U.S. AI Safety Institute
National Institute for Standards and Technology
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, Maryland 20899-2000

Re: NIST AI 800-1 2pd, Managing Misuse Risk for Dual-Use Foundation Models

The Center for Democracy & Technology (CDT) respectfully submits these comments in response to the U.S. AI Safety Institute's (AISI) request for comments on the second public draft of its guidance on Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1 2pd). 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 artificial intelligence (AI), and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems. CDT is also an active member of NIST's AI Safety Institute Consortium.

We applaud AISI's continued commitment to developing concrete, evidence-based guidance for managing the risk of foundation model misuse. **In our view, this updated draft is a marked improvement over AISI's initial public draft.**[1] We are heartened to see that AISI has incorporated several of the themes we emphasized in our comments on that initial draft into this update.[2] We particularly appreciate that, while the focus of this guidance remains on developers, this draft includes clear, actionable recommendations for other actors in the AI value chain as well. Both our earlier comments and prior research have emphasized that actors across the AI value chain must all act responsibly in order to effectively address AI risks.[3] As such, we applaud AISI for recognizing the role of actors other than model developers in AI risk management, and for taking steps toward providing those actors with concrete guidance for minimizing the risk of serious misuse. We are also glad to see this draft give developers more robust guidance on how to weigh the potential benefits of a model against its risks when deciding whether to deploy or continue developing it.

We appreciate that AISI has made the limited scope of its guidelines more explicit, since the document understandably does not aim to address every important risk associated with foundation model development. It does not, for instance, help developers prevent improper biases in their models or the facilitation of unlawful discrimination. Nor does it describe how to

[1] National Institute of Standards and Technology, "Managing Misuse Risk for Dual-Use Foundation Models: Initial Public Draft" (2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[2] *See* "Comments on NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models," *Center for Democracy & Technology* (2024), https://cdt.org/wp-content/uploads/2024/09/Final-Comments-CDT-DS-NIST-800-1-FM-Misuse.pdf.
[3] *Id.*; Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," *arXiv* (2022), https://arxiv.org/abs/2108.07258.

reduce the risk of accidental harms from foundation models.[4] We recognize that it would be impractical for a single document to cover every type of risk, so we thank AISI for explicitly noting this guidance's limited scope. However, we emphasize, as AISI does, that **the exclusion of certain risks from this guidance should not deter developers from continuing to mitigate those risks,** including by drawing on existing guidance from NIST, such as its AI Risk Management Framework and that Framework's Generative AI Profile. It is worth noting, moreover, that many of AISI's recommendations in this guidance can help mitigate risks beyond the specific set of risks to public safety that motivate this guidance. Anticipating the potential risks and impacts of a model in advance, developing an advance plan to manage them, thoroughly testing the model before release, and monitoring for harms after release — all of which are recommended by this guidance — are important mechanisms for managing many risks, not just the public-safety-related misuse risks discussed in this document.

We recommend that AISI continue to clarify and expand on three aspects of this guidance. In particular, AISI should:

1. Emphasize the importance of independent domain experts from diverse disciplines in managing misuse risk.
2. Ensure that when developers are encouraged to produce documentation, they produce artifacts that benefit the stakeholders they are meant for.
3. Clarify the importance of considering privacy during post-deployment monitoring.

I. EMPHASIZING THE IMPORTANCE OF INDEPENDENT DOMAIN EXPERTS FROM DIVERSE DISCIPLINES

One of the strengths of this draft guidance is its repeated insistence that developers' management of misuse risk must be grounded in empirical evidence. As the guidance recognizes, one key source of evidence for developers is input from *domain experts* — researchers with specific expertise in the fields most relevant to the ways in which a model could be misused. The guidance correctly identifies several points in the risk management process where domain experts' involvement would be especially valuable (namely, the creation of threat profiles, the design of capability evaluations, and red-teaming). However, the guidance should further clarify the role of domain experts in two ways.

First, AISI should explicitly encourage developers to seek input from *independent* domain experts at relevant points throughout the risk management process. For a variety of reasons, developers may find it useful to rely on domain experts who are employed or closely affiliated with the developer itself. For instance, creating accurate threat profiles might require access to sensitive information that cannot be shared safely with external parties. Similarly, developers might hesitate to give external red-teamers sensitive details about model design.[5] However, developers should be discouraged from relying *exclusively* on internal experts. Involving independent experts, who — unlike employees — lack a strong interest in the outcome of a risk

[4] Usman Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," *arXiv* (2024), https://arxiv.org/abs/2404.09932.

[5] This draft guidance (rightly) encourages developers to give red teams access to such details in order to realistically simulate certain misuse scenarios.

assessment, provides an important layer of scrutiny and accountability to a developer's risk management processes, as prior CDT research has emphasized.[6] As such, AISI should encourage developers to make use of independent domain experts throughout the misuse risk management process, putting in place protections as needed to guard against disclosure or misuse of any proprietary or sensitive information. Specifically, developers should be encouraged to consult independent experts when creating threat profiles and designing capability evaluations, include independent experts in their red teams, and make their models available to independent evaluators prior to deploying them.

Second, AISI should clarify that domain experts may come from a variety of disciplines, including the social sciences. While technical researchers' instinct may be to rely exclusively on technical domain experts, non-technical and social-scientific experts can also provide important input into misuse risk management. The empirical evidence that informs a developers' threat profiles ought not to be purely technical information about a model's capabilities — equally relevant is information about the *social environment* into which that model would be deployed. For instance, information about the characteristics and likely behavior of rogue actors who might leverage a highly-capable foundation model to conduct large-scale cyberattacks, or information about the most common types of CSAM or NCII and how they are typically generated and disseminated, would be vital for creating empirically-informed threat profiles. Social scientists, more so than technical researchers, are well-suited for gathering this information, and AISI should explicitly recommend that developers make use of their expertise.

II. ENSURING THAT DOCUMENTATION BENEFITS RELEVANT STAKEHOLDERS

A key feature of this draft guidance is that it includes recommended documentation and disclosure practices to correspond to each recommendation it makes. Documentation is a key plank of any AI governance effort, and we are heartened to see the degree of emphasis AISI places on it in this guidance. However, some types of documentation are more effective than others. As past CDT research has emphasized, overly vague or high-level documentation artifacts can easily fail to achieve their intended goals.[7]

In order to ensure that the documentation the guidance recommends creating is most useful, AISI should articulate the specific rationale behind each documentation artifact it recommends that developers create. The rationale behind an artifact significantly affects what form it ought to take: for instance, the optimal format for documentation meant to be shared with external stakeholders is quite different from that for documentation meant for internal record-keeping.[8] By clarifying what it sees as the role of each documentation artifact it recommends, AISI can help developers produce these artifacts in the manner most likely to support the intended goals.

---

[6] Miranda Bogen, *Assessing AI: Surveying the Spectrum of Approaches to Auditing and Understanding AI Systems*, *Center for Democracy & Technology* (2025), https://cdt.org/wp-content/uploads/2025/01/2025-01-15-CDT-AI-Gov-Lab-Auditing-AI-report.pdf.

[7] Amy Winecoff and Miranda Bogen, *Improving Governance Outcomes Through AI Documentation: Bridging Theory and Practice*, *Center for Democracy & Technology* (2024), https://cdt.org/wp-content/uploads/2024/09/CDT-AI-Documentation-Report-092424-final.pdf.

[8] *Id.*

Moreover, where relevant, developers should be urged to produce documentation artifacts in consultation with the stakeholders that are their intended audience. When documentation artifacts are not developed with their intended use in mind, they tend to neglect stakeholder needs and thus fail to fulfil their goals and end up minimally useful to their intended recipients.[9] Developing these artifacts in consultation with relevant stakeholders is an important means of avoiding this failure mode. In a similar vein, developers could also be urged to create avenues through which external stakeholders can provide feedback on the form and usefulness of documentation artifacts.

Lastly, AISI should urge actors other than model developers to play an active role in promoting the responsible documentation of developers' misuse risk management. For instance, model-hosting platforms, such as Hugging Face, could be encouraged to establish norms and best practices regarding documentation of misuse-relevant risk management, analogous to how they have established norms regarding system cards.[10]

III. MAINTAINING PRIVACY DURING POST-DEPLOYMENT MONITORING

A final dimension of this draft guidance that ought to be strengthened is its recommendations regarding post-deployment monitoring. Detailed recommendations for post-deployment monitoring support a crucial element of effective risk management, and we agree that "distribution channels" — third-party platforms that make foundation models available to users, a major path through which many users interact with foundation models — may have an important role in monitoring for misuse and sharing relevant information with developers.

However, AISI should clarify that post-deployment monitoring must be carefully balanced against considerations of user privacy. While monitoring for misuse is important, its importance does not justify invasive methods that would require developers to indiscriminately access user interactions with their models — especially because users may be prone to input highly sensitive information during those interactions. Thankfully, developers need not rely on such invasive methods. Instead, they can use, and continue to develop, privacy-preserving techniques for monitoring for the potential malicious use of a deployed model; indeed, at least one major foundation model developer has begun to develop such techniques for post-deployment monitoring at scale.[11] These techniques expose only anonymized, aggregated user prompts to the developer. AISI should encourage all developers to adopt privacy-preserving, robust techniques for post-deployment monitoring that allow them to detect dangerous misuse while not infringing on the rights of their users.

***

We appreciate AISI's continued solicitation of feedback from stakeholders and affected communities on these important matters. For additional information, or any inquiries, please contact Miranda Bogen (mbogen@cdt.org), Director of CDT's AI Governance Lab.

---

[9] *Id.*
[10] *See* Ezi Ozoani, Marissa Gerchick, and Margaret Mitchell, "Model Card Guidebook," *Hugging Face* (2022), https://huggingface.co/docs/hub/en/model-card-guidebook.
[11] Alex Tamkin et al., "Clio: Privacy-Preserving Insights Into Real-World AI Use," *arXiv* (2024), https://arxiv.org/abs/2412.13678.