HUGGING FACE

Hugging Face Comments on NIST AI 800-1: Managing Misuse Risk for Dual-Use Foundational Models

Hugging Face commends the US AI Safety Institute (AISI) on the AI 800-1 document: Managing Misuse Risk for Dual-Use Foundation Models. This comprehensive framework identifies key objectives and practices for managing risks associated with foundation models. We offer recommendations to strengthen this document based on our experiences in democratizing good AI and characterizing risks of systems as an open platform for state-of-the-art (SotA) AI systems. Our comments are organized by objectives and practices as outlined in the document. Where we do not have specific, actionable feedback on a section or practice, we have not highlighted it.

About Hugging Face

Hugging Face is a community-oriented company based in the U.S. and France working to democratize good Machine Learning (ML), and has become the most widely used platform for sharing and collaborating on ML systems. We are an open-source and open-science platform hosting machine learning models and datasets within an infrastructure that supports easily processing and analyzing them; conducting novel AI research; and providing educational resources, courses, and tooling to lower the barrier for all backgrounds to contribute to AI.

Executive Summary

Hugging Face commends the US AI Safety Institute (AISI) for its comprehensive AI 800-1 document to address managing misuse risks for dual-use foundation models. Based on our experience as a leading open AI platform, we offer the following key recommendations to enhance the framework's effectiveness:

1. **Joint Management Across the AI Supply Chain**: Risk management should be a shared responsibility among all stakeholders, including data providers, infrastructure providers, and end-users, rather than solely on individual model developers. Encourage open, collaborative approaches where diverse stakeholders contribute to defining risk thresholds and management strategies.
2. **Enhance Transparency, Accountability, and Ongoing Risk Identification**: Implement regular transparency reporting on risk management practices, and establish mechanisms for meaningful accountability, including sharing information with

independent entities. Improve scientific and regulatory visibility on risk profiles by providing clear guidelines for reporting and categorizing misuse risks, and support external safety research through vulnerability disclosure policies and safe harbor provisions.

3. **Tailor Risk Interventions and Balance Security with Accessibility**: Different risks require context-specific, flexible safeguards. Ensure that interventions are tailored to the nature of the risk, whether it be non-consensual intimate imagery or CBRN threats. When considering model access restrictions, balance the need for security with the benefits of openness to foster innovation and research.
4. **Recognize and Support Open Foundation Models**: Explicitly acknowledge the unique characteristics and benefits of open foundation models, including their role in enabling external scrutiny, mitigating monoculture, and fostering innovation. Develop guidelines that support responsible open-source AI development while managing associated risks.

Objectives and Practices to Manage Misuse Risks

Anticipating and Measuring Risk

(Objectives 1 and 4)

Holistic Approach to Risk Assessment

To effectively manage and safely deploy AI models, it is essential to adopt a holistic approach to risk assessment that encompasses both technical and societal factors. While this document primarily focuses on safety risks, it is important to recognize that mechanisms designed to address these risks can also be applied to broader societal concerns. Failing to integrate these considerations could lead to missed opportunities for creating more comprehensive and effective risk management strategies. For instance, the focus on preventing "model theft" might overlook the value of broad, inclusive participation in identifying model biases and other potential harms. A framework for measuring risks that comprehensively addresses foundational models' impacts on people and society would ensure that safety measures do not inadvertently neglect significant societal implications – or make addressing them through other means more difficult.

Collaborative Risk Definition

The current approach to risk definition, which relies on individual developers to define and assess risks, introduces several challenges. It creates disparities between corporate and collaborative developers, complicates comparison between models from different developers,

and deviates from established scientific processes. Instead, we advocate for a more standardized, collaborative approach. This approach could involve a centralized system where diverse stakeholders contribute to and maintain a comprehensive list of potential malicious uses or unintended consequences of foundation models, akin to the Common Use Enumeration (CUE) component we have proposed for structured harm reporting. This collaborative approach offers several benefits:

- **Broader Expertise:** It leverages a wider range of expertise, including domain-specific knowledge that individual companies may lack.
- **Resource Equity:** It reduces disparities in resources and incentives for risk evaluation between corporate and collaborative developers.
- **Consistent Assessment:** It enables more consistent risk assessment across different models and developers, facilitating meaningful comparisons and industry-wide progress on safety.
- **Scientific Alignment:** It aligns more closely with established scientific processes for risk assessment.
- **Comprehensive Risk Identification:** It reduces duplicated efforts across organizations and is likely to identify a more comprehensive and nuanced set of potential risks than any single organization could alone.

By moving towards a more inclusive, standardized approach, the industry can establish a more robust, scientifically rigorous, and comprehensive system for anticipating and managing potential misuse risks in foundation models. Such an approach corresponds to Practice 1.1 by improving the standardization and inclusivity of risk assessments.

Transparency in Capability Estimation

Transparency is crucial in estimating model capabilities before deployment (Practice 4.1). We recommend publicly documenting the methods used for capability estimation, including any limitations or uncertainties. Developing and using open benchmarks, leaderboards, and evaluation tools will enable more standardized and comparable capability assessments across the industry. Periodic measurement of capabilities throughout the development process is commendable, but this should be viewed as an opportunity to focus more on upstream development choices, such as dataset selection and model scaling, rather than merely increasing capabilities. For open models, leveraging their openness for collaborative risk identification is key.

More Effective Red Teaming

Red teaming is a critical component in identifying vulnerabilities and strengthening defenses against potential misuse (Practice 4.2). To further enhance the effectiveness of red team exercises, we suggest establishing guidelines for open participatory processes that allow the

public and strategically selected experts to red team a model safely, including safe harbor clauses. Additionally, industry-wide mechanisms for sharing anonymized findings should be implemented. This collaborative effort would improve collective understanding of emerging threats, foster the development of more effective mitigation strategies, and create a culture of shared responsibility within the AI community. Red teaming should be used as part of a broader suite of AI accountability tools, including algorithmic impact assessments, external audits, and public consultations.

Establishing Plans for Managing Misuse Risk

(Objective 2)

Standardized Risk Thresholds and Community Input

To enhance Practice 2.1, AISI should provide comprehensive guidelines for defining and assessing acceptable levels of misuse risk in various contexts. Drawing from collaborative governance approaches like the BigCode project, we recommend the following:

- **Multi-stakeholder advisory mechanisms**: Encourage the formation of advisory groups with representatives from academia, industry, and civil society to review and refine risk thresholds periodically.
- **Tiered risk categorization systems**: defining risk thresholds that accommodate varying acceptable risk levels across different contexts and stakeholder groups based on potential impact and likelihood of occurrence.
- **Open participation processes**: Outline best practices for mechanisms that enable diverse stakeholders to contribute insights on risk thresholds and management strategies.
- **Structured community input**: Recommend establishing regular feedback cycles where stakeholders can provide input on risk thresholds and mitigation strategies, modeled after collaborative processes like data inspection sprints.
- **Public documentation standards**: Encourage transparent documentation of risk threshold decisions and development processes to enhance accountability and trust.

Collaborative Risk Management Roadmaps

To improve Practice 2.2, AISI should provide guidelines for organizations to adopt iterative approaches to risk management, treating roadmaps as living documents that adapt to emerging threats. Recommendations include:

- **Stakeholder impact assessment tools**: Develop tools for stakeholders to assess their involvement or impact within AI systems, such as BigCode's "Am I in The Stack" tool that can help address privacy and software security risks.
- **Regular review cycles**: Establish best practices for periodic reviews of risk management roadmaps, engaging both internal teams and external advisors.
- **Lessons learned documentation**: Maintain detailed logs of risk-related insights from each iteration or release, documenting improvements and changes in mitigation strategies. For example, the Starcoder project continuously improved its data curation process, enhancing personally identifiable information (PII) redaction and opt-out mechanisms based on community input and evolving best practices.
- **Transparency reporting**: Issue regular reports outlining updates to risk assessment methodologies, changes in identified risks, and strategies to mitigate them.
- **Collaborative knowledge sharing**: Promote the creation of platforms or processes for sharing best practices in risk management, enabling collaboration between developers, researchers, and other stakeholders.

Managing Risks and Ensuring Responsible Model Release

(Objectives 3 and 5)

Reframing Model Theft and Managing Misuse Risks in Open Models

Efforts to protect valuable AI assets must balance the need for open science and collaboration. For open foundation models, the traditional concept of "model theft" requires re-evaluation. Open-weight models hold minimal or nonexistent risk of theft, depending on license and permissive use. Instead, the focus should be on responsible sharing and usage. We recommend that AISI recognize that this objective should not inadvertently deter the development and utilization of open models, which benefit from transparency and community collaboration. Efforts to prevent misuse should be proportional to the actual risks posed by open models compared to closed ones. Emphasizing the development of clear usage guidelines, ethical frameworks, and community standards for responsible AI development and deployment will be more effective. This approach aligns with Practice 3.1 by advocating for a shift from theft prevention to responsible use.

Context-Specific Safety Measures

A one-size-fits-all approach to risk management is inadequate for addressing the diverse range of threats associated with AI models. Different types of risks require tailored interventions. For instance, managing non-consensual intimate imagery involves distinct strategies compared to addressing Chemical, Biological, Radiological, and Nuclear (CBRN) risks. Implementing

safeguards proportionate to the model's misuse risk (Practice 5.2) necessitates flexible, context-specific measures. A tiered system of safeguards, adaptable based on the deployment context and model impact, allows organizations to tailor their risk management strategies effectively. Safeguards should be rigorously tested, with evidence of their effectiveness established before deployment. Clear and transparent criteria for what constitutes "adequate" risk management (Practice 5.3) are crucial, and these criteria should be regularly reviewed and updated to reflect new research and emerging threats.

Proactive and Contextual Risk Management

Pre-deployment risk management is crucial for the responsible development of foundation models. To strengthen this approach, we advocate for a holistic and inclusive strategy in assessing potential misuse risks (Practice 5.1). This assessment should involve cross-functional teams, including technical experts, legal professionals, ethicists, and communications specialists. By integrating diverse perspectives, these teams can comprehensively evaluate risks, considering technical vulnerabilities, ethical implications, public perception, and legal considerations. This comprehensive approach ensures that deployment risk assessments are robust and address all facets of potential AI incidents.

Managing misuse risks in open foundation models requires proactive strategies, especially since the release of model weights enables a wide range of downstream applications. Effective risk management extends beyond the model itself to include platform-specific considerations. A "safety by design" approach is essential, where risks are assessed before broadening access, and staged releases—such as gated models —allow controlled distribution and user verification. Model distribution safety techniques such as SafeTensors enable secure dissemination of open models. Comprehensive documentation, such as governance cards, should outline anticipated risks and mitigation strategies, empowering users to adapt models responsibly. Community engagement through discussion forums and transparent content moderation guidelines, further supports responsible deployment. By combining these strategies, platforms can manage misuse risks effectively while fostering safe and responsible AI development.

Standards for Ongoing Risk Identification and Transparency

(Objectives 6 and 7)

Distributed Responsibility for Misuse Identification and Reporting

Current guidance places substantial responsibility for identifying and responding to misuse on model developers (Practice 6.1). This approach can be particularly challenging for developers of both open models and closed models with commercial APIs serving a large and diverse user

base, who might not even have all the information required about downstream systems and use cases to effectively and independently implement risk management measures. We recommend adopting a distributed responsibility model, involving all relevant stakeholders—data providers, infrastructure providers, individual model developers, downstream application developers, and end-users—in monitoring and reporting misuse. Clear definitions of roles and responsibilities for each stakeholder should be established to facilitate effective communication and collaboration in identifying and addressing misuse.

Coordinated Response Mechanisms

To enhance Practice 6.2, AISI could establish clear incident response protocols that incorporate a collaborative responsibility model, tiered harm classification, and distributed responsibility across the AI supply chain. Drawing lessons from the mature Common Vulnerabilities and Exposures (CVE) ecosystem in cybersecurity practices, these protocols should include comprehensive steps for initial assessment and triage of reported misuses, procedures for escalation based on severity and potential impact, and guidelines for timely and transparent communication with affected parties. AISI could additionally develop standardized documentation formats for incident response, including model metadata that can be pulled from standardized model cards, incident timelines, root cause analysis, mitigation measures, and lessons learned. Templates for post-incident reports should balance transparency with the protection of sensitive information, in accordance with responsible disclosure principles. Additionally, an independent adjudicator should be designated to resolve disputes between reporters and vendors, ensuring impartial issue resolution. Similar to established practices in cybersecurity vulnerability disclosure, AI harm reporting guidelines should address safe harbor protections for reporters disclosing potential misuses in good faith, providing legal protection and promoting a culture of transparency and cooperation.

Transparency Reporting Standards

To improve *Practice 7.1* and *Practice 7.2*, we recommend the development of a standardized template for AI transparency reports. These reports should include:

- Sections on identified misuse risks, implemented safeguards, and a summary of misuse incidents and responses (excluding sensitive details).
- Guidance on ongoing monitoring efforts, with recommendations on the appropriate frequency and level of detail based on the model's capabilities and deployment context.
- Examples that illustrate how to present complex technical information in an accessible manner for diverse stakeholders.
- Guidelines for supply chain transparency, covering the documentation of training data origins, model architectures, key algorithms, and third-party components.

Currently, Practice 7.3 does not adequately address the lack of standardization and monitoring in AI incident reporting websites. For example, the AI Incident Database predominantly lists news reports, leaving out crucial findings identified through red teaming, bug bounties, or independent research. Standardization and regular maintenance of these reporting systems are crucial to ensure comprehensive coverage and effective incident management.

Conclusion

Hugging Face remains committed to the responsible development and deployment of AI technologies. We greatly appreciate the opportunity to provide insights on this document and we look forward to ongoing collaboration with NIST and AISI, other industry partners, researchers, and policymakers to refine and implement best practices for managing the risks associated with dual-use foundation models.

Submitted by:
Avijit Ghosh, Applied Policy Researcher, Hugging Face
Yacine Jernite, ML and Society Lead, Hugging Face
Irene Solaiman, Head of Global Policy, Hugging Face