Comments of the World Privacy Forum

Regarding

NIST AI 100-4 Draft for Public Comment, Reducing Risks Posed by Synthetic Content, An Overview of Technical Approaches to Digital Content Transparency

Sent via email to: NIST-AI-100-4@nist.gov

ATTN:
National Institute of Standards and Technology
100 Bureau Drive Mail Stop 8900
Gaithersburg MD 20899–8900

2 June 2024

The World Privacy Forum appreciates the opportunity to comment on the work underway at NIST to establish standards for detecting, authenticating, labeling, and tracking the provenance of synthetic content.[1] These comments respond to *NIST Draft AI 100-4, Draft for Public Comment, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*, published April 2024 on the NIST website, available at: https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf.

The World Privacy Forum is a non-partisan 501(c)(3) public interest research group focused on conducting research, analysis, and education in the area of privacy and complex data ecosystems and their governance, including in the areas of identity, AI, health, and others. WPF works extensively on data governance and privacy across multiple jurisdictions, including the U.S., India, Africa, Asia, the EU, and additional jurisdictions. For more than 20 years WPF has written in-depth, influential research regarding systemic data issues. These include medical identity theft, India's Aadhaar identity ecosystem, and an early influential report on machine learning and consumer scores (*The Scoring of America, 2014*). Most recently, WPF published *Risky Analysis,* a 2023 report on AI Governance Tools that establishes the beginnings of an evaluative environment for these tools. WPF co-chairs the UN Statistics Data Governance and

[1] NIST AI 100-4 Draft for Public Comment, *Reducing Risks Posed by Synthetic Content, An Overview of Technical Approaches to Digital Content Transparency*, (April 2024), https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf.

Legal Frameworks working group. At the OECD, WPF researchers participate in the OECD.AI AI Expert Groups, among other activities. WPF participated in the core group of AI experts that collaborated to write the OECD Recommendation on Artificial Intelligence, now widely viewed as the leading normative principles regarding AI. WPF research on complex data ecosystems governance has been presented at the National Academies of Science, the Mongolian National Academies of Science, and the Royal Academies of Science. See our reports and other data at World Privacy Forum: https://www.worldprivacyforum.org.

The data governance and privacy implications of synthetic content as expressed in various types of systems are significant, and have not yet been adequately measured or mapped, and for the most part the larger ecosystem remains largely ungoverned.[2] Early research suggests that the synthetic data detection, authentication, labeling and provenance tracking ecosystem is likely to create new forms of digital identifiers and identity signals, and also could create new, potentially negative downstream impacts including new ways to expose sensitive dataThese comments discuss several aspects of the challenges, with a particular focus on the standards development process itself and issues related to metadata, among other issues.

I. Normative codes of conduct for Standards Development Organizations need to be applied in the NIST AISIC development context

As the foundational aspects of AI governance and policy are designed, standards bodies have a central role in facilitating the creation of fair, ethically developed standards that abide by well-understood standards development codes of conduct and openness mechanisms.[3] WPF has concerns regarding ad hoc standards developed outside of SDO codes of conduct such as those established at ISO, among other entities. Quasi-standards and methods created primarily by entities that may be subject to the those very standards or evaluation processes should be assessed by normative ethical guidelines created in an environment of non-dominance. WPF understands the need for urgent work on solutions regarding content provenance and synthetic data. However, we are concerned about a race to solutions without an adequate evaluative environment that fully documents, tests, and explores the reliability of such techniques in an inclusive way.

One key aspect to this normative work, discussed later in these comments, can be characterized as the need to establish "meta measurement," or metrology … measuring the measurements. This would facilitate the establishment of evaluative environments that fully document, test, and explore the infrastructure of the reliability of synthetic content detection, authentication, labeling, and provenance tracking techniques as well as the reliability of related governance and privacy methods.[4] These types of infra-evaluative environments will need meaningful involvement from multistakeholder groups representing the data use, governance,

---

[2] NIST AI RMF, NIST, https://www.nist.gov/itl/ai-risk-management-framework.

[3] For example, see: *ISO Code of Ethics and Conduct*, as approved under Council Resolution 11/2023, adopted on 23 February 2023, ISO. https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100011.pdf . Specific NIST ethical guidance extends beyond standards development, but the guidance is still helpful: *US Department of Commerce NIST Summary of Ethics Rules*, U.S. Department of Commerce, 2022. https://www.commerce.gov/sites/default/files/2024-04/nist-summary_of_ethics_rules-2022_0.pdf .

[4] See Section VII of these comment regarding Metrology. See also: Márcio José Sembay, Douglas Dyllon Jeronimo de Macedo, Laércio Pioli Júnior, Regina Maria Maciel Braga, and Antonio Sarasa-Cabezuelo, *Provenance Data Management in Health Information Systems: A Systematic Literature Review,* J Pers Med. 2023 Jun; 13(6): 991. Published online 2023 Jun 13. Doi: 10.3390/jpm13060991 .

and privacy-related considerations of civil society, among a wide range of national and international stakeholders, including vulnerable and often under-represented populations, and other communities, including Indigenous communities. Ensuring that the normative ethical principles of conduct for Standards Development Organizations are applied —such as those from NIST, ISO, IEEE, and ANSI, among other similar SDOs — will be necessary to achieve good performative results.

WPF urges NIST to set up an internal team to ensure that normative standards are in place as the work at AISIC proceeds,[5] and that the evaluation of the application of normative codes of conduct in the resulting standards development is an integral part of the process. The fairness of the development process itself needs to be measured and made transparent.

II. Comments on Metadata, responsive to NIST Draft AI 100-4, Sections 3 and 4

A. Metadata identifiers and the important role of existing identity standards and codes of conduct

NIST's recognition of the need for further research "to understand how metadata recording may impact user privacy and security, security of the metadata itself" (pages 18-19) is welcome. As noted in NIST's research, metadata generated by, used or shared in systems that detect, authenticate, label and track provenance of synthetic content could include a variety of sensitive data such as personal or group-related identifiers,[6] geographic locations visited by content creators, and other data that could be used to reveal identity. The metadata flowing through these systems could potentially be shareable and accessible to multiple downstream parties, depending on the architecture of the system and its guiding legal framework, if / when applicable. Various stakeholders have different interests in downstream metadata; WPF suggests that these interests be mapped, quantified and addressed through robust socio-technical frameworks, including standards, and quite possibly legal frameworks when standards are insufficient.

For example, data indicating precise locations and times when an individual or group created or altered a content file such as a job application, immigration applications, medical care — even MRI content,[7] financial forms, and other types of content could be embedded into a

---

[5] NIST AISIC, https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic .

[6] Group related identifiers are a meaningful risk in metadata uses. WPF notes that, for example, medical forms of research as well as other research that may fall outside of the Common Rule or other regulations may have significant implications for the collective privacy of groups, including protected groups that have experienced stigmatization. See W. Nicholson Price II and Glenn Cohen, *Privacy in the age of medical big data,* Nature Medicine 25, 37–43 (2019). https://doi.org/10.1038/s41591-018-0272-7. See also: Alessandro Mantellerò, *From Group Privacy to Collective Privacy: Towards a New Dimension of Privacy and Data Protection in the Big Data Era*. In: Taylor, L., Floridi, L., van der Sloot, B. (eds) 2017. Group Privacy. Philosophical Studies Series, vol 126. Springer, Cham. https://doi.org/10.1007/978-3-319-46608-8_8. See also: *All of Us Research Program Tribal Consultation Final Report March 2021*, National Institutes of Health. March 2021. https://allofus.nih.gov/all-us-research-program-tribal-consultation-final-report. See also Pam Dixon, forthcoming: *Collective Privacy Rights and Data Sovereignty in Indigenous Approaches to Privacy*, Paper workshopped at the Privacy Law Scholars Conference 31 May 2024. Available upon request to NIST.

[7] Taro Langner, *Machine Learning Techniques for MRI Data Processing at Expanding Scale*, April 2024. Book chapter preprint. https://arxiv.org/abs/2404.14326 .

variety of content files and made accessible to varying degrees.[8] The NIST draft astutely recognizes that watermarking methods enabling changes to content could create verifiable signals that can be identified and extracted. (p. 7, lines 9-13.)

This is not unlike other privacy-related challenges created by digital fingerprinting methods that gather and analyze data such as details of device settings in order to estimate or triangulate user identity.[9] Similarly, the document states that "digital fingerprints, which are hashes that are predictably generated from the content itself, can also be used to generate unique identifiers to which metadata can be associated externally to the content itself." (p. 16, lines 1-18) NIST also rightfully notes that hashing methods or other "encryption schemes and digital signatures are not foolproof." (p. 18, line 37)

Regarding NIST's future research regarding the potential use of unique identifiers in synthetic content detection, authentication, labeling and provenance systems and metadata, it will be helpful to reference existing regulations and guidance on digital identifiers and their uses around the world including those from NIST, OECD, ID4Africa, and others. Guidance needs to encompass the developing and the developed world. [10] [11] [12]

B. Metadata sharing, inaccuracy, and other vulnerabilities

As indicated in NIST Appendix B (p. 71), which lists more than 30 Technical Tools related to digital content transparency, NIST has recognized the possibility that synthetic content detection and tracking could incorporate the use of multiple systems operating in conjunction to form a complex ecosystem (s). For instance, content and data provenance metadata may not be embedded in watermarks, but rather stored in separate systems that are referenced by watermark systems. As a result, content and data provenance metadata may be duplicated and shared across a number of distinct systems, creating multiple windows for legitimate or illegitimate access, sharing and exposure.

---

[8] John C. Simmons, Joseph M. Winograd, *Interoperable Provenance Authentication of Broadcast Media using Open Standards-based Metadata, Watermarking and Cryptography*. Submitted May 2024 to Cryptography and Security (cs.CR); preprint: https://arxiv.org/abs/2405.12336 .

[9] For a discussion of digital fingerprinting, see: Marcos Oliveira, Jonathan Yang, Daniel Griffiths, Denis Bonnay, Juhi Kulshrestha, *Browsing behavior exposes identities on the Web*, Computers and Society, 2023. https://doi.org/10.48550/arXiv.2306.14735 ; See also: Antonio Desiderio, Anna Mancini, Giulio Cimini, and Riccardo Di Clemente, *Recurring patterns in online social media interactions during highly engaging events*, Physics and Society (physics.soc-ph), 26 June 2023. arXiv:2306.14735 .

[10] The ID4D — or ID for Development — movement is one which seeks to provide legal identity to all persons by 2030. See: *United Nations Legal Identity Agenda*, https://unstats.un.org/legal-identity-agenda/ . Although spearheaded by various entities, such as the UN, there are a variety of specific ID4D principles that articulate what constitutes beneficial guardrails on identity. See for example *ID4Africa Code of Ethics*, ID4Africa, https://id4africa.com/code-of-ethics/ . The ID4Africa multistakeholder code specifically articulates how identity must be handled in identity ecosystems at the regional, national, and subnational level in the African region.

[11] *OECD Recommendation on the Governance of Digital Identity*, OECD, June 2023. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0491#backgroundInformation.

[12] *NIST SP 800-63-3 Digital Identity Guidelines*, NIST. June 2017. https://csrc.nist.gov/pubs/sp/800/63/3/upd2/final.

NIST also writes that watermarking model source code could be manipulated or altered to change or falsify data and metadata related settings and controls. (p. 14, lines 31-41). It will be important to establish appropriate processes for evaluating methods and processes used to verify metadata, control access to the metadata, and ensure appropriate data governance and privacy in watermarking model source code.

C. Evaluating governance and privacy methods for synthetic content detection

The methods used to enact data governance and protect data privacy and security often function inside the deep layers of synthetic content detection, authentication, labeling and provenance tracking systems. These may be opaque to a number of ecosystem participants, and will require adequate evaluation through full documentation, testing, and exploration of methods to ensure transparency, robustness, and system reliability. Many data encryption and privacy-enhancing techniques are not always reliable, especially when used out of context. Off-label use of PETs may create meaningful tradeoffs that can negatively impact people, including children and vulnerable or minority groups, including victims of crime.

For example, as the document notes, retrieval-based synthetic content detection methods need further evaluation, as they have been found to be limited in accuracy and to produce meaningful privacy impacts by exposing the previously-generated text data or large language model response data of all users.[13] The NIST document rightfully acknowledges that "user privacy—especially for vulnerable populations—could be at risk" (p.18, lines 12-25) without appropriate metadata methods and controls. WPF sees this problem as one that can be solved, particularly with early attention to developing rigorous testing and evaluative frameworks well-fit for the task. It may be that some content areas need tools specifically adapted to the content; health and medical data are one such category.

D. Use limits and data minimization

As noted by NIST, (pages 18-19) metadata embedded in content may be stripped in order to preserve privacy.[14] Despite calls for some type of opt-in or opt-out controls regarding the use of these systems, the document indicates that there is support for ensuring that the metadata in these systems is not stripped. Thus, the desire to ensure that the metadata in these systems is not "wiped" or "stripped" in order to facilitate effectiveness could be in conflict with the need to reduce identifiable and sensitive data created and shared by them.

Opt in and opt out is an extremely difficult governance mechanism in the metadata context. As discussed in *Risks associated with use of identifiers in synthetic content detection*, (p. 16) authentication, labeling and provenance tracking systems and metadata risks are compounded when considering potential downstream transfer and processing of this information as it moves across borders, complicating approaches to trusted data flows.

There are a number of solutions to this problem. WPF has hope for a technical solution. However, it is possible that a socio-legal governance mechanism (s) will be necessary to mitigate the risks posed. As time goes by, it will be very difficult to police all of the existing

---

[13] Zhengyuan Jiang, Jinghuai Zhang, Neil Zhenqiang Gong, *Evading Watermark based Detection of AI-Generated Content,* In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3576915.3623189 .

[14] See section 3.1.2. of the document addressing Metadata Recording.

content contained in metadata. Solutions such as back end usage limitation through technical forms of data minimization and other procedural and administrative tools need to be considered and carefully and robustly tested for feasibility, applicability, and effectiveness. It is possible that a complex mesh overlay of multiple techniques may need to be in place. This will need to be automated given the scale of many systems.

IV. Biometric data in synthetic content metadata

The NIST draft document does not specifically address biometric data; WPF sees this as an oversight, particularly considering OMB Memorandum M-24-10, which sets an important set of guardrails and a concrete framework for biometrics. Biometrics no matter where they are encountered are an important category of data that could be passed through synthetic content detection and provenance tracking systems. Biometric data such as face print data, fingerprints or iris scans (or combinations thereof) could readily be included in the metadata associated with synthetic content detection and provenance tracking systems and methods.

WPF notes that Office of Management and Budget (OMB) Memorandum M-24-10, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence (March 28, 2024) sets forth how U.S. Federal Agencies and Executive Departments shall govern their use of Artificial Intelligence. In the Memo, OMB specifically sets forth biometric systems as "rights impacting," and sets forth mandatory guidance for such systems. The memorandum implements Executive Order 14110 on AI, which directed its publication.

The memorandum creates extensive AI governance requirements ranging from how procurement of AI systems is conducted to risk assessment of AI and informing the public and giving the public choice in regards to some government uses of AI. The memorandum applies to all agencies as defined in 44 U.S.C and the AI in Government Act of 2020, and excludes elements of the Intelligence Community. The OMB memorandum provides an extensive articulation of emergent guardrails around modern AI. There are many points of interest to discuss, but the most striking includes the thread of biometrics systems guidance throughout the memorandum and continuing on in the associated Fact Sheet and materials. Additionally, the articulation of minimum practices for safety -impacting and rights- impacting AI will likely become important touch points in regulatory discussions in the U.S. and elsewhere.

Regarding the establishment of minimum practices for "Safety-impacting and rights-impacting Artificial Intelligence," the required minimum practices are extensive, and include an expansion of the discussion of AI Impact Assessments and how they are to be conducted throughout the AI lifecycle, notably including assessment of AI systems using metrics, which the memorandum notes should be "quantifiable measures of positive outcomes for the agency's mission," in addition to other measures. WPF's December 2023 report, Risky Analysis: Assessing and improving AI governance tools, discusses such metrics in detail, and provides a global index of these nascent tools and measures.

The OMB memo includes additional minimum practices for rights-impacting AI such as requirements for consulting and incorporating feedback from affected communities and the public. For example:

> "B. Consult and incorporate feedback from affected communities and the public." Consistent with applicable law and governmentwide guidance, agencies must "consult affected communities, including underserved communities, and they must"solicit public feedback, where appropriate, in the design, development, and use of the AI and use such feedback to inform agency decision-making regarding the AI The consultation and feedback process must include seeking input on the "agency's approach to

> implementing the minimum risk management practices established in Section 5(c) of this memorandum, such as applicable opt-out procedures." (Page 22.)

The discussion of opt-out here is notable, because the memorandum explicitly includes the Federal government's use of biometrics as an AI use case that impacts rights. In such cases, Agencies using face recognition and other biometric systems are specifically and clearly directed to take extensive extra steps, which will encompass the expanded AI Impact Assessments, public feedback, monitoring for discrimination, providing notice to affected members of the public, and additional responsibilities.

WPF sees the OMB Memo as establishing an important baseline framework, one which NIST should incorporate in its work. WPF encourages NIST to include biometrics in its report, and cite the OMB framework and also biometric guardrails specifically, as is only fair given the existence of this Memorandum and its relevance.

V. Bias in Watermarking and Synthetic Content Detection

NIST is correct in recognizing the need for further research regarding false positives and false negatives that can occur when using watermarks to authenticate the origin of content. (p. 14, lines 14-44) Recent scholarly research shows that deepfake detection methods vary in accuracy and may be built with imbalanced data of different races and genders that can result in large disparities in predictive performances across races. [15] [16] These errors could perpetuate distrust in the accuracy of watermarks and the watermarking process. Some researchers have begun to address these problems by proposing methods for improving fairness[17] and robustness[18] of existing deepfake detectors.

VI. Risks to human rights

In NIST's evaluation of synthetic content detection, authentication, labeling and provenance tracking methods and systems, it will be essential to document and address the inherent risks to human rights created by these systems in the U.S. and in other jurisdictions where risks may be heightened. Unintended outcomes such as compromising or stifling medical care, stopping the creation and dissemination of content such as religious texts, artwork, or journalistic reports from people fighting government repression — all are a real possibility. Synthetic content authentication systems could also impact children who may not pass age-based content creator requirements.

---

[15] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, *Deepfakes generation and detection: State-of- the-art, open challenges, countermeasures, and way forward, Applied Intelligence*, pp. 1–53, 2022.

[16] Y. Xu, P. Terhoïst, K. Raja, and M. Pedersen, *A comprehensive analysis of AI biases in deepfake detection with massively annotated databases*, arXiv preprint *arXiv:2208.05845*, 2022.

[17] Ju, Yan & Hu, Shu & Jia, Shan & Chen, George & Lyu, Siwei, *Improving Fairness in Deepfake Detection,* 2023 Arxiv, https://arxiv.org/pdf/2306.16635.pdf.

[18] Nadimpalli, A.V., & Rattani, A., *On Improving Cross-dataset Generalization of Deepfake Detectors*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 91-99. https://arxiv.org/abs/2204.04285.

An extensive analysis and modeling of potential harms associated with use of synthetic content detection and associated techniques was published by the Coalition for Content Provenance and Authenticity (C2PA) in 2023.[19] WPF applauds this effort to be transparent about the risks. A follow-on analysis regarding this system from WITNESS -- an organization that trains and equips activists and citizens around the world to use media in their fight for human rights -- found that the system could be used to "enforce journalistic identity in laws in a jurisdiction or demand additional information on media posted on social media" in ways that suppress speech or reduce media diversity. WITNESS also found that data from the C2PA system could be misused to curtail freedom of expression (e.g. political speech)."

Given that data is available regarding human rights impact from the system creators and from observers, WPF urges NIST to address the human rights risks of content provenance systems with rigorous study and the application of effective governance and other socio-technical controls.

VII. Metrology: "Meta Measurement " or evaluation or detection and provenance metrics

Metrology[20] in this context is measuring the evaluation techniques used in synthetic content detection and content authentication according to criteria such as effectiveness, fairness and explainability, including techniques and metrics that generate quantifications and scores [21] used to assess system accuracy, quality and robustness. Metrology will be core to the NIST efforts. Although there is a meaningful literature on metrology, WPF has not seen a substantial literature of metrology regarding evaluation or detection regarding provenance metrics.

Some intriguing literature parallels do exist, particularly in the health context.[22] In an important paper by Brown (2021), he proposes the concept of metrology as "*measuring measurement*." He notes that metrology is often invisible, a form of "infra-technology" supporting the quality infrastructure. WPF agrees with the arguments of the paper that quality control at a meta level is essential to the robustness and health of ecosystems. WPF requests that metrology be adopted as a key area of work in AISIC.

VIII. Meaningful Stakeholder Involvement

---

[19] *C2PA Harm, Misuse, and Abuse Assessment, PHASE II - Initial Adoption* (*Version 1.1 - April, 2023),* C2PA, https://c2pa.org/specifications/specifications/1.0/security/_attachments/Initial_Adoption_Assessment.pdf . See also: *C2PA Harms Modeling*, C2PA, https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html#_harms_considerations_for_c2pa_stakeholders .

[20] *Fundamentals of Metrology,* NIST, Office of Weights and Measures, 2019. https://www.nist.gov/pml/owm/fundamentals-metrology.

[21] *NIST AI 100-4, Draft for public comments, Appendix E.2.. Background: A Common Testing Experiment Framework under Testing and Evaluation.* NIST, April 2024. https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf . Pages 87-90.

[22] Richard J.C. Brown, *Measuring measurement: What is metrology and why does it matter*? Measurement (Land), 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7471758/ .

WPF appreciates the large and diverse stakeholder group that NIST has assembled to work on AI issues through the AISIC. Work by external NGOs, academics, civil society groups, and others to assess and address potential risks associated with synthetic content detection, authentication, labeling, and provenance tracking methods is essential, and adds important and helpful tension to the system, when accomplished well.

Processes for evaluating synthetic content detection, authentication, labeling, and provenance tracking systems, including in relation to metadata governance and privacy methods, is an area of work that needs to be conducted with balanced input from all stakeholders, including users, civil society and other groups affected by these systems. The principles of non-dominance should apply to this work. Any and all standards development in content provenance and synthetic media detection must occur within the ethical guidelines and codes of conduct common to all respected Standards Development Organizations. In addition, informal "standards" built outside of ISO, IEEE, ANSI, and NIST standards of ethical requirements for standards, including regarding conflicts-of-interest, must not be used as prefabricated foundations for formal standards.

IX. Conclusion

WPF appreciates the opportunity to comment on this important area of standards development. As discussed in these comments, synthetic data detection, authentication, labeling, and provenance tracking systems carry with them inherent and downstream risks and potential impacts that could affect people in developed and developing jurisdictions in a variety of challenging ways, including ways that could affect human rights.

The World Privacy Forum stands ready to assist NIST in designing an evaluative environment to test for a range of meaningful concerns regarding these systems. WPF is interested in rigorous standards, as well as rigorous metrology infrastructures. Thank you for NIST's important work in this area.

Respectfully submitted,

Pam Dixon, Executive Director World Privacy Forum
Kate Kaye, Deputy Director World Privacy Forum