First, CDT recommends that the report provide a path forward through a multi-stakeholder process that will more fully address the complexity and scale of coordination required to effectively implement the techniques discussed in the report and protect individual rights. The current draft covers the benefits and risks of each individual technique and duly notes the difficulty with implementing each, but the report does not currently attempt to discuss what a comprehensive approach to labeling synthetic content at scale and across platforms would look like or point to the necessary multi-stakeholder engagement over time needed to create such an approach. Second, the report frequently calls for additional research. CDT recommends that the report also identify entities that should be providing independent researchers access to data to accomplish the ambitious research agenda the report lays out. Finally, CDT appreciates the thorough discussion of detection and mitigation strategies for Child Sexual Abuse Material (CSAM) and Non-Consensual Intimate Images (NCII). CDT offers a few clarifying suggestions for expanding and adding detail to that section of the report.

Scaling and Coordination Challenges

The report begins by advising that a risk-based and human-centered approach to synthetic content detection and labeling will be critical to success.[3] Given the degree of complexity with implementation of many of the labeling and detection techniques the report reviews, as well as their significant limitations, synthetic content detection or labeling will not be a cure-all, and each individual technique could not be implemented by a single stakeholder effectively.[4] Instead, with the right safeguards, managed expectations, and transparency measures, these techniques could represent important tools for harm reduction if coordination across necessary stakeholders is a priority from the outset and human rights, particularly privacy and free expression, are front and center. To that end, the report, in addition to outlining the specific benefits and risks of each individual technique, should also analyze the complexity of using these tools across production and distribution platforms at scale and recommend a process of multi-stakeholder engagement to ensure the needed coordination and oversight to maximize benefits while minimizing human rights risks.

The report repeatedly raises three concerns with implementing the labeling and detection techniques it analyzes: implementation at scale, protecting privacy, and ensuring equitable outcomes. It also points out that coordination among multiple stakeholders and additional research are needed to implement many of the techniques effectively.

As the report thoroughly discusses, different tools are more effective for certain types of content, but are ineffective for others, creating challenges for implementing them at scale. For example, the report details that indirect watermarking is most effective for digital images, though it still has

---

[3] *Id*. at 3

[4] See also, Access Now, Identifying Generative AI Content: When and How Watermarking Can Help Uphold Human Rights (Sept. 2023), https://www.accessnow.org/wp-content/uploads/2023/09/Identifying-generative-AI-content-when-and-how-watermarking-can-help-uphold-human-rights.pdf.

significant limitations even for that type of content, including the potential for removal or alteration, complicating implementation.[5] On the other hand, watermarking is far less effective for synthetic text and is particularly ineffective for academic writing and non-english languages. Indirect watermarking therefore can be done, but only somewhat effectively for certain content, and even then, not without challenges.[6] The limitation on text watermarking for certain types of content also highlights equity concerns with the implementation of watermarking in AI-generated text that must be addressed.

Similar issues of differential effectiveness and implementation at scale arise with synthetic content detection techniques. Synthetic image detection can be relatively accurate when searching for content made by a specific image generator, but far less so when attempting to detect synthetic content made by a variety of image generators.[7] Synthetic video detection technologies may perform well in training environments, but far less so in real world scenarios where video quality and other characteristics vary significantly.[8] Ensuring detection technologies work equally effectively across skin tones and genders will require robust investment in representative training and testing datasets.[9] Effectiveness in real world scenarios is also a challenge for detecting synthetic audio. Additionally, most audio detection methods are designed and tested in English.[10] This limitation again points to the risk that deployment of some of these techniques may leave non-English speaking communities out of risk mitigation plans, or, worse, could disproportionately mis-identify their speech as synthetic if tools fail to account for language differences.

The report fairly and accurately accounts for these challenges, and more, in detail. From the draft, it is clear that implementation of these techniques at scale across multiple platforms for generating and distributing content could compound the risks the report identifies. Error rates for detection and labeling could balloon, with authentic content labeled incorrectly and synthetic content going unaddressed. Bad actors also could easily evade or corrupt some interventions, further undermining their effectiveness. Haphazard or poorly coordinated approaches to instituting synthetic content detection and labeling techniques could also harm free expression and user trust.

The report should more directly discuss how all stakeholders, including Generative AI developers, Generative AI deployers, social media services, web search providers, app stores, open source software platforms, technologists, civil society, researchers, and technology users, will need to work together to optimize efforts to mitigate risks posed by synthetic content. Tools to detect and label content should be interoperable for distributor platforms and they should be

---

[5] Draft Report, *supra* note 1, at 20
[6] *Id*.
[7] *Id*. at 25
[8] *Id*. at 27.
[9] See Hibaq Farah, Deepfake detection tools must work with dark skin tones, experts warn, The Guardian (Aug. 17, 2023), https://www.theguardian.com/technology/2023/aug/17/deepfake-detection-tools-must-work-with-dark-skin-tones-experts-warn.
[10] Draft Report, *supra* note 1, at 32

legible to the users who are intended to benefit from them. For each intervention, multiple actors will need to participate to ensure maximal benefit while minimizing collateral risks. The report should call for that collaboration as a step toward responsible implementation.

The report should also point to potential processes, procedures, and spaces where this multistakeholder engagement can thrive. Mitigating risks posed by synthetic content will be an iterative process, requiring ongoing adjustment as technology changes and methods of evasion improve. Continued engagement across stakeholders will help ensure effective interventions that are equitable across differences in content and in end users. No single platform, search engine, or generative AI company can accomplish this task alone and even together their efforts would be poorly informed without the perspective of their users, researchers, and civil society expertise. The report should note the need for the creation of a space for meaningful and accountable multi-stakeholder engagement as a next step toward risk mitigation.

Researcher Access to Data

The draft report calls for additional research into various issues no less than eight times.[11] Potential subjects for research include watermarking's impacts on public perception of the content, how certain techniques can be abused by adversarial actors across modalities, and the effectiveness of synthetic content labeling and detection in reducing the harms of CSAM and NCII. These are all worthy and necessary areas of exploration. The National Plan to End Gender Based Violence report from the White House also recently called for additional research into technology-enabled gender based violence.[12] Similar calls were made by the U.S.-E.U. Trade and Technology Council in its Joint Principles on Combating Gender based Violence in the Digital Environment.[13]

While it is useful for companies to leverage their own resources to produce reliable research on these subjects, in this context, where coordination across stakeholders is critical to success, independent research would produce the most reliable and translatable results. In order to conduct such research, researchers need access to relevant data and information about the AI systems that generate synthetic content, the user-generated services where that content can be shared, and the technologies for labeling and detection. Unfortunately, the types of data necessary to conduct this research has in recent years become more difficult to access.[14] For

---

[11] Draft report, *supra* note 1, at 14, 15, 19, 21, 27, 31, 35, 42.
[12] White House Task Force to Address Online Harassment and Abuse, Final Report and Blueprint, at 15 (2024) https://www.whitehouse.gov/wp-content/uploads/2024/05/White-House-Task-Force-to-Address-Online-Harassment-and-Abuse_FINAL.pdf.
[13] U.S.- EU Trade and Technology Council (TTC), Joint Principles on Combatting Gender based Violence in the Digital Environment (Apr. 5, 2024) https://digital-strategy.ec.europa.eu/en/library/us-eu-trade-and-technology-council-ttc-joint-principles-combatting-gender-based-violence-digital.
[14] See Morten et al., Berkeley Technology Law Journal, Researcher Access to Social Media Data: Lessons from Clinical Trial Data Sharing, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4716353; Shapiro et al., "New Approaches to Platform Data Research", https://www.netgainpartnership.org/resources/2021/2/25/new-approaches-to-platform-data-research

example, platforms are restricting access to tools that facilitate analysis of how content travels over their services.[15] Meta is deprecating its CrowdTangle tool, which had been critical to researchers seeking to understand online harms.[16] The removal of these tools has, among other things, made it more difficult for researchers who study and attempt to mitigate CSAM networks.[17]

There are silver linings, however. The Digital Services Act in the European Union contains requirements for researchers to be able to access data to further societal understanding of how platforms influence our daily lives.[18] As online services subject to the DSA begin to comply with this provision, there may be opportunities to leverage those researcher access programs for projects outside of the EU.[19] Similarly, policymakers here in the United States, including members of Congress and Administration officials, are increasingly recognizing the importance of independent research and considering policies to support it.

Researcher access to AI systems is also limited. For example, there is strong evidence that some generative AI systems have CSAM in their training data, and can and have been used to generate CSAM.[20] Researchers are interested in getting access to the models and data used to train generative AI models in order to get a sense of how they could be used, including for the generation of CSAM and NCII.[21] Yet this problem too may be getting more difficult to research. More than 300 AI researchers cosigned a letter claiming that companies have suspended accounts and even threatened legal reprisal to good faith researchers seeking to understand the capabilities and potential impacts of generative AI systems.[22] And AI companies provide even less information about how users are actually using their systems.[23]

The Draft Report identifies numerous avenues for future research but it does not specify how that research might be conducted or which entities should be involved in producing the needed analyses. The draft report could be improved by noting that generative AI developers and

---

[15] Sarah Greve Gotfredson, Q&A: What Happened to Academic Research on Twitter, Columbia Journalism Review (Dec. 6, 2023). https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php

[16] Civil Society Letter to Meta: https://foundation.mozilla.org/en/campaigns/open-letter-to-meta-support-crowdtangle-through-2024-and-maintain-crowdtangle-approach/.

[17] David Thiel, Renee DiResta, Alex Stamos, Cross-Platform Dynamics of Self-Generated CSAM (June 7, 2023), https://purl.stanford.edu/jd797tp7663.

[18] European Union Digital Services Act, Article 40.

[19] Platform Transparency Tools & the Brussels Effect: https://iddp.gwu.edu/platform-transparency-tools-brussels-effect.

[20] David Thiel, Jeffrey Hancock, Identifying and Eliminating CSAM in Generative ML Training Data and Models (Dec. 2023) https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

[21] E.g., Gebru et al., 2018; Mitchell et al., 2018

[22] https://sites.mit.edu/ai-safe-harbor/

[23] Aylin Caliskan and Kristian Lum (2024) https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purpose-ai/; Zeve Sanderson, Josh Tucker (2024) https://www.techpolicy.press/beyond-red-teaming-facilitating-user-based-data-donation-to-study-generative-ai/.

deployers, social media platforms, and other industry actors should provide independent researchers in academia and civil society with access to data and other information needed to carry out the types of research called for in the report and work with them to design relevant research projects, while also ensuring privacy and human rights are protected.

CSAM / NCII

The draft report specifically examines methods for preventing and mitigating the harm in the creation and distribution of synthetic Child Sexual Abuse Material (CSAM) and non-consensual intimate images (NCII).[24] The harms related to both categories of content are quite similar, but the interventions and risk mitigations appropriate to address them may, at times, be quite different. The report notes some of these complexities when it acknowledges that it can be difficult to determine whether consent existed when analyzing whether a particular image is NCII. However, the draft report also at times assumes that interventions that may be appropriate for CSAM are also appropriate for NCII without accounting for these added complexities. Globally, CDT recommends disaggregating interventions for CSAM and NCII and addressing each category of content separately to allow for the requisite nuance in the use of each tool to be more clearly presented.

- *Training data filtering*

Broadly the report outlines the benefits and risks of filtering training datasets to ensure that they do not contain CSAM or NCII. The report argues that filtering of training data could be a method of preventing models from generating CSAM and NCII.[25]

With respect to CSAM, the report appropriately notes the risk inherent in building systems that would detect unknown CSAM (e.g., CSAM not already present in a hashed database) through trained classifiers because merely possessing such content in a classifier data set may be a crime.[26] Safe harbors for engaging in such detection and filtering may need to be created to ensure effective removal of CSAM from training data.

By suggesting that removing NCII from training data sets could help ensure that models cannot be used to generate NCII, the report over-indexes an intervention that might be useful for reducing CSAM harms[27] to imply it could also be helpful for NCII. Removing known-NCII (i.e., content which a developer knows the person depicted did not consent to its production or distribution) from training data is a best practice worth encouraging. However, further suggesting that content filtering tools for training data might be used to attempt to identify and remove from training data novel-NCII or sexual content generally, which can often be indistinguishable from

[24] Draft Report, *supra* note 1, at 36-42.
[25] *Id*. 36.
[26] *Id*. at 36.
[27] See Grossman, Shelby, Pfefferkorn, Riana, et. al, The Strengths and Weaknesses of the Online Child Safety Ecosystem at 76 (Apr. 22, 2024) (recommending platforms use hash-based CSAM detection systems and establish and integrate content provenance and authenticity standards) https://purl.stanford.edu/pr592kc5483.

NCII, could create problems for free expression and for the usefulness of model outputs.[28] For those reasons, CDT has significant concerns with that suggestion. As the report rightly points out, removal of all sexual content from a training data set could degrade model output, including educational content, medically accurate content, consensually created sexual images, and entirely AI-generated sexual content that does not depict real persons. It could also suppress particular speech or degrade use cases involving educational, artistic, or scientific content. For NCII specifically, determination of consent is non-trivial. It may not be clear from a particular image whether the person depicted consented to the creation or to the disclosure of the image. Classifiers that attempt to guess whether consent is present could lead to over-censorship of sexually-related content, degrading model output and jeopardizing important use cases.

- *Image input / output filtering*

Image input and output filtering are common-sense post training interventions, but they must be implemented carefully to ensure free expression is protected. To that end, the report notes that merely possessing a classifier that attempts to identify CSAM that would be necessary to facilitate such filtering presents legal risks, as noted above, and safe harbors may need to be designed to permit this use case.[29] The report further notes that classifiers struggle to identify novel forms of abuse, e.g., unknown NCII. However, this type of output filtering might not be desirable as a method of preventing NCII because such a filter would likely have no way to assess whether the creation and distribution of the image was consensual and could suppress other legal and desirable use cases.

The report could be improved by exploring whether interstitial messages presented when certain input or output filters are triggered could be implemented to introduce friction into the process of creating content that might turn out to be synthetic CSAM or NCII, potentially mitigating some risk. The report could also explore whether generative AI services do or could ask whether persons intended to be depicted in a sexual image consented to the depiction before generating such an image. Assuming that verifiable consent is too burdensome to obtain, merely asking whether a person seeking to generate an image had consent to do so would be an imperfect and, likely, easily circumvented intervention, but could potentially reduce NCII production in models that employ it. More research is likely needed to understand whether these interventions would reduce the creation of NCII.

- *Hashing Confirmed Images*

As noted above, developers can use methods to filter CSAM and known NCII out of their training data to a certain degree. Similar techniques, known as hashing, can be used to detect known CSAM and known NCII once they have been distributed. Once an image has been identified and hashed, the hash can be used to identify the image across various platforms. The report rightly notes that there is a need for coordination among entities as they become aware of

[28] Draft Report, *supra* note 1, at 37.
[29] *Id*. at 39.

CSAM and NCII and create hashes of that content and that there is an equally pressing need to protect the privacy of hash databases for that content to prevent re-traumatizing victims.[30]

The report notes the need for coordination among stakeholders to address the problem, but does not analyze existing efforts. CDT suggests that the report should specifically address existing shared hash databases for NCII and CSAM, including Take It Down[31] and Stop NCII.[32] The report should outline the challenges inherent in these efforts and discuss methods by which these efforts could be improved to more meaningfully achieve their goals of removing harmful content while protecting free expression and privacy, including through greater transparency regarding implementation and mulitstakeholder engagement.

- *Provenance and Data tracking*

Finally, the report notes that all of the concerns and challenges of provenance and data tracking mentioned earlier apply with equal force to CSAM and NCII.[33] CDT would also note that provenance and data tracking might be useful for identifying synthetic CSAM or synthetic NCII and might be helpful in tracking that harm, but it will not mitigate the harm of the distribution of that content nor will it help to mitigate the harms of the distribution of authentic CSAM and NCII. Being able to detect synthetic versions of this content may have some value in some cases, but it will be far from the only intervention necessary to address the inherent harms in the creation and distribution of this content.

Conclusion

The Draft Report is a thorough and thoughtful overview of the risks and benefits of synthetic content labeling and detection techniques. CDT appreciates the opportunity to offer these suggestions for additional improvement to the report and looks forward to the publication of the final version. Please contact Kate Ruane, kruane@cdt.org, with any questions.

[30] *Id*. at 40.
[31] Take It Down, https://takeitdown.ncmec.org/.
[32] Stop NCII, https://stopncii.org/.
[33] Draft Report, *supra* note 1, at 41.