COMMENTS OF PUBLIC KNOWLEDGE

Date: June 2, 2024
Docket: NIST-2024-0001
Re: NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile Initial Public Draft;
NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency Draft Report
Author: Nicholas P. Garcia, Policy Counsel, nick@publicknowledge.org

Public Knowledge (PK) thanks the National Institute of Standards and Technology (NIST) for the opportunity to comment on the initial public draft of the Generative Artificial Intelligence (GAI) Profile companion resource for the AI Risk Management Framework (AI RMF). PK is strongly supportive of the consensus-driven, open, transparent, multi-stakeholder approach to the development of the RMF and GAI Profile, and PK hopes to continue to contribute to the evolution of these and other resources for voluntary standards and best practices for the governance, management, and development of AI technologies. Overall, the GAI Profile presents a strong evaluation of existing GAI risks and actions for governing, mapping, measuring, and managing those risks.

Drawing upon PK's expertise in law and policy, especially around intellectual property, privacy, and information systems, these comments primarily focus on the GAI Profile, though NIST's Synthetic Content Report (NIST AI 100-4) also informs these comments and some recommendations apply to both documents. This comment focuses on two main points of feedback:

(1) Intellectual Property (IP) risks have particular qualities distinct from other risks presented in the Profile, and NIST should therefore consider changes to language about, and actions oriented towards, IP risks.

(2) The GAI Profile's Appendix's recommendations on provenance data tracking do not sufficiently address risks to privacy and free expression that may be enabled by provenance tracking techniques or the limitations of data tracking approaches. NIST should directly incorporate limitations discussed in the Synthetic Content Report and point Profile users towards more holistic solutions for addressing information integrity issues beyond data tracking.

(1) Challenges and Opportunities in Addressing IP Risks

Potential IP violations pose a different profile of risk compared to most of the other risks covered in the Profile. Unlike many of the risks of GAI, IP-related risks are almost purely economic, are covered by a very well-developed system of law and policy, present a vector for adversarial risks to AI actors, and can be over-indexed and internalized in a way that collaterally presents significant adverse effects on other critical risk areas.

IP Harms Are Economic; Not Harms to Safety, Rights, or Society.

IP violations are overwhelmingly economic in nature, both for the AI actor and for the potential rightsholder. This stands in strong contrast to the other risk categories where the potential harms to individuals or society are simultaneously more severe and/or abstract. Other GAI risks impact physical safety (as with CBRN information, dangerous or violent recommendations, environmental impact), civil or human rights (as with privacy harms and degrading, abusive, or biased outputs) or undermine fundamental aspects of society (as with information integrity, confabulation, and human-AI configuration risks). IP infringement risks, by contrast, are purely limited to theoretical economic losses for rightsholders (i.e. losses from hypothetical licensing revenues) and quantifiable economic damages and fines that an AI actor might incur as a result of adverse infringement findings.

There is a potential dignitary component to some IP infringement risks (i.e. a creator's desire to have control over their work), but U.S. IP frameworks are not the most suitable policy tool for addressing these concerns. For example, when looking at the question of how to honor content owner preferences regarding web scraping we can see that the prevailing system is voluntary and exists outside of the context of IP law.[1] Similarly, as the Profile notes, harms related to the appropriation of an individual's image or likeness, beyond cases where such likeness has a commercial value, are not currently protected by IP laws.[2] Such dignitary or personal harms to name, image, or likeness appropriation are therefore perhaps best thought of in risk categories related to privacy or obscene, degrading, or abusive content rather than through IP.[3]

This different risk profile should be explicitly addressed because, as described further below, over-indexing on IP risks is very likely given that it is easily internalizable because of ease of quantification and actual legal obligations, and such overemphasis carries with it negative consequences for other GAI risks that are less internalizable.

[1] https://www.robotstxt.org/faq/legal.html
[2] Perry Jackson, *Hey, That's My Voice! – The Significance of the Right of Publicity in the Age of Generative AI*, Public Knowledge (Aug. 14, 2023), https://publicknowledge.org/hey-thats-my-voice/.
[3] *Id.*

IP Risks Are Legal Risks and Thus Strongly Internalizable, Quantifiable, and Predictable.

The main IP risk for AI actors is a finding of infringement. Unlike many of the other risks covered by the profile, this risk flows from a well-established system of law and policy with actual legal obligations. This renders IP risk more predictable and manageable than more unsettled or developing areas of GAI risk management—even accounting for ongoing litigation and potential policy-changes surrounding IP law. Copyright and trademark infringement both provide for statutory damages, making the potential cost of an infringement finding strongly quantifiable compared to the difficult-to-evaluate cost of something like biased outputs or harms to the information environment from misinformation or user overreliance on confabulation.

In addition to being more predictable, it is also critical to highlight that IP risk mitigation is also mandatory and enforceable, whereas many of the other areas of GAI risk covered by the Profile remain the domain of voluntary compliance and best practices. Even where legal compliance is required to address components of risk (such as compliance with existing civil rights, consumer protection, or privacy laws) those compliance costs fall very unevenly across different actors.

IP, as an economic right, also presents direct incentives for rightsholders to aggressively and adversarially attempt to over-enforce their rights to extract economic benefits. There is well-documented history of the Digital Millennium Copyright Act's "notice and takedown" provisions being abused;[4] copyright and trademark trolls pursuing frivolous or opportunistic litigation in pursuit of settlement payouts;[5] and large platforms adopting copyright enforcement policies that disproportionately advantage alleged rightsholders in ways that negatively affect users and free expression.[6]

This legal grounding for IP risk does carry with it distinct advantages compared to other risks: there is a mature and well-developed body of experts that can advise AI actors on their IP risks and mitigation strategies. The cost of consulting legal experts, like potential damages, is also strongly quantifiable and predictable.

[4] Electronic Frontier Foundation, Takedown Hall of Shame, https://www.eff.org/takedowns.

[5] Matthew Sag, *Copyright Trolling, An Empirical Study,* Iowa Law Review (August 24, 2014), available at SSRN: https://ssrn.com/abstract=2404950.

[6] See e.g., Timothy Geigner, *YouTube's Content ID System Flags, Demonetizes Video Of Cat Purring*, Techdirt (Feb. 14, 2022), https://www.techdirt.com/2022/02/14/youtubes-content-id-system-flags-demonetizes-video-cat-purring/.

This combination of obligation and quantification operates to make IP risks strongly internalizable for AI actors; meaning that they can and must build the costs and risks of IP into their operations. In contrast, many of the other risks addressed by the GAI Profile remain, in the absence of clear legal or regulatory guidance, strongly externalizable. The result is that among the risks posed by the Profile, AI actors may **readily overemphasize addressing IP risks to the detriment of other risks** with minimal need for encouragement through resources like the Profile. Indeed, the best use of the Profile may be to point out how to best balance managing IP risks with other associated risk areas.

There Are Significant Collateral Risks Created by Over-Indexing on IP Risks.

An overemphasis on IP risks in GAI risk management will lead to negative consequences in other critical areas.

Firstly, as the GAI Profile acknowledges, diverse and representative training data is crucial for the optimal performance of AI models.[7] When access to a wide range of data is restricted due to IP concerns, it can significantly hamper the ability of AI systems to learn from varied sources, ultimately affecting their effectiveness and the richness of their outputs. This limitation directly impacts the model's ability to mitigate other key GAI risks such as toxicity, bias, homogenization, model collapse, and data memorization. For example, limited training datasets can lead to models that do not adequately represent or understand diverse perspectives or contexts, which can perpetuate biases or lead to toxic outputs.

Publicly accessible information often serves as a vital resource for training AI models. However, overly risk averse approaches to IP can limit the use of such data, thus narrowing the scope of information that AI systems can learn from. This not only affects the diversity and representativeness of AI outputs but also potentially stifles accessibility and representation in AI development (e.g. among lower-resourced or noncommercial actors). By overly focusing on the risks associated with IP, there is a risk of creating an environment where the fear of potential IP infringements may deter researchers and developers from more robustly addressing other critical GAI risks.

IP rights can also operate directly in opposition to free expression rights. Limitations and exceptions to copyright are essential to ensure that free expression rights are protected. Policies that give undue deference and weight to IP protection concerns beyond the bounds of the law thus begin to encumber free expression. For example, GAI systems that are designed to limit outputs in a way that goes beyond preventing directly

[7] MP-2.3-002; MS-2.11-007

infringing works to also prohibit content that bears stylistic similarity to existing works is severely limiting the expression of users beyond what is required by law. While this may be a reasonable design choice based on the principles or preferences of a given AI actor, it should not be encouraged from the perspective of IP risk mitigation because of the risk of such private IP self-policing continuing to balloon and swallow the rights of users.

Specific Recommendations

The GAI Profile's introduction to IP risks should explicitly address the risk of over-indexing on IP compliance and the institutional pitfalls that may lead to that result. It should also discuss the collateral adverse effects of overly-restrictive IP policies.

NIST should explicitly clarify that conducting legal evaluations regarding Fair Use and other limitations and exceptions to copyright that may apply is an acceptable strategy for mitigating IP risks for training data and outputs.

NIST should remove or revise the language after "legal fora" starting on line 3 of page 9. The existing language can be read to incorrectly imply that there is or should be a right to compensation for the use of journalistic content; there is considerable and heated public debate regarding these issues, and the GAI Profile should avoid language that may point towards a particular perspective without more complete analysis.

There are a few references to "open source" AI models, systems, or components. The definition of open source is currently ambiguous in the AI context and NIST should include a document-specific definition in the Glossary to clarify the usage of that term.

Recommendations regarding specific IP-related Actions:

- MP-4.1-017: Use trusted sources for training data that are licensed or open source and ensure that the entity has the legal right for the use of proprietary training data.
    - Revise to "Use trusted sources for training data and conduct an independent legal review of the permissible use of that data."
    - Reasoning: Neither licensing nor being open source is independently determinative of the permissible uses of a given training data set depending on use case. There may also be trusted data sources that are neither licensed nor open source (e.g. a controlled access data set of permissively licensed or public domain material).

- MS-1.1-018: Track the number of training and input data items covered by intellectual property rights (e.g., copyright, trademark, trade secret).
    - Consider cutting or revising to: “Assess or evaluate training and input data sources for data items that may be covered by intellectual property rights (e.g., copyright, trademark, trade secret).”
    - Reasoning: It is incredibly difficult to make an accurate evaluation of whether a given data item is actually covered by IP rights without detailed fact-specific legal analysis. Especially given the scale of training data sets, such a measurement effort aimed at tracking the IP status of each individual component is technically and legally infeasible.

- MS-2.8-001 Compile and communicate statistics on policy violations, take-down requests, intellectual property infringement, and information integrity for organizational GAI systems: Analyze transparency reports across demographic groups, languages groups, and other segments relevant to the deployment context.
    - Revision: insert “alleged” in front of “intellectual property infringement.”
    - Reasoning: As discussed above, IP systems are subject to incentives that may result in significant inaccurate or weaponized use of IP infringement claims to attempt to extract economic value, impermissibly dampen competition, or suppress free speech. The GAI Profile should ensure that claims of IP infringement are viewed with appropriate skepticism and scrutiny, particularly when it comes to analysis and reporting.

- MG-3.1-007: Review GAI training data for CBRN information and intellectual property; scan output for plagiarized, trademarked, patented, licensed, or trade secret material.
    - Revision: change “plagiarized” to “copyright infringing”.
    - Reasoning: Reduce intentionality and anthropomorphization of model outputs.

(2) Privacy, Free Expression, and IP Risks Created by Provenance Data Tracking

The GAI Profile should take a broader view of content provenance issues in the Appendix, and must clearly call for AI actors to evaluate risks to privacy, free expression, and IP raised by implementing content provenance data tracking systems.

In previous comments addressed to NIST regarding the development of this Profile, PK emphasized that “Generative AI tools could be leveraged for the purposes of consumer manipulation, impersonation and deception, and enable never before seen scalability in the production of mis- and disinformation."[8] This remains a critical concern, and it is important that the Profile addresses these risks head-on and presents a variety of actions for addressing these concerns. In the Appendix on primary considerations, the profile specifically addresses Content Provenance, specifically through the lens of provenance data tracking to identify generated or synthetic content.

However, NIST has simultaneously produced a detailed report on synthetic content that accurately identifies additional risks and critical limitations to technical content provenance-based approaches. The GAI Profile should directly incorporate the conclusions of that report:

> "[E]ach [technical approach to digital content transparency] has important limitations that are both technical and social in nature. It is vital to note that none of these techniques can be considered as comprehensive solutions; the value of any given technique is use case and context specific. In order for digital content transparency to succeed, the application of provenance data tracking and synthetic content detection approaches must occur in tandem with various social efforts and initiatives to affirm content authenticity."[9]

As acknowledged in the Synthetic Content Report, provenance data tracking can create risks to user privacy through the inclusion of potentially sensitive information about a system's users. These risks can be exacerbated by measures designed to make provenance authentication systems more resilient and robust, such as by obscuring the existence or collection of metadata or provenance tracking measures from the user, or by using technical measures to make tracking data difficult to remove or alter.

Additionally, provenance data that purports to account for creators, sources, and modifications can create risks from an IP perspective by offering potentially incomplete or inaccurate information that may be used to automate or substantiate inaccurate IP enforcement actions.

Specific Recommendations

PK would advise that for ***both the GAI Profile and the Synthetic Content Report*** updated versions should acknowledge the risks associated with provenance tracking, disclose and discuss the limitations of provenance tracking, encourage minimization of

[8] Public Knowledge Comments on NIST AI EO Actions (Feb. 2, 2024), https://publicknowledge.org/policy/nist-comments-on-ai-executive-order/.
[9] Synthetic Content Report at 45, https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf.

risks, and should recommend careful evaluation on the part of AI actors about if data tracking provides sufficient benefit to warrant the associated risks.

As PK recommended in its previous comments, it is critical to direct AI actors to “focus on robust information ecosystem scale solutions. There is no silver bullet for dealing with disinformation generated by malicious actors, but a combination of trusted and reliable fact-based journalism institutions, user-facing tools for transparency and information, AI and information literacy and education, robust content moderation standards, and real oversight and accountability for the digital platform and AI sectors, would all contribute towards developing a healthier information environment with users that are more resistant to the dangers of malicious disinformation.”[10]

[10] Public Knowledge Comments on NIST AI EO Actions at 4.