AI Alliance

The AI Alliance Comment on NIST AI 800-1 Initial Public Draft:

"Managing Misuse of Dual-Use Foundation Models"

Introduction

The AI Alliance is a diverse community of organizations, large and small companies, academic and non-profit institutions, representing developers, researchers, and business leaders who are focused on accelerating and disseminating open innovation across the AI technology landscape. They aim to improve foundational capabilities, safety, security, and trust in AI, and to responsibly maximize benefits to people and society everywhere. The AI Alliance brings together a critical mass of compute, data, tools, and talent to accelerate open innovation in AI.

We are encouraged by the Federal Trade Commission[1] (FTC)'s recent statement, which concludes that "open-weights models have the potential to drive innovation, reduce costs, increase consumer choice, and generally benefit the public – as has been seen with open-source software." The U.S. Department of Commerce and the National Telecommunications and Information Administration (NTIA)[2] report on dual-use foundation models with widely available model weights states that "Dual-use foundation models with widely available model weights … introduce a wide spectrum of benefits. They diversify and expand the array of actors, including less resourced actors, that participate in AI research and development. They decentralize AI market control from a few large AI developers. And they enable users to leverage models without sharing data with third parties, increasing confidentiality and data protection."

The content in this response is provided by the AI Alliance and is not intended to reflect the views of any particular member organization. We value the opportunity to provide feedback on NIST AI 800-1.

Our response collectively urges NIST to address both open and closed foundation models, consider the entire AI ecosystem, provide practical and proportionate guidelines, align with established marginal risk standards, and promote harmonization across various risk management guidelines.

- ***Section A*** points out the importance of considering not only model developers but the full ecosystem who all have a role to play in mitigating risk. We recommend NIST AI 800-1 include the whole ecosystem or focus on risks that model developers are uniquely positioned to mitigate. This would avoid a situation in which model developers are required to micromanage the downstream uses of models by deployers – an approach that is at odds with fundamental principles of open source – and would allow each player in the ecosystem to focus on risks that it is best placed to mitigate.
- ***Section B*** notes that several identified risks in NIST-AI 800-1 do not have broadly established safety protocols and we recommend further research to develop those protocols, rather than anticipating that developers could apply protocols that do not yet exist.
- ***Section C*** encourages NIST to align its guidelines with the marginal risk standard adopted by the NTIA's recent report and independent experts, ensuring that the US Government's approach is

[1] Federal Trade Commission (FTC),On Open-Weights Foundation Model. 10 July 2024 https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models.
[2] NTIA Report. "Dual-Use Foundation Models with Widely Available Model Weights", July 2024: https://www.ntia.doc.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

consistent across work products and that the strategies for managing misuse risks in foundation models are practical and proportionate to the real risks they pose.

- ***Section D*** urges NIST to ensure that its final report provides guidance rooted not only in the risks but also in the benefits associated with foundation models, including the positive impacts these models can have on innovation, economic growth, and societal advancement
- ***Section E*** discusses the value of open foundation models, including features that enhance transparency, prevent market concentration, and help to address safety concerns and recommends NIST AI 800-1 consider the impact of its guidance on open models and focus on model-neutral recommendations that can be flexibly implemented.
- ***Section F*** recommends that NIST/AISI promote harmonization and integrate and expand upon existing frameworks such as NIST AI 600-1, the White House Executive Order, ML Commons AI safety benchmarks, and FMF's best practices for frontier AI safety evaluations.
- ***Section G*** suggests how the draft could be updated to more clearly address open foundation models in the context of the seven objectives articulated in NIST AI 800-1.

A. NIST AI 800-1 should incorporate the roles and responsibilities of the many actors in the value chain to manage misuse risk and not focus exclusively on 'initial developers'.

The current scope of NIST AI 800-1 is defined by the following statement:

"*The practices in this document are principally focused on the central role that foundation models' initial developers have in the supply chain for their models. These developers contribute most to determining how their models are made available, the models' capabilities, and safeguards against their misuse. ….. Other parties also play important roles in managing misuse risks, but they are not the focus of this document. They include downstream developers and deployers, third-party evaluators and auditors, civil society organizations, and government agencies. Relevant stakeholders throughout the AI supply chain are encouraged to share information and collaborate to understand and mitigate misuse risks, including to integrate appropriate risk mitigations into downstream systems that rely on foundation models*."

As participants in the open source AI ecosystem, including model developers, model deployers, end users, and researchers, we are concerned that this narrow scope does not adequately consider the roles of stakeholders other than models' initial developers – and potentially injects those developers into the independent decisions of deployers, end users, and researchers in ways that would be inappropriate. To illustrate the criticality of the downstream activities, consider the example of an organization that does not build its own foundation model (e.g. LLM), but chooses to use an existing commercial or open

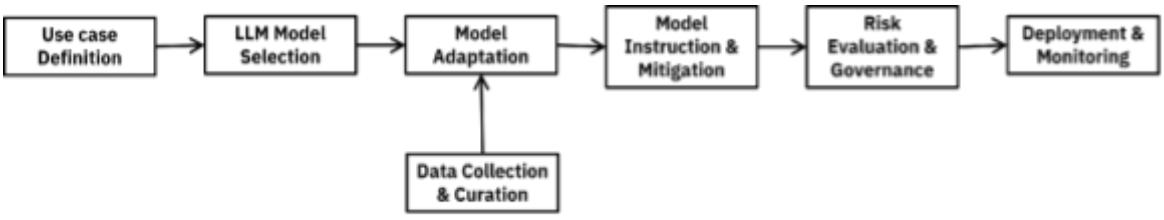

Figure 1: Typical steps to build an LLM based application.

foundation model as the starting point. Figure 1 describes the key steps in the LLM application development process for business/mission critical use.

These are the short definitions of the steps included in the diagram above:

- Use Case Definition: Clear enumeration of the task GenAI is supposed to do from the mission/business perspective and the related risk factors (e.g., accuracy, expected response time, various trustworthy AI attributes such as fairness).
- LLM Model Selection: Selection of an LLM from a library based on a set of evaluation criteria, such as fairness, explainability, uncertainty quantification, performance, etc. These are intrinsic model attributes that need to be assessed before a model is chosen, typically under the scope of the model developers.
- Data Collection and Curation: The collection and curation of both labeled and unlabeled data needed for the appropriate model adaptation technique of choice, such as prompt engineering, Retrieval Augmented Generation (RAG), fine-tuning, etc.
- Model Adaptation: Prompt engineering, RAG, fine-tuning, reprogramming, etc.
- Model Instruction and Mitigation: Model teaching, improvement, learning from human or AI feedback, human alignment.
- Risk Evaluation and Governance: Usage guidance, risk assessment, fact collection, model audit, policy packs, safeguards, etc.
- Deployment and Monitoring: Unlike traditional software, model monitoring is important in GenAI applications due to potential drifts due to unfamiliar data or unexpected emergent behavior.

Since these steps are undertaken by different actors – regardless of whether a model is or is not open sourced – a proper risk evaluation would need to consider the roles of each of the parties in the ecosystem.[3]

This consideration is particularly important because the same technical task (e.g., summarization) can have different risk expectations based on the specific use case in an application domain. As an example, summarizing the available lunch options in a local restaurant has a different risk compared to summarizing potential threats from a terrorist organization, demonstrating that the use case is a critical element of the risk assessment. While the model developer cannot feasibly anticipate all possible use cases, in many instances the downstream developer or deployer is better positioned to mitigate risks. In addition, each of the steps in Figure 1 can introduce new risks and also provide additional opportunities to mitigate these risks. To effectively address misuse risks, NIST should provide guidance on how to allocate risk evaluation responsibilities across downstream developers, deployers, and users, recognizing their unique positions to implement context-specific risk mitigations.[1]

Recommendation: *NIST should either broaden NIST AI 800-1 to encompass all ecosystem players or simplify the burdens placed on the initial model developed by requiring identification and mitigation of downstream misuse risks. This would avoid requiring model developers to micromanage downstream uses by deployers, align with open-source principles, and allow each ecosystem actor to focus on their most manageable risks.*

B. NIST should advocate for research to develop risk protocols, instead of presuming that developers can implement protocols that do not yet exist.

In Section 3, NIST AI 800-1 correctly identifies seven key challenges to risk management of foundation models, recognizing that many of these solutions do not yet exist. Despite this, NIST AI 800-1 proposes recommendations in Section 4 that assume these solutions are available to implement. For example, AI

[3] M. Srikumar, et al., Partnership on AI: Risk Mitigation Strategies for the Open Foundation Model Value Chain. 11 July 2024:
https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/.

models learn from training data and consequently there is no program code to define the application behavior linking inputs to outputs. Moreover, the document is asking developers to measure threat profiles without providing guidance on how to develop or prioritize them. Instead of doing this, NIST AI 800-1 should provide recommendations of key areas for further research and investment to guide industry in solving the most pressing problems.

***Recommendation**: NIST should recommend conducting research to develop safety protocols for the risks identified in NIST AI 800-1, as these protocols are not yet broadly established. This approach is more practical than expecting developers to implement non-existent protocols.*

C. NIST should adopt a marginal risk approach.

NIST should be consistent with the growing consensus that marginal risk[1,4] is the more appropriate and tested way to measure the risk. Marginal risk[5] is defined as "the risk presented by a new technology relative to risks posed by existing technologies." Practice 5.2.1 of NIST AI 800-1 mandates developers to "Implement safeguards designed to protect the model from misuse" without adequately weighing the benefits against the risks or considering the marginal risk. This approach goes against NTIA guidance[4]: "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 Frontier AI Forum (FMF)[6] also recommends the use of marginal risk: "When evaluations are intended to directly evaluate the risk posed by a system, in many cases they should consider evaluating the marginal risk relative to other applications." NIST should be consistent with its prior guidance and other experts and focus on marginal risk.

***Recommendation:** In its final report, NIST should explicitly align its guidelines with the marginal risk standard used by NTIA and independent experts, ensuring that the measures for managing misuse risk in foundation models are both practical and proportionate to the actual marginal risks posed.*

D. NIST AI 800-1 should present a balanced view of AI risks and rewards.

NIST should strike a balance between AI's risks and its vast benefits, fostering a policy environment that champions safe, innovative, and open AI development to boost innovation, economic growth, and societal progress.

The excessive focus on potential risks without recognizing the potential significant benefits will result in an overly restrictive environment that will deter responsible AI developers from innovating. At the same time, less responsible developers, potentially subject to different guidelines, are likely to continue to push forward cutting edge and potentially less safe models that do not align with democratic values. This, in turn, could prevent downstream deployers from accessing the latest models for new applications and use cases, limiting user choices. Moreover, it could discourage researchers from pursuing new AI applications, ultimately impeding innovation and technological progress. The AI Alliance strongly believes that a broad selection of AI models should be made available to empower developers, deployers, and users to innovate and select the best model for their specific use case.

---

[4] A. Basdevant, et al. "Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness in Artificial Intelligence." arXiv preprint arXiv:2405.15802 (2024).

[5] Harry Law, The marginal risk of AI: On evaluation, misinformation, and moral panic https://www.learningfromexamples.com/p/the-marginal-risk-of-ai

[6] Frontier Model Forum, Issue Brief: Early Best Practices for Frontier AI Safety Evaluations, 31 July 2024: https://www.frontiermodelforum.org/updates/early-best-practices-for-frontier-ai-safety-evaluations/.

A predominantly risk-focused approach may also slow down the economic growth anticipated from AI, which could increase global GDP by $10 trillion, or by as much as 10 percent.[7] Countries that adopt more openness in AI policies could outpace others, potentially leading to shifts in global economic and technological leadership, with the United States at risk of falling behind. It is central for economic vitality to create an environment that balances both the benefits and risks, promoting open and responsible AI development that contributes to economic prosperity.

The risk and benefits analysis is also explicitly outlined in the Executive Order[8] that directs the Department of Commerce to "submit a report to the President on the potential benefits, risks, and implications of dual-use foundation models for which the model weights are widely available, as well as policy and regulatory recommendations pertaining to those models."

Public perception of AI is also significantly shaped by the discourse surrounding it. An overly risk-focused narrative, without sufficient emphasis on the benefits, may amplify public fear and resistance towards AI that could otherwise offer significant societal advantages. It is important to develop a narrative that is balanced, emphasizing not only the risks but also the immense potential AI holds to enhance lives and address complex societal challenges.

Recommendation: *NIST AI 800-1 should present a balanced view of how AI can positively impact innovation, the economy, and society while ensuring proposed policies provide meaningful guidance for a safe and responsible AI.*

E. NIST AI 800-1 should not adopt policies that unnecessarily impact open foundation models.

The AI Alliance was created because of the value that open models create for a vast ecosystem of organizations, large and small companies, academic and non-profit institutions and society as a whole. Open foundational models, which consistently rank among the top performers on various leaderboards,[9] are creating positive societal impact and new business opportunities. Open models[10] provide broader access and greater customizability, enabling local adaptation and inference ability. They also mitigate monoculture and market concentration.

Open foundation models improve safety through greater transparency and access. Anyone can assess internal work and look for vulnerabilities that might be hidden in closed models. By allowing a broader community to evaluate, test and refine AI models independently, the models become more reliable and safe for the whole community. It is possible to maintain a balance between openness and safety, ensuring that open foundation models are not only innovative but also robust and trustworthy.

---

[7] JPMC Report. "Is generative AI a game changer?" February 14, 2024 https://www.jpmorgan.com/insights/global-research/artificial-intelligence/generative-ai

[8] Executive Order 14110. "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence": https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[9] Hugging Face, Open LLM Leaderboard: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

[10] S. Kapoor, et al., "On the Societal Impact of Open Foundation Models", https://arxiv.org/pdf/2403.07918v1

Multiple agencies in the US government have already recognized the value of open foundation models. The Cybersecurity and Infrastructure Security Agency's (CISA)[11] states: “We see significant value in open foundation models to help strengthen cybersecurity, increase competition, and promote innovation.” The FTC[12] has recognized the advantages of open models for innovation and competition and highlighted its potential benefits for the public. The NTIA report on Dual-Use Foundation Models with Widely Available Model Weights similarly finds[13] that “the government should not restrict the wide availability of model weights for dual-use foundation models at this time. Instead, the U.S. government should actively monitor and maintain the capacity to quickly respond to specific risks across the foundation model ecosystem, by collecting evidence, evaluating that evidence, and then acting on those evaluations.”

NIST AI 800-1 makes recommendations that disproportionately impact open models including recommending that model developers are able to rescind model access, monitor its usage, or moderate its behavior during and after deployment (see Practices described in Objective 5 and Objective 6). This is inconsistent with the very nature of open foundation models and much of the value they create originates from not being subject to centralized control. These recommendations would also mean that model developers would have to be intimately involved in the operation of independent businesses in challenging ways. These issues could be resolved if NIST AI 800-1 recognized that some of these recommendations could be more easily implemented by other participants in the AI ecosystem.

Open foundation models also have safety advantages, by providing greater transparency and access. Open models are easier to do research on and can facilitate better understanding of vulnerabilities applicable to all models. By allowing a broader community to evaluate, test, and refine AI models independently, the models can become more reliable and safer for the whole community.

NIST AI 800-1 should focus on neutral frameworks that do not unnecessarily single out a specific type of model and assign responsibility to the party best able to mitigate that risk.

Recommendation: *NIST should recognize the value of open foundation models, including from a safety perspective, and consider the impact of its recommendations on open models and focus on flexible risk mitigating recommendations consistent with different types of models and deployments.*

F. The NIST AI 800-1 should drive consistent risk management guidelines.

AI transcends national boundaries, making it crucial for frameworks to be aligned across jurisdictions to promote interoperability. Fragmentation and undue restrictions on AI can have significant repercussions, potentially reversing the very progress that characterizes this dynamic field. Fragmentation is particularly challenging for small and medium sized companies.

- Fragmented frameworks can lead to costly compliance burdens that smaller companies cannot afford, further consolidating the AI ecosystem with the largest players. This can discourage the widespread adoption and use of AI models, hindering the open exchange of ideas and innovations needed for rapid progress in the field. Low-resource organizations particularly

[11] CISA. With Open Source Artificial Intelligence, “Don’t Forget the Lessons of Open Source Software”: https://www.cisa.gov/news-events/news/open-source-artificial-intelligence-dont-forget-lessons-open-source-software.

[12] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models

[13] NTIA Report. "Dual-Use Foundation Models with Widely Available Model Weights", July 2024: https://www.ntia.doc.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

struggle in this environment. The uncertainty and risks introduced by inconsistent policies across jurisdictions make it challenging for startups to plan strategically and secure investments. Higher barriers to market entry can reduce competitiveness and introduce challenges in scaling, innovation, and collaboration that may negate or minimize the ability of open models to provide the greater competition necessary to avoid monopolies.

- Restrictive guidelines could result in access limitations. Historically, one of the pillars of AI advancement has been open innovation. This openness allows a broad range of researchers and developers to test, iterate, and innovate, thereby accelerating the pace of advancements and broadening the application of AI technologies across different sectors and making them safer. However, if access is limited to a select few, the potential for these widespread benefits is curbed.

NIST AI 800-1 can play an important role in synthesizing the guidance issued from the different government bodies and organizations.

Recommendations: *To overcome the challenges posed by fragmentation and leading to undue restrictions, NIST AI 800-1 should synthesize and build upon existing frameworks such as NIST AI 600-1, the White House Executive Order, ML Commons AI safety benchmarks, and FMF's best practices for frontier AI safety evaluations.*

G. Recommendation on the seven objectives of NIST AI 800-1 in light of open foundation models.

The sections above highlight our concerns with NIST AI 800-1 and the way that it is unnecessarily restrictive of open foundation models. In this section we offer specific suggestions for the seven objectives in the guidance.

Objective 1: Anticipate potential misuse risk

Though it is challenging for open foundation model developers to anticipate all potential misuse risks as NIST requires, they can and should identify the most significant ones and implement practical measures to minimize potential misuse risk. NIST can help with this by identifying the risk areas foundation model developers should focus on.

Objective 2: Establish plans for managing misuse risk

Managing plans for misuse risk for open foundation models as required by NIST needs to take into account unique characteristics of open foundation models, which means that the guidance should include collaborative, adaptive, and evolving risk mitigation strategies. Foundational model developers should concentrate on addressing the risks they are most equipped to mitigate, for which NIST can offer guidance.

Objective 3: Manage the risks of model theft

NIST AI 800-1 should recognize that many models are intentionally open source, and as such, the risk of model theft is minimal or nonexistent. It is important that this objective does not inadvertently imply that models should be closed rather than open source. The focus should be on appropriately managing

misuse risks without discouraging the development and use of open models, which often benefit from transparency and community collaboration. Therefore, the objective should be revised to clarify that the emphasis on confidentiality is only necessary for proprietary models and when the risk of model theft is significant and cannot be mitigated by other means.

Objective 4: Measure the risk of misuse

To be more helpful, NIST AI 800-1 should provide a framework to measure the risk of misuse.

The misuse can be due to various reasons such as an accidental action of a downstream deployer due to inadequate knowledge or experience with AI, or the deliberate act of someone to cause harm. Aurora-M[14] is an example of an open-source multilingual model fine-tuned on human-reviewed safety instructions, thus aligning its development not only with conventional red-teaming considerations, but also with the specific concerns articulated in the Executive Order 14110. Kaur et al.[15] introduced a unique dataset containing adversarial examples in the form of questions, called AttaQ, designed to provoke harmful or inappropriate responses. These are examples of AI that help to assess potential misuse. NIST should support the development of a framework that enables tools to measure and better understand these misuse risks.

Objective 5: Ensure that misuse risk is managed before deploying foundation models

NIST should recommend that open foundation model developers do redteaming to target potential misuse and appropriate finetuning of the model before it is released. NIST should also encourage model developers to provide user guides and safety tools for the downstream deployers to adapt the model safely for their use. While the open foundation model developers have a responsibility to focus on designing and training their models to operate safely and responsibly, they cannot control how others may choose to use or modify their work.

Objective 6: Collect and respond to information about misuse after deployment

Instead of discouraging the broad distribution of AI model weights, the NIST final report should promote practical mechanisms that can be implemented by developers of open models. Specifically, NIST should recommend:

- Establishment of output feedback mechanisms: Developers should set up systems to collect feedback on potentially problematic outputs, such as content related to criminal activities, regulated substances, or harmful behavior (e.g., hate speech or bullying). This feedback mechanism will help in identifying and mitigating issues promptly.

[14] T. Nakamura, et al. "Aurora-m: The first open source multilingual language model red-teamed according to the us executive order." arXiv preprint arXiv:2404.00399 (2024).

[15] G. Kour, et al. "Unveiling Safety Vulnerabilities of Large Language Models." arXiv preprint arXiv:2311.04124 (2023).

- Creation of bug bounty programs: Developers should encourage the community to report security vulnerabilities including through bug bounty or reporting programs. This not only helps in enhancing the security of the models but also engages the community in a constructive way.
- Commitment to continuous improvement: Making AI model weights widely available should be coupled with a commitment to address and resolve issues as they arise. Open models promote a transparent feedback loop, allowing for continuous improvement in safety, fairness, and security. This approach builds trust over time as the organization demonstrates its dedication to responsible practices and elevating standards for AI releases.

Objective 7: Provide appropriate transparency about misuse risk

There are various methods to ensure transparency regarding misuse risk. NIST's recommendations should be flexible, not mandating specific methods that may be incompatible with certain models or use cases, but rather accommodating diverse approaches that achieve the same objectives and are adaptable to different models. Many techniques cited in Practice 7.1 for transparency are already available for many open foundation models. These include, "Holistic Evaluation of Language Models"[16] (HELM) metrics for pre-release evaluation of model performance and Foundation Model Transparency Index[17] across 23 subdomains, such as data, compute, risks, mitigation, etc. However, it can be difficult to capture experiences of misuse incidents and hazards (discussed in Practice 7.3) involving open foundation models during deployment. For example, it may require reviewing media reports and examining social media accounts. Structured feedback mechanisms can help to address misuse risk.

Conclusion

The AI Alliance values this opportunity to highlight the importance of safeguarding widely available foundation model weights within the framework of the 'Managing Misuse of Dual-Use Foundation Models' guidance. We look forward to additional opportunities to demonstrate how open innovation is crucial to realizing many of the benefits of AI advancements.

[16] P. Liang, et al. "Holistic evaluation of language models." arXiv preprint arXiv:2211.09110 (2022).

[17] R. Bommasani, et al., "The Foundation Model Transparency Index v1. 1: May 2024." arXiv preprint arXiv:2407.12929 (2024).