9 September 2024

Artificial Intelligence Safety Institute
National Institute of Standards and Technology
Department of Commerce
Washington, DC 22314

In re: IEEEs response to Request for Comments on the Initial Public Draft of the Managing Misuse Risk for Dual-Use Foundation Models guidelines (Docket Number 240802-0209)

IEEE is pleased to submit the following response to the Initial Public Draft and AISI's questions laid out in the RFC. This response includes input from both IEEE-USA and the IEEE Standards Association.

In November 2023, IEEE-USA published a *Position Statement on Effective Governance of AI,* with the explicit recommendation of developing "a diverse toolkit of strategies to manage" AI. The Artificial Intelligence Safety Institute's (AISI) publication of this initial public draft represents an important step towards fulfilling this governance gap and, moreover, setting the stage for "international cooperation for ethical, trustworthy AI systems."

A characteristic of general-purpose technologies, such as AI, is their potential for dual-use that can be prone to misuse. To effectively mitigate these risks, it is necessary to build trust by understanding how to test and evaluate these technologies. As outlined in IEEE-USA's Position Statement, it is important to "calibrate public trust, understanding, and discourse about AI systems" and "develop mechanisms for identifying and accounting for the features of AI systems that could cause current testing, evaluation, certification, and investigation methods to misinform decision makers or the public about the risk of system deployment or the causes of system malfunction." In IEEE-USA's opinion, this Initial Public Draft contributes to those objectives.

Please do not hesitate to contact Erica Wissolik at e.wissolik@ieee.org or (202) 360-5023 if you have questions and wish to discuss further.

Sincerely,

Keith Moore
IEEE-USA President

IEEE-USA | 2001 L Street, N.W., Suite 700, Washington, D.C. 20036-4928 USA
Office: +1 202 785 0017 | Fax: +1 202 785 0835 | E-mail: ieeeusa@ieee.org | Web: https://ieeeusa.org

IEEE-USA's Response:

1. What practical challenges exist to meeting the objectives outlined in the guidance?

Section 3 of the document outlines several challenges to meeting the guidance objectives and advises organizations to "take appropriate steps to address" them. Yet, missing from the draft are recommendations as to what those appropriate steps are, how they impact the recommended objectives and practices, or recommendations on future research & development (R&D) activities needed to address these challenges.

Some of the recommendations refer to practices or taxonomies outside of the guidance that may not be broadly understood, are emergent or still maturing, and lack references that can be used by adopters to implement the practice. Examples include:

- Taxonomies of "capabilities of concern" and "threat actors" (Practice 1.1, 4.1);
- Estimating performance based on characteristics and which characteristics to consider (Practice 1.3, 4.1);
- References to recommended "security practices" and "safeguards", for example, *NIST SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile* (Practice 2.2, 3.3);
- Monitoring for misuse while maintaining privacy (Practice 6.1);
- Monitoring mechanisms for widely available weights (Practice 6.1); and
- Machine-readable format for reports of misuse (Practice 7.3).

Addressing these challenges, or deepening our collective understanding of them, is an opportunity for the AISI, its Consortium, NIST more broadly, and the AI community at-large. Including a section at the end of the guidelines that summarizes future R&D or guidance needs could encourage prioritization and collaboration of these specific needs. In other words, when the AISI (or NIST) is producing guidance for an emerging area, and they identify gaps, this is an opportunity to clearly communicate those gaps to those that may be able to fill them. This recommended action is inspired by academic research that typically identifies areas for further research, which in turn inspires other researchers.

Using pre-existing artifacts (such as templates and evaluation suites) and tools (such as AI red team automation frameworks) can help create a common baseline, reduce implementation cost, and free up limited resources for other activities. For example, the use of existing automation frameworks and testing scripts for AI Red Teaming (Practice 4.2) can free up limited AI Red Teaming resources to find new and novel risks and gaps in mitigations. While AISI may not want to recommend specific artifacts and tools, highlighting practices where these exist could increase the practicality of the guidance and reduce implementation cost, especially for smaller and less resourced actors. The AISI could also consider leveraging the AISI Consortium to collect, maintain, and publish a catalog of these artifacts, acknowledging that the catalog is provided without endorsement.

2. How can the guidance better address the ways in which misuse risks differ based on deployment (e.g., how a foundation model is released) and modality (text, image, audio, multimodal, and others)?

One of the easiest ways the guidance can be applied in differing scenarios is by mapping the practices to the characteristics (such as deployment method and modality), with the intersection of these dimensions indicating whether the practice:

- Applies;
- Must be implemented upstream (e.g. this practice can only be implemented by the developer because it requires changes to the training process);
- Must be implemented downstream (e.g. this practice can only be implemented downstream because it requires knowledge of the specific usage context);
- Requires downstream action (e.g. this practice is implemented by the developer, but some action is required downstream for the risk to be fully mitigated);

- Does not apply; or,
- R&D is needed (e.g. this practice does not have an implementation that works with a specific modality, once it does it would apply).

To avoid adding complexity to the objectives and practices, which may become overwhelming, these mappings could be provided as an appendix, or as a standalone document. These mappings should evolve over time as new characteristics emerge or the implementation of a given practice changes.

3. How can the guidance better reflect the important role for real-world monitoring in making risk assessments?

The guidance could expand on the different signal sources that real-world monitoring could consume and how the insights from the monitoring should be fed back into the risk management process. For example, developers should monitor the AI ecosystem for information about emerging threats and mitigation techniques and use those insights to adapt their risk assessments and mitigation implementations. Ensuring that there is a feedback loop from real-world monitoring to the risk assessment and mitigation processes is essential to stay ahead of new and novel threats. Creating these feedback loops is consistent with the following NIST AI RMF elements:

- MANAGE 3: "AI risks and benefits from third-party entities are managed;"
- MANAGE 4.2: "Measurable activities for continual improvements are integrated into AI system updates…";
- GOVERN 5.1: "Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team…"; and
- GOVERN 5.2: "Mechanisms are established to enable the team that developed or deployed AI systems to regularly incorporate adjudicated feedback…".

4. How can the guidance's examples of documentation better support communication of practically useful information while adequately addressing confidentiality concerns, such as protecting proprietary information?

The appropriate level of transparency is a careful balance of the value it provides, concerns such as confidentiality, production and maintenance cost, and the potential for documentation (for example, results from AI red teaming and penetration tests) to be misused by malicious actors. There is no single answer on how to achieve this balance, but some techniques that can be applied include:

- Critically examining how the documentation is (or could be) used by downstream actors in the AI supply chain;
- Determining the specific information needed to support those use cases;
- Identifying when summarized, aggregated, or otherwise redacted information would be sufficient; and,
- Modeling how this information may assist attacks by malicious actors.

This level of detail would be impractical to produce for this guidance; however, the technical standards development process is ideally suited to producing this information. The consensus process used to develop technical standards brings together different perspectives and typically produces outcomes that balances the value and concerns. Often during this process, alternative solutions are discovered that achieve a similar outcome with fewer concerns.

Section 3 of the guidance should explicitly include the challenge of achieving the appropriate level of transparency as a key challenge, and the AISI should identify the transparency documentation that should be prioritized for future technical standards development activities.

5. How can the guidance better enable collaboration among actors across the AI supply chain, such as addressing the role of both developers and their third-party partners in managing misuse risk?

One of the inherent challenges of any dual-use technology is the knowledge required to determine whether certain activities can be categorized as "misuse". The developer of a foundation model, which are by definition, general purpose ("applicable across a wide range of contexts"), may have little to no visibility into the context where a system is employed. The further up the AI supply chain a developer is, the more limited their visibility will be, and in some scenarios, such as widely available weights, they will have almost no visibility and extremely limited enforcement ability. While some activities will be clearly misuse, the more "dual use" an activity is, the more difficult it will be to determine its risk of harm, and, in these cases, actors downstream will have an increasingly important role to detect and prevent misuse. Highlighting the subset of practices and transparency documentation that can be performed or produced by developers only may help developers prioritize their efforts to unblock downstream actors. For example, practices that need to be performed prior to development or deployment (Practice 1.3, 3.2, …) and practices that require control over the model weights (Practice 3.3, …) can *only* be implemented by developers.

While possibly out of scope for this guidance, describing how practices and transparency documentation should be used by downstream actors, especially when action is needed from them, and how they should assess and mitigate any residual or contextual risks would provide a more complete risk mitigation process. For example, Objective 6 is more effective when downstream actors adopt practices to report incidents of misuse to upstream actors. This guidance could also be expanded to cover other actors in the AI supply chain, such as platform services including AI safety services, model registries, and inference services.

Aligning the structure of the guidance closer to the NIST AI Risk Management Framework (AI RMF) may make it easier for actors that are using the AI RMF to incorporate the recommended practices into their risk management process. Grouping the objectives by AI RMF function and/or tagging each practice with the AI RMF function, category, or subcategory it relates to will create a strong alignment with the AI RMF and help adopters understand how the recommended practices relate to other practices in their AI RMF implementation. While it's preferred that this is incorporated into the body of the guidance, if this is not possible then a mapping table or diagram showing the relationship with the AI RMF would be beneficial.

****

The IEEE Standards Association (IEEE SA) welcomes the opportunity to provide its comments to NIST on the U.S. Artificial Intelligence Safety Institute (AISI) draft document: Managing Misuse Risk for Dual-Use Foundation Models with Widely Available Model Weights**,** which proposes guidelines for improving the safety, security, and trustworthiness of dual-use foundation models consistent with the National AI Initiative Act and the White House Executive Order 14110.

IEEE SA is a globally recognized standards-setting body within IEEE, the largest organization of technology professionals in the world. We develop consensus standards through an open process that engages industry and brings together a broad stakeholder community.

The draft document is intended to identify best practices to map, measure, manage, and govern misuse risks from dual use foundational Artificial Intelligence (AI) models, and to provide transparency into the management of identified risks.

IEEE SA recognizes that this is an important step in identifying the risks associated with the misuse of dual use AI foundational models and in framing approaches to address the wide-ranging implications for risk management and risk mitigation, however the framing of the report raises several concerns, particularly from both a non-proliferation and risk methodology perspective.

The report states that it focuses on risks when "such models will be deliberately misused to cause harm." While it correctly notes that the methods by which models can be misused will continue to evolve, it seems to, intentionally or not, overlook the fact that significant harm is equally likely to come from non-deliberate misuse,

not least because of how little is known, understood, and obfuscated in the deployment of these technologies and their embedding into critical safety systems.

Given the nature of these systems, non-deliberate misuse could easily pave the way for deliberate misuse to cause harm. This crucial link is implied but not explicitly addressed in the report, which we suggest is a gap that needs to be filled. Addressing both deliberate and non-deliberate misuse is essential, especially given that the report's examples of "known examples of misuse" refer to chemical, biological, radiological, and nuclear (CBRN) weapons.

While IEEE SA recognizes the growing evidence that generative models might devise ways to weaponize in unexpected ways, it is important to note that there is currently limited evidence that these models are being misused in the context of Chemical, Biological, Radiological and Nuclear (CBRN) weapons. That does not mean that this is not a significant risk, but the report would benefit from some nuance and clarification here.

However, the potential for harm remains significant, and the range of risks associated with these models is far more extensive than just CBRN risks. Focusing exclusively on CBRN threats without addressing the broader spectrum of potential misuses may narrow the scope of the report's risk assessment.

If the report is to focus on CBRN risks, we suggest that it should emphasize that the Biological Weapons Convention[1], while primarily addressing deliberate use, also underscores the importance of robust biosafety and biosecurity measures to prevent accidental or non-deliberate misuse of biological agents.

Strengthening these frameworks to ensure generative models cannot be used for such purposes would be more effective than managing these risks solely through a generic framework for generative models.

While safeguards might be nascent in their application to generative models, there is extensive practice and experience with safeguards for sensitive technologies embedded and used in critical settings. Many of those practices also apply to generative models. It might strengthen the core arguments to refer to existing safeguards regimes that also uphold the balance between promotion (innovation) and protection (safeguards).

Similarly, we suggest that concerns about these models being used to automate cyber operations should be addressed with reference to existing international legal frameworks, such as the Budapest Convention on Cybercrime[2]. International law, including the Laws of War, clearly applies to digital technologies and the cyber realm.

Therefore, the paper would benefit from referencing established frameworks, such as **IEEE 7000™, Standard Model Process for Addressing Ethical Concerns during System Design**, highlighting not only the risks but also the fact that mechanisms are already in place that need to be strengthened. This would help ensure that misuse—whether deliberate or non-deliberate—of these systems is not merely seen as an issue of non-compliance within the AI lifecycle or a problem in the actors' supply chain, but as a violation of existing laws concerning their applications and generated outputs.

This need for clarity becomes even more urgent given the fast-expanding cyber threat surface, as more and more generative models and LLMs are being embedded and maintained in ways that can cause great and proven concerns.

While the harms resulting from deliberate and non-deliberate misuse may be similar, the political and regulatory solutions required to mitigate these risks are different. By acknowledging these, the document can help provide a more comprehensive and realistic approach to managing the risks associated with generative models - and in so doing also think deeply about what standards apply and identify gaps. It would underscore that these risks are not emerging in a regulatory vacuum but rather in an environment where legal structures already exist, albeit requiring reinforcement, a strong family of standards, and adaptation to new technological realities.

It is well known that the lack of transparency, the lack of independent reviews of data filtering and model training, the lack of independent monitoring, verification, and compliance of these systems, and issues related to red team results not being released or shared all contribute to the challenges in managing these risks. The draft report notes that "Consider providing red teams with available legal protections for their tasks, such as waiving terms of service and indemnifying them for legal liability for their interactions with the model." This is a hard ask given that companies will still be liable for downstream processors of data by international data protection regulation. We suggest for consideration re-writing this recommendation to suggest that red teams are offered commercial reasonable terms. The recent review of compliance rate with White House issued guidance[3] demonstrated an extremely poor compliance, especially when doing systems critical tasks such as red teams and other transparency efforts.

Furthermore, methodologies for assessing harms and risks are often conducted internally and not shared, leading to a lack of transparency. It remains unclear whether issues around CBRN and the cyber threat surface are sufficiently understood, discussed, and prioritized within developer companies and their supply chains.

Integrating these considerations into the report will help provide a clearer, more structured analysis of the actual risks and the relevant legal and political frameworks that need to be strengthened to effectively address both deliberate and non-deliberate misuse of generative models.

IEEE SA would like to note that it is important for policymakers to be aware of the evolving nature of AI, the complexities in standardizing such technologies and the balance between enabling innovation and ensuring ethical considerations are followed. While regulation is essential, rigid regulation can hinder innovation. Well-designed international standards could minimize disparities in how foundational AI systems are deployed and associated risks are managed as well as provide the foundation for compliance.

At IEEE SA, our community has developed resources and standards globally recognized in applied AI ethics and systems engineering and continue to develop accessible and sustainable approaches and solutions for pragmatic application of AIS principles and frameworks. For example, the first edition of the IEEE SA, "Ethically Aligned Design[4]" was published in 2018, and included a glossary[5] that defines AI ethics.

IEEE SA would like to point out that its portfolio of AI standards[6] and projects offer open and transparent mechanisms that address the management of AI systems, security, threats, and risks.

The value of adhering to a standards-based approach in *developing and deploying product development and compliance to standards can help mitigate potential risks as well as to enable interoperability between products and within systems.*

IEEE has a portfolio of standards that may be of interest to the NIST and the U.S. Artificial Intelligence Safety Institute as it looks to revise its draft.

These include:

IEEE 7001-2021™ Standard for Transparency of Autonomous Systems establishes measurable, testable levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined.

IEEE 7002-2022™ Standard for Data Privacy Process contains requirements for a systems/software engineering process for privacy-oriented considerations regarding products, services, and systems utilizing employee, customer, or other external user's personal data.

IEEE P7003™ Algorithmic Bias Considerations describes specific methodologies to help users certify how they worked to address and eliminate issues of negative bias in the creation of their algorithms, where "negative bias" infers the usage of overly subjective or uniformed data sets or information known to be inconsistent with

legislation concerning certain protected characteristics (such as race, gender, sexuality, etc); or with instances of bias against groups not necessarily protected explicitly by legislation, but otherwise diminishing stakeholder or user well-being and for which there are good reasons to be considered inappropriate.

IEEE P7009™ Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems establishes a practical, technical baseline of specific methodologies and tools for the development, implementation, and use of effective fail-safe mechanisms in autonomous and semi-autonomous systems. The standard includes (but is not limited to): clear procedures for measuring, testing, and certifying a system's ability to fail safely on a scale from weak to strong, and instructions for improvement in the case of unsatisfactory performance.

IEEE P7009™ Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems establishes a practical, technical baseline of specific methodologies and tools for the development, implementation, and use of effective fail-safe mechanisms in autonomous and semi-autonomous systems. The standard includes (but is not limited to): clear procedures for measuring, testing, and certifying a system's ability to fail safely on a scale from weak to strong, and instructions for improvement in the case of unsatisfactory performance.

IEEE 7010-2020™ Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being provides specific and contextual well-being metrics that facilitate the use of a Well-Being Impact Assessment (WIA) process in order to proactively increase and help safeguard human well-being throughout the lifecycle of autonomous and intelligent systems (A/IS).

IEEE P7010.1™ Recommended Practice for Environmental Social Governance (ESG) and Social Development Goal (SDG) Action Implementation and Advancing Corporate Social Responsibility provides recommendations for next steps in the application of IEEE Std 7010, applied to meeting Environmental Social Governance (ESG) and Social Development Goal (SDG) initiatives and targets. It provides action steps and map elements to review and address when applying IEEE 7010. This recommended practice serves to enhance the quality of the published standard by validating the design outcomes with expanded use. It provides recommendations for multiple users to align processes, collect data, develop policies and practices and measure activities against the impact on corporate goals and resulting stakeholders.

IEEE P7011™ Standard for the Process of Identifying and Rating the Trustworthiness of News Sources provides semi-autonomous processes using standards to create and maintain news purveyor ratings for purposes of public awareness. It standardizes processes to identify and rate the factual accuracy of news stories in order to produce a rating of online news purveyors and the online portion of multimedia news purveyors.

IEEE P7012™ Standard for Machine Readable Personal Privacy Terms identifies/addresses the manner in which personal privacy terms are proffered and how they can be read and agreed to by machines.

IEEE P7014™ Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems defines a model for ethical considerations and practices in the design, creation and use of empathic technology, incorporating systems that have the capacity to identify, quantify, respond to, or simulate affective states, such as emotions and cognitive states. This includes coverage of 'affective computing', 'emotion Artificial Intelligence' and related fields.

IEEE P7015™ Standard for Data and Artificial Intelligence (AI) Literacy, Skills, and Readiness establishes an operational framework and associated capabilities for designing policy interventions, tracking their progress, and empirically evaluating their outcomes. The standard includes a common set of definitions, language, and understanding of data and AI literacy, skills, and readiness.

IEEE P2840™ Standard for Responsible AI Licensing describes specifications for the factors that shall be considered in the development of a Responsible Artificial Intelligence (AI) license. Possible elements in the specification include (but are not limited to): (1) What a 'Responsible AI License' means and what its aims are (2) Standardized definitions for referring to components, features and other such elements of AI software, source

code and services (3) Standardized reference to geography specific AI/Technology specific legislation and laws (such as the EU General Data Protection Regulation - GDPR) as well as identification of violation detection, penalties, and legal remedies. (4) The specification lists domain specific considerations that may be applied in developing a responsible AI license.

IEEE P2863™ Recommended Practice for Organizational Governance of Artificial Intelligence specifies governance criteria such as safety, transparency, accountability, responsibility and minimizing bias, and process steps for effective implementation, performance auditing, training and compliance in the development or use of artificial intelligence within organizations.

IEEE P3119™ Standard for the Procurement of Artificial Intelligence and Automated Decision Systems establishes a uniform set of definitions and a process model for the procurement of Artificial Intelligence (AI) and Automated Decision Systems (ADS) by which government entities can address socio-technical and responsible innovation considerations to serve the public interest.

IEEE 2089-2021™ Standard for an Age Appropriate Digital Services Framework (Based on the 5Rights Principles for Children) sets out processes through the life cycle of development, delivery and distribution, that will help organizations ask the right relevant questions of their services, identify risks and opportunities by which to make their services age appropriate and take steps to mitigate risk and embed beneficial systems that support increased age appropriate engagement.

IEEE P2890™ Recommended Practice for Provenance of Indigenous Peoples' Data outlines the core parameters for providing and digitally embedding provenance information for Indigenous Peoples' data. The recommended practice establishes common descriptors and controlled vocabulary for provenance, including recommendations for metadata fields that can be used across industry sectors, including machine learning (ML) and artificial intelligence (AI) contexts, biodiversity and genomic science innovation and other associated databases, and supports proper and appropriate disclosure of originating data information.
For the complete list of IEEE AI/S Standards and current Projects please see:
https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/

We would look forward to working with the U.S. Artificial Intelligence Safety Institute on its next iteration of this draft.

---

[1] https://www.state.gov/biological-weapons-convention/
[2]https://www.justice.gov/opa/pr/united-states-signs-protocol-strengthen-international-law-enforcement-cooperation-combat
[3] https://www.whitehouse.gov/omb/briefing-room/2023/11/09/biden-harris-administration-releases-final-guidance-to-improve-regulatory-analysis/
[4] https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e.pdf
[5] https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e_glossary.pdf
[6] https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/