SIIA

September 9, 2024

Ms. Elizabeth Kelly
Executive Director
U.S. Artificial Intelligence Safety Institute
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

Re: Request for Comment on NIST 800-1

Dear Ms. Kelly:

On behalf of the Software & Information Industry Association (SIIA) we write to provide feedback on the U.S. Artificial Intelligence Safety Institute's (AISI) initial public draft of NIST 800-1, *Managing Misuse Risk for Dual-Use Foundation Models* ("800-1 Draft").

SIIA is the principal trade association for companies in the business of information. Our membership of nearly 400 companies reflects the broad and diverse landscape of digital content providers and users in academic publishing, education technology, financial information, software, platforms, data analytics, and information services. Our members include upstream and downstream AI designers, developers, and deployers of AI systems across various environments.

We appreciate AISI's attention to the challenges associated with managing and measuring misuse risk and its initiative in proposing a set of voluntary best practices to mitigate misuse risk. As detailed in this submission, we believe this is a productive first step in developing guidance to mitigate misuse risk. We provide several recommendations to improve the utility and uptake of AISI's recommended practices and further the overall goal of improving the safety, security, and trustworthiness of dual-use foundation models.

1. AISI Should Calibrate Recommended Practices Based on Recognized Challenges in Mapping and Measuring Misuse Risk and Existence of Technical Guidance

We appreciate the AISI beginning 800-1 Draft with a clear recognition of the limitations of measurement science in mapping and measuring misuse risks. (800-1 Draft, at 2-3 (Section 3).) Among the challenges are nascent methods to evaluate safeguards, limitations in measuring the potential of harm, and the broad applicability of foundation models. This is important because technical guidance and formal standards for achieving the objectives set out in 800-1 Draft do not yet exist. While there are robust efforts underway to understand the scope of misuse risk and develop mitigation methods, it is fair to describe these as in their infancy.

We recommend that AISI incorporate a more robust recognition of these limitations throughout the suggested practices. This is a threshold issue that feeds into the ability of an organization to develop its own "risk tolerance" level. The utility of the AISI's recommended practices will be strengthened if those practices are calibrated to what is technically feasible.

Relatedly, we recommend that the AISI provide technical guidance on each of the recommended practices. Where no technical guidance is available, we recommend that 800-1 describe the practice as "aspirational" and provide an indication about the future development of relevant technical guidance.

2. AISI Should Calibrate Recommended Practices Based on Different Characteristics of Dual-Use Foundation Models and Foreseeability of Risk

The framework in 800-1 Draft presumes a degree of uniformity among dual-use foundation models and the ability to identify (and thereafter manage) potential misuse risks that we believe does not accurately account for the variations in these models.

The degree of openness of a model is one characteristic that may require a different approach to several of the recommended practices in 800-1 Draft. Understanding the degree and type of openness in a model is fundamental to the ability to foresee misuse risk associated with that model. The National Telecommunications and Information Administration (NTIA) addressed this issue in its recent report, *Dual Use Foundation Models with Widely Available Model Weights*. The NTIA report recommends a marginal risk framework, stating: "The consideration of marginal risk is useful to avoid targeting dual-use foundation models with widely available weights with restrictions that are unduly stricter than alternative systems that pose a similar balance of benefits and risks."

The marginal risk framework is one that aligns with input SIIA provided during the comment period. We addressed the fundamental need to conduct a risk assessment of these models across the level-of-access gradient, from fully closed models to those that provide one or more of the following: hosted access, API access, API access for fine tuning, access to weights, access to training data, access to code with use restrictions, access to features without restrictions, and so on. We urged NTIA to defer to NIST on developing an AI RMF use-case profile for generative AI to guide their assessment of openness, noting, among other things, that certain "risks can be mitigated through various measures, including staged release; less than fully open access; limitations on who can access the weights (e.g., through license and user restrictions); and limitations on how the assets can be used (e.g., use restrictions and contract terms)." We cautioned "against a one-size-fits-all approach to mitigating risks for open models due to the gradient of openness, the differences among models, and differences around model training data. In addition, advances in foundation models, risk mitigation techniques (TEVV, auditing, red-teaming, and so on) and the capabilities of bad actors mean that any approach must be sufficiently flexible and agile to adapt."

Our submission to NTIA considered just one set of characteristics of foundation models – openness. There are many other characteristics of these models that will bear on the risks associated with those systems, and factor into the actions of various actors involved in the design, development, and deployment of foundation models. These include, for example, system architecture (generative adversarial networks, variational autoencoders, autoregressive models, diffusion models, transformer-based models, and so forth), training mechanisms, applications, training libraries, computational requirements, and other features.

We believe 800-1 would benefit from further attention to different characteristics of dual-use foundation models and their impact on key safety and security risks that may be associated generally with dual-use models, particularly in the ability to foresee misuse risk. This would help in developing best practices that are tailored to different types of models and help the NIST AISI to advance a robust approach to model safety in its engagement with AISIs in partner nations.

3. AISI Should Delineate Guidance Across the Entire AI Value Chain

Although 800-1 Draft is addressed specifically to model developers, effectively managing misuse risk requires engagement across the full AI value chain. Actors across the AI value chain have critical roles in mitigating risk especially in the context of dual-use foundation models which, as NIST recognizes, are broadly applicable and may be used in ways that the developers did not intend. The approach of 800-1 Draft contrasts with NIST AI 600-1, which acknowledges the importance of a comprehensive perspective that spans the entire AI lifecycle.

This is particularly relevant in the context of open-source AI models, where the responsibility for managing risks extends beyond the initial developers which may include those who deploy and use these models in various contexts. For instance, 800-1 Draft's Objective 4, which aims to measure the risk of misuse, is heavily dependent on how models are used and by whom. Likewise, Objective 6, concerning the gathering of information about misuse after deployment, underscores the need for a broader focus that includes all stakeholders in the AI value chain. Given this, we recommend NIST reframe 800-1 to avoid unrealistic expectations of developers and promote an approach that is calibrated to improve risk assessment and mitigation of dual-use models.

4. AISI Should Consider the Downstream Implications of All Recommended Practices Prior to Finalizing 800-1

Policymakers across the United States and around the world have viewed NIST as a thought leader in AI risk management since NIST commenced the AI Risk Management Framework process in 2021. One of the consequences of this is that policymakers have begun to rely on and incorporate NIST guidance into law (for example, in Colorado's new AI law). For this reason, it is important that the recommended practices in 800-1 take into account the possibility that regulatory agencies or jurisdictions may require organizations to comply with the individual practices, even if there are significant challenges associated with carrying out the practices.

As an illustration, Objective 5 proposes recommendations that will be extremely difficult for any developer to realize, and even harder (if not impossible) for an open model developer. Objective 5 would require developers to avoid taking actions to increase access to a model unless misuse risks are adequately managed. While the proposed benchmark - an organization's own risk tolerance - is subjective, there will be an expectation that managing risk means reducing the possibility of harm to zero. The effect will be more pronounced with respect to open source models. By their very nature, developers of these models retain less control of how the models may be used. Yet there are many benefits of open models, as NTIA has recently explained in depth, for public safety, competition, research, government use, and more.

Similarly, aspects of Objective 7 related to transparency could hinder the release and deployment of foundation models, as companies may be hesitant to expose themselves to legal liability. This could stifle the development and deployment of innovative AI technologies, ultimately impacting the broader industry. For example, 7.1 is overly broad when noting that information should only be disclosed "without introducing risks to public safety" and details should be shared "without rendering the safeguards ineffective." By publicly disclosing key descriptors regarding misuse risks and how they are managed, based on the way NIST recommends documenting and evaluating these risks, this could open the door for threat actors and serve as a roadmap for malicious behavior.

Summary of Recommendations

Given the concerns outlined above, we recommend the following:

1. **Engage with the NIST AISI Consortium Before Releasing 800-1.** We believe the 800-1 Draft requires further input from the community of stakeholders before it is ready to be finalized. We would recommend that NIST engage closely with Consortium Task Force 5.1 to build out the approach and detail in this publication.
2. **Incorporate Language Reflecting the Limitations of Measurement Science and Technical Standards.** To the extent 800-1 will serve as a collection of aspirational best practices, it would be more effective, and limit the downstream effects, if it identifies scientific and technical limitations to the proposed best practices. At the same time, NIST should prioritize the development of measurement science for misuse risk, providing clear and actionable guidance that can be applied consistently across different AI systems.
3. **Consolidate Technical Guidance on Identifying and Managing Misuse Risk.** Given the robust engagement of the academic and industry communities on identifying and assessing misuse risk, we recommend that NIST take steps to collect key research and use those to help establish a shared understanding of the challenges, promote consistent approaches to managing these risks, and identify current limitations. This research would serve as a valuable resource for AI developers, policymakers, and other stakeholders by providing a basis for a consistent and informed approach to AI safety.
4. **Promote Harmonization Across Jurisdictions:** We encourage NIST to continue engaging in active dialogue with other U.S. government agencies and international bodies to harmonize the approaches to managing AI risks. By doing so, NIST can help ensure that AI safety standards are coherent and interoperable, reducing the burden on companies operating in multiple jurisdictions and already complying with existing marginal risk standards.

Thank you for considering our comments. We look forward to continued collaboration with the AISI in shaping a balanced and effective approach to AI safety and promoting responsible AI development and use.

Respectfully submitted,

Paul Lekas
Senior Vice President, Head of Global Public Policy and Government Affairs
Software & Information Industry Association

Bethany Abbate
Manager, AI Policy
Software & Information Industry Association