IST Institute for SECURITY + TECHNOLOGY

IST Leadership

Mike McNerney
Chair, Board of Directors

Philip Reiner
Chief Executive Officer

Megan Stifel
Chief Strategy Officer

Steve Kelly
Chief Trust Officer

Institute for Security and Technology
195 41st Street #11045
Oakland, CA 94611

March 14, 2025

National Institute of Standards and Technology
100 Bureau Dr
Gaithersburg, MD 20899

Subject: Comments on NIST AI 800-1, *Managing the Risk of Misuse for Dual-Use Foundation Models*, Second Public Draft; NIST-2025-0001.

Dear Ms. Chambers,

The Institute for Security and Technology (IST) appreciates the opportunity to file comments in response to a request for public feedback on the second draft of its guidelines on **Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1)**. We submit for consideration elements of the following IST reports: "*A Lifecycle Approach to AI Risk Reduction*: *Tackling the Risk of Malicious Use Amid Implications of Openness*" (June 2024),[1] "*The Implications of Artificial Intelligence in Cybersecurity"* (October 2024), *"Navigating AI Compliance: Tracing Failure Patterns in History"* (December 2024),[2] and *"Navigating AI Compliance: Risk Mitigation Strategies for Safeguarding Against Future Failures"* (pending publication in March 2025). The study process leading to the final two reports involved participation from a working group of 20 stakeholders from leading AI labs, industry, academia, and civil society.

Consistent with Figure 1 on page 7 (NIST AI 800-1) depicting the AI lifecycle, IST has also adopted an "AI Lifecycle Framework" for mapping specific policy or technical risk mitigation strategies to the relevant lifecycle stage for implementation. However, in consultation with our working group of AI/ML engineers and technical experts, we adopted a more granular framework consisting of the following seven (7) stages (See "*A Lifecycle*

[1] Louie Kangeter, "A Lifecycle Approach to AI Risk Reduction," Institute for Security and Technology, June 2024, https://securityandtechnology.org/wp-content/uploads/2024/06/A-Lifecycle-Approach-to-AI-Risk-Reduction.pdf
[2] Mariami Tkeshelashvili, Tiffany Saade, "Navigating AI Compliance, Part 1," Institute for Security and Technology, https://securityandtechnology.org/wp-content/uploads/2024/12/Navigating-AI-Compliance.pdf

Approach to AI Risk Reduction: *Tackling the Risk of Malicious Use Amid Implications of Openness*," p. 7):

- Data collection and preprocessing
- Model architecture
- Model training and evaluation
- Model deployment
- Model application
- User interaction
- Ongoing monitoring and maintenance.

We recommend revisiting your publication's choice of the Organization for Economic Cooperation and Development's (OECD) approach as we believe it does not fully characterize the real-world stages of AI development, deployment, and use. Also, its "not necessarily sequential" approach and overlapping stages might hamper its utility when attempting to develop and implement specific risk reduction strategies targeted to the most relevant lifecycle stage, as has become our approach.

In addition to the NIST assessment made in Chapter 4, pages 4 and 5 on the key challenges of managing misuse risks, IST concluded that the AI ecosystem will face novel challenges related to the development of AI agents. IST recommends expanding the list of challenges by recognizing that the AI agents will blur the lines of liability in the automated world. The proliferation of AI agents and the rise of multiagent environments can create feedback loops in which decisions based on past data may influence future outcomes, and any causal connection between the original deployer's intent and future outcomes will inevitably attenuate (See "*The Implications of Artificial Intelligence in Cybersecurity*," p. 33). This scenario could enhance and reinforce biases or inaccuracies, or worse yet, leave the human out of the loop altogether (See *"Navigating AI Compliance: Tracing Failure Patterns in History,"* pp. 19-20).

IST aligns with NIST's Practice 3.1, #3 (line 20) on insider threats, and recommends the publication also incorporate the concept of an AI model itself constituting an insider threat. While this is a newer concern, our initial inquiries into AI agents (discussed above) and discussions with experts on the topic of "AI control" brings to mind scenarios in which increasingly capable and autonomous AI agents operating within an authorized context might self-evolve and eventually deviate from their intended

scope—potentially even colluding with other agents to bypass controls. This is an emerging, and as yet underdeveloped, area of research. This topic may be best aligned with your Objective 6 on “Monitor and respond to misuse.”

IST commends NIST’s emphasis on monitoring and responding to misuse in Objective 6 (Practices 6.1-6.5) as a critical component of managing risks associated with dual-use foundation models. Given the rapidly evolving nature of AI development, we propose enhancements to these practices to ensure that they not only track AI misuse incidents, but also assess the effectiveness of detection and mitigation strategies over time. This adaptive approach, supported by insights from our reports, *"A Lifecycle Approach to AI Risk Reduction Tackling the Risk of Malicious Use Amid Implications of Openness"* (June 2024) and *"Navigating AI Compliance Part 1: Tracing Failure Patterns in History"* (December 2024), aims to streamline compliance with NIST’s objectives while optimizing resource allocation, resulting in more effective risk mitigation efforts.

Accordingly, IST recommends adding a new practice under Objective 6 on instituting a mechanism to evaluate the performance of misuse identification and mitigation efforts, and thereafter adapting strategies based on those learnings. It might read as follows:

Practice 6.6: Monitor and refine counter-misuse practices.

Continuously monitor the efficacy of misuse detection, response, and mitigation approaches; regularly update practices to close identified performance gaps.

Recommendations for Practice 6.6:

1. Monitor and evaluate misuses that are successfully detected, those that evade initial detection, why certain mitigations fail, and response procedures.

2. Regularly refine misuse detection, response, and mitigation practices to close identified gaps.

3. Document and share learnings and best practices with other entities across the AI supply chain to inform improvements across the ecosystem.

Considering NIST AI 800-1 contains voluntary guidelines, IST notes that adopting these practices can generate return on investment (ROI) for AI users and builders. While perhaps not appropriate for a technical publication, IST would like to share for the record our findings in *"Navigating AI Compliance: Risk Mitigation Strategies for Safeguarding Against Future Failures"* on the following forms of ROI when adhering to risk management best practices:

- Given the rapid proliferation of AI tools, industries utilizing these technologies are expected to face increasing scrutiny from regulators. Proactively implementing safety, security, privacy, transparency, and anti-bias measures can help prevent unexpected and costly harms, their associated litigation, and reputational implications.
- Strong compliance practices provide a competitive advantage for both the AI system builders and enterprises adopting it. A recent report from Bain reveals that organizations with an effective approach to responsible AI doubled their profit impact from their AI efforts compared to those organizations that lack such an approach.
- The United States government's procurement policies and preferences make and shape markets. A company that complies with the relevant standards in the AI space will be better prepared to compete in the government procurement-shaped markets. As the AI market becomes one of the largest and most valued in the geopolitical and economic race to the top, governments will likely increase their investments into the development of frontier models, and will likely favor companies whose products are safe to use, with robust security standards in place.
- Organizations that prioritize responsible AI development and deployment practices have an edge in attracting top talent who increasingly seek workplaces committed to responsible innovation. A strong ethical framework enhances employee morale and loyalty, fostering an environment where skilled professionals want to contribute and grow. This talent pipeline is crucial for both model capability development, as well as scaling AI products into new markets worldwide.

- By investing in responsible AI practices, companies can build stronger relationships with customers, partners, and employees, leading to higher satisfaction and loyalty. For customers, this translates to increased lifetime value to the company, as satisfied customers are more likely to return and advocate for the brand, ultimately boosting long-term profitability. Proactively addressing compliance concerns in AI can safeguard an organization's reputation over time. Companies that navigate these challenges effectively are better positioned to withstand scrutiny and maintain public trust, ensuring their brand remains resilient against potential controversies.

- Enterprises that demonstrate compliance, particularly in emerging technologies like AI, are likely to attract more investment, as stakeholders increasingly consider security risks. A rigorous compliance program alludes to a lower risk threshold profitable scenario for new investors to come in, and for existing investors to sustain their investments.

I and my team welcome an opportunity to discuss our work and these comments with you. Thank you for considering them as you further refine this important publication.

Regards,

Steve Kelly
Chief Trust Officer