PARTNERSHIP ON AI

Partnership on AI's comments on NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency

Background

Partnership on AI (PAI) is a nonprofit partnership of academic, civil society, industry, and media organizations creating solutions to ensure that AI advances positive outcomes for people and society. PAI studies and formulates sociotechnical approaches aimed at achieving the responsible development of artificial intelligence (AI) and machine learning (ML) technologies. Today, we connect over 100 partner organizations in 14 countries to be a uniting force for the responsible development and fielding of AI technologies.

PAI develops tools, recommendations, and other resources by inviting multistakeholder voices from across the AI community and beyond to share insights that can be synthesized into actionable guidance. We then work to promote adoption in practice, inform public policy, and advance public understanding. We are not an industry or trade group nor an advocacy organization. We aim to change practice, inform policy, and advance understanding.

The information in this document is provided by PAI and is not intended to reflect the view of any particular Partner organization of PAI. The comments provided herein are intended to provide evidence-based information, based on PAI's research, about several aspects of the draft NIST report.

Comments

PAI welcomes the release of NIST's draft report, and the opportunity to provide this further input to NIST's work under section 4.5 of the Executive Order on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (the "AI Executive Order"). PAI previously provided a written response to NIST's earlier RFI, and convened a listening session for PAI stakeholders to inform NIST's work on this issue. This submission does not repeat the content of that session and our RFI response, but provides some brief comments on several aspects of the draft report.

PAI submits that NIST's report on Reducing the Risks Posed by Synthetic Content should be amended to address the following:

1. **Consistently using the terms Direct and Indirect Disclosure to emphasize social implications.**

   We make 2 proposed recommendations:

   a. **Ensure consistent use of, and distinction between, the concepts of Direct and Indirect Disclosure**
   b. **Amend the infographic in Figure 1 on Page 6, to explicitly reference the difference between Direct and Indirect Disclosure methods and explain the difference between them more clearly.**

   PAI welcomes the clear explanation of, and distinction between, the concepts of Direct and Indirect Disclosure contained in Section 3 of the draft report (Page 5). This terminology mirrors the language PAI has used in our work on this issue. However, this language is not always used throughout the draft report. **It would be helpful for the report to use this terminology consistently**, including in the sections describing the different types of disclosure methods and the infographics/figures throughout the document. This would be a foundation for drawing out more clearly the social implications of the technical methods described in the report, by grounding a discussion of the fact that human audiences will ultimately be the people encountering *Direct* Disclosures drawn from *Indirect* Disclosures. The draft report contains (in the "Additional Issues for Consideration" section) some discussion of how people interact with digital content transparency approaches; there is an opportunity to embed this idea throughout the report by describing (where relevant) that interaction as Direct Disclosure informed by Indirect Disclosure. This approach would be helpful specifically throughout the report's assessment of the more technical (Indirect Disclosure) methods discussed.

   Further, while NIST emphasizes the technical focus of the draft report, it is crucial to not lose sight of the clear connection between how Indirect Disclosure ultimately informs Direct Disclosures that contribute to audience understanding and content authenticity. This is integral to the impact of this report on the societal challenges it's trying to address.

2. **Breaking out Different Types of Metadata.**

   NIST's draft report identifies the different types of metadata currently used to differentiate between authentic and synthetic/manipulated content.

   We make 2 proposed recommendations:

   a. **Discussion of cryptographically-signed metadata should be separated into a standalone section, with additional detail**. This is one of the most discussed metadata-based methods for providing disclosure. While it is currently addressed in the draft report, given its importance in public discourse as a standalone tool for providing disclosure, PAI believes it warrants separate, more detailed treatment — both its limitations and opportunities. This section should also address the complementary role this category of metadata can play in broader standards.

   b. **NIST should include a more detailed breakdown of the different types of metadata included in the chart on Page 16.** This should include a precise analysis of the impact and privacy, security, trustworthiness and integrity, and management and quality issues for consideration.

3. **Completeness and Clarity of the "Additional Issues for Consideration" and "Testing and Evaluation" sections: highlighting social impact.**

   We make 3 proposed recommendations:

   a. **Elaborate on social challenges and opportunities associated with Indirect Disclosure mechanisms (e.g. Privacy, Trustworthiness and Integrity) and set out key questions that could inform work on alleviating them.** This should include where NIST might need to collaborate with wider US government agencies to address these issues, and ensure social factors and considerations are integrated into NIST's work program on synthetic content.

   b. **Ensure NIST's synthetic content workstream identifies issues warranting further research and collaborates with the newly established NIST ARIA where appropriate.**

   c. **Further development of the "Trustworthiness and Integrity" heading** (see proposed inclusions below e.g. comparing the efficacy of described Direct Disclosure signals to context tools).

   The draft report does a great job of providing many of the technical risks and opportunities associated with various Indirect Disclosure mechanisms, and includes some high-level commentary about their social implications (e.g. Privacy, Trustworthiness and Integrity) that connect clearly to Direct Disclosure. We encourage NIST to further elaborate on some of these social challenges, the types of institutions that might be able to play a role in alleviating them, and key questions that should inform such work.

Sections 3.1.1.2 ("Additional Issues for Consideration" for digital watermarking) and 3.1.2.3 ("Additional Issues for Consideration" for metadata recording) are two particular sections that discuss issues warranting further research (which could potentially involve collaboration with the recently announced NIST ARIA (Assessing Risks and Impacts of AI) Program that will focus on sociotechnical testing.) The discussion of these issues could be further refined in section 4.1 (which addresses Testing and Evaluating Provenance Data Tracking and Synthetic Content Detection Techniques). Some specific examples of opportunities for further development that could be highlighted under the "Trustworthiness and Integrity" heading in section 3.1.2.3 are:

- Evaluating whether or not signals of content synthesis with AI, without any identity signal, support audience understanding of content.
- Testing and capturing how much real-world content features inaccurate and manipulated Indirect and Direct Disclosure signals (e.g. a fake provenance label from C2PA on an image; doctored watermark embedded in a file).
- Comparing the efficacy of described Direct Disclosure signals (content labels, visible watermarks, and disclosure fields) to other sets of Context tools, like community notes and additional methods for providing context about content.
- Developing a benchmark for how to monitor and gauge the deployment of manipulated Indirect and Direct Disclosure methods.

4. Reference to emerging challenges related to human-AI distinguishability, not just synthetic content distinguishability.

We make 1 proposed recommendation:

a. Include references related to complementary transparency methods that not only convey if content has been AI-generated, but if it comes from an authentic human – emphasizing privacy.

Consistent with the requirements in section 4.5 of the AI Executive Order, the draft report discusses standards, tools, methods, and practices for authenticating content, tracking its provenance, and labeling and detecting synthetic content.

The draft report clearly describes the importance of methods for distinguishing if content has been altered or AI-generated. In the Summary section, there is reference to the importance of "asserting ownership of content.” However, there is room for the report to highlight another distinct societal challenge that will be exacerbated by AI: how people can know they are interacting with a real human, while emphasizing privacy preservation. How might we build complementary transparency methods that not only convey if content has been AI-generated, but if it comes from an authentic human? This is emerging as a key concern across civil society, industry, media, and beyond. Referencing this distinct issue more clearly in the paper will better prepare the field to investigate key questions related to identity and personhood, with a focus on privacy, on the web.

Conclusion

PAI would be happy to provide further information about any of the matters discussed in this submission. We look forward to NIST’s final report.

For any further information and questions related to this submission, please contact claire@partnershiponai.org and policy@partnershiponai.org.