Center for Research on Foundation Models

Stanford | RegLab

Stanford University Human-Centered Artificial Intelligence

Department of Commerce
National Institute of Standards and Technology
U.S. Artificial Intelligence Safety Institute
Docket No. 240802-0209
XRIN 0693-XC137
Request for Comment on the U.S. Artificial Intelligence Safety Institute's Draft Document: Managing Misuse Risk for Dual-Use Foundation Models

September 9, 2024

Introduction

Overall, we agree with and support the U.S. AI Safety Institute's (US AISI) draft guidelines (hereafter "the guidelines") for improving the safety, security, and trustworthiness of dual-use foundation models, which were issued in line with obligations under the October 2023 Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI). The guidelines provide useful and actionable recommendations for model developers to manage misuse risk. We encourage the US AISI to develop similar guidance for other actors in the foundation model supply chain as we believe model developers can contribute to, but are not best positioned for, mitigating all types of risk.[1] In addition, while not captured by the Executive Order's focus on foundation model misuse, we encourage the US AISI to develop guidance for non-misuse risks. While elements of the guidance (e.g., API monitoring) are more appropriate for some foundation model release strategies, we encourage the US AISI to explicitly affirm that these should not dissuade the open release of foundation models absent evidence of marginal risk.[2]

---

[1] Sarah Huiyi Cen et al., "AI Supply Chains," updated May 5, 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4789403; Rishi Bommasani et al., "Ecosystem Graphs: The Social Footprint of Foundation Models," March 28, 2023, https://arxiv.org/abs/2303.15772; Arvind Narayanan and Sayash Kapoor, "AI Safety Is Not a Model Property," March 12, 2024, https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property; Madhulika Srikumar, Jiyoo Chang, and Kasia Chmielinki, "Risk Mitigation Strategies for the Open Foundation Model Value Chain," July 11, 2024, https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/.

[2] National Telecommunications and Information Administration (NTIA). "Dual-Use Foundation Models with Widely Available Model Weights". July 30, 2024. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf; Response led by Stanford-Princeton on Open Foundation Models. Request for Comment on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights. March 27, 2024. https://hai.stanford.edu/sites/default/files/2024-03/Response-NTIA-RFC-Open-Foundation-Models.pdf; Response by AI Policy and Governance Working Group. Request for Comment on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights. March 27, 2024. https://www.ias.edu/sites/default/files/AIPGWG-Response_NTIA-RFC-on-Open-Foundation-AI-w-Available-Model-Weights_Updt_Mar2024.pdf

We make three targeted recommendations to improve the (draft) guidelines:

1. The guidelines should guide developers on *how* to make their evaluations reproducible.
2. The guidelines should guide developers to actively support *third-party* evaluations.
3. The guidelines should guide developers on *how* to monitor model usage post-deployment.

1. Strengthen guidance on reproducible evaluations

Reproducible evaluations are more trustworthy because third parties can replicate and verify evaluation results. Reproducible methodologies have, for example, facilitated the discovery of errors in science.[3] As part of the documentation for Practice 4.1, the guidelines recommend the inclusion of "a methodological description for each evaluation in enough detail to reproduce it," which advances reproducible evaluations.[4] However, by not clarifying what a "methodological description … in enough detail to reproduce it" would entail, the guidelines are insufficient for attaining the full benefits of reproducibility. Therefore, we recommend that the guidance include *specific* information that should be reported, including artifacts that go beyond the methodological description.

First, we consider a "methodological description" to be too vague to be actionable: Developers may interpret this language differently and, unintentionally, fail to provide key elements required for reproducing evaluations. Therefore, we recommend that this language should enumerate key elements for evaluation reproducibility such as data collection, preprocessing, prompts, inference parameters, evaluator models, etc.

Second, we consider "methodological description" to be often insufficient: For many evaluations conducted by model developers, a simple description of the methodology will not enable the reproduction of results. For example, according to the May 2024 Foundation Model Transparency Index, only three of 14 major developers conduct evaluations for intentional harm that are externally reproducible.[5] Prior work demonstrates that seemingly minor implementation details can have a large impact on the evaluation outcomes.[6] Therefore, the release of additional artifacts like evaluation code, evaluation data, proxy models/tasks, and evaluator models is crucial.[7] Ideally, these resources would be made publicly available to reduce the cost for third parties to reproduce the evaluation. Since other considerations, such as proprietary information contained in codebases, may countervail full release, the guidelines could provide guidance on

[3] Sayash Kapoor et al., "REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science," *Science Advances* 10, no. 18 (May 2024), https://www.science.org/doi/full/10.1126/sciadv.adk3452, eadk3452.
[4] Practice 4.1, Line 34.
[5] Rishi Bommasani et al., "The Foundation Model Transparency Index," May 2024, https://crfm.stanford.edu/fmti/.
[6] Stella Biderman et al., "Lessons from the Trenches on Reproducible Evaluation of Language Models," updated May 29, 2024, https://arxiv.org/abs/2405.14782.
[7] Hakan Inan et al., "Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations," December 7, 2023, https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations.

how developers could better navigate potential trade-offs by providing partial access. For example, in lieu of the full repository, developers could release code for processing evaluation inputs into prompts, which would at least benefit reproducibility without including proprietary information or data.

The release of evaluation artifacts offers benefits beyond reproducibility and will further safety research. For example, public evaluation artifacts allow the scientific community to scrutinize and improve standards for evaluation.[8] Xie et al. (2024) analyze existing datasets for language model safety training and find that "prior datasets are often built upon course-grained [sic] and varied safety categories, and that they are overrepresenting certain fine-grained categories." Based on these insights, they propose "a fine-grained 45-class safety taxonomy across 4 high-level domains."[9]

In addition to the generic benefits of these evaluation artifacts, there are specific benefits due to the unique position of model developers. Model developers, especially those who monitor the downstream distribution channels through which their models are used (e.g., via APIs), are best positioned to understand how models are used in practice and to design more ecologically valid evaluations informed by this context. Developers can conduct evaluations using data and insights from real-world post-deployment usage, which are not usually available to academic researchers. Therefore, by sharing evaluation artifacts broadly, developers can expand who has access to this rarefied information and contribute to greater collective safety.

Third, we emphasize that reproducibility is desirable yet insufficient for realizing the full benefits of safety evaluations. To properly interpret evaluation results, the AI community needs to understand the relationship between the training data for the foundation model and the evaluation data. Prior work has shown that (dangerous) capabilities may be overestimated because of leakage between training and test data.[10] Therefore, we support the guidelines' existing language on minimizing overlap between training and evaluation data,[11] but advocate for it to be strengthened to cover the public reporting of information about train-test overlap given overlap may not always be successfully minimized.

Overall, the guidelines for reproducible evaluations should apply to *all* relevant evaluations. In particular, estimates of model capabilities (Practice 1.3), model red-teaming (Practice 4.2), and

[8] Bochuan Cao et al., "Defending Against Alignment-Breaking Attacks via Robustly Aligned LLM," updated June 12, 2024, https://arxiv.org/abs/2309.14348.
[9] Tinghao Xie et al., "SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal Behaviors," June 20, 2024, https://arxiv.org/abs/2406.14598.
[10] Arvind Narayanan and Sayash Kapoor, "GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question," March 20, 2023, https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks; Arvind Narayanan and Sayash Kapoor, "Leakage and the Reproducibility Crisis in Machine Learning-Based Science," *Patterns* 4, no. 9, September 8, 2023, https://www.cell.com/patterns/pdfExtended/S2666-3899(23)00159-9.
[11] Practice 4.1, Lines 28-29.

assessment of the efficacy of safeguards (Practice 5.2) are all evaluations that will be more useful if they are reproducible. Just as before, reproducibility isn't all or nothing: Guidelines should push model developers to make a best effort toward reproducible evaluations while accounting for other considerations (e.g., proprietary data, exploitation by malicious actors).

2. Strengthen guidance on third-party evaluations

Third-party evaluations play a fundamental role in providing greater scrutiny and accountability: They naturally complement developer evaluations and are necessary even when developer evaluations are reproducible.[12] In general, we distinguish (i) the reproducibility of a developer's evaluation from (ii) the ability for third parties to conduct evaluations.[13] The guidelines should encourage developers to allow third parties to evaluate, because third parties can independently *specify* what to evaluate for, which provides greater accountability than reproducibility.

Prominent developers have already taken decisive steps to support—or indicate their support for—third-party evaluations. To solidify these claims, the guidelines should clearly and completely address the protections required for researchers to perform third-party evaluations safely.[14] Along these lines, we are pleased to see the inclusion of Practice 6.4, recommending developers to "provide safe harbors for third-party safety research."[15] As this guidance reflects, we should proactively avoid past failures: Researchers have less effectively researched social media platforms due to terms-of-service restrictions, and researchers have faced legal threats for their good-faith computer security work. As Longpre et al. (2024) have proposed, AI research may experience similar "chilling effects" and "incentives to tackle the wrong problems" in the absence of appropriate protections.[16] In particular, the guidance can be strengthened by recommending a safe harbor and specifying critical elements in such protections. For example, developers should provide mechanisms for researchers to appeal terms-of-service violations, specify what information is sufficient to deem research as good faith, justify their enforcement actions, and state any expectations of advanced disclosure clearly. In addition, developers should make clear what specific terms in their terms-of-service do not apply to demonstrably good-faith research to reduce confusion and unintended suppressive effects on such research.

The guidelines should also address the intersection between reproducible evaluations and third-party evaluations. In particular, the guidelines should encourage developers to push for reproducible and interpretable evaluations when working with third-party evaluators. For

[12] Inioluwa Deborah Raji et al., "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, July 27, 2022, https://dl.acm.org/doi/abs/10.1145/3514094.3534181.
[13] Bommasani et al., "The Foundation Model Transparency Index."
[14] Kevin Klyman, Sayash Kapoor, and Shayne Longpre, "A Safe Harbor for AI Researchers: Promoting Safety and Trustworthiness through Good-Faith Research," Federation of American Scientists, June 28, 2024, https://fas.org/publication/safe-harbor-for-ai-researchers/.
[15] Practice 6.4, Line 14.
[16] Shayne Longpre et al., "A Safe Harbor for AI Evaluation and Red Teaming," March 7, 2024, https://arxiv.org/abs/2403.04893.

example, when developers like Anthropic work with biosecurity firms like Gryphon to measure offensive biological capabilities, these assessments will be more valuable if they can be rigorously understood by the broader scientific community.[17] Third-party evaluations can provide valuable insight and interrogation that is distinct from developer evaluations, but these benefits will only be realized fully if the broader scientific community can understand the results and judge them to be credible.

3. Clarify guidance on post-deployment monitoring

While many AI policy efforts have focused on pre-deployment evaluations and risk mitigation, we welcome the focus on post-deployment usage and societal outcomes in the US AISI guidelines. In particular, many aspects of model risk are difficult to predict: Understanding and releasing information about model usage will allow developers, researchers, and the government to study harms that materialize. For example, research on model usage information can allow policymakers to "rely less on their potentially misguided intuitions about risk and more on data about where those risks are actually occurring."[18]

To better understand post-deployment outcomes, we recommend a multipronged strategy. Many conceptual and practical challenges complicate the understanding of post-deployment outcomes. By way of distribution channels, developers can better model usage and communicate this information in aggregate to the public (e.g., via transparency reports).[19] We emphasize that social media has established precedent that this is possible for major digital technology: Social media platforms regularly release reports on platform usage and, specifically, the prevalence of harmful content.[20] In concert with developer-centric interventions, we encourage the guidelines to specifically endorse adverse event reporting for the misuse of foundation models. Adverse event reporting would enable regulators to better track emerging risks,[21] building on the National Artificial Intelligence Advisory Committee's recommendation to pilot an adverse event reporting system for "post-deployment events stemming from AI systems."[22] Overall, the guidelines

---

[17] Anthropic, "Frontier Threats Red Teaming for AI Safety," July 26, 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

[18] Gabriel Nicholas, "Grounding AI Policy: Towards Researcher Access to AI Usage Data," Center for Democracy & Technology, August 13, 2024, https://cdt.org/insights/grounding-ai-policy-towards-researcher-access-to-ai-usage-data/.

[19] Arvind Narayanan and Sayash Kapoor, "Generative AI Companies Must Publish Transparency Reports," Knight First Amendment Institute, June 26, 2023, https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports; Rishi Bommasani et al., "Foundation Model Transparency Reports," February 26, 2024, https://arxiv.org/abs/2402.16268.

[20] Narayanan and Kapoor, "Generative AI Companies Must Publish Transparency Reports." These transparency reports are also legally codified in the European Union under the Digital Services Act.

[21] Neel Guha et al., "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," *George Washington Law Review* (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634443.

[22] The National Artificial Intelligence Advisory Committee (NAIAC), "RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting," November 2023, https://ai.gov/wp-content/uploads/2023/12/Recommendation_Improve-Monitoring-of-Emerging-Risks-from-AI-through-Adverse-Event-Reporting.pdf.

should more aggressively reduce the informational gaps that pervade the ecosystem on how foundation models are used in practice, including for malicious purposes.

Alongside improved monitoring, developers should continually reassess the practices in Objectives 1, 4, and 5 post-deployment. After deployment, new threats may emerge (Practice 1.1), the impacts of known threat profiles may change (Practice 1.2), model capabilities may change with newly discovered inference-time enhancements (Practice 4.1), new exploits for models may be discovered (Practice 4.2), and new safeguards may be discovered (Practice 5.2). Risk assessment should be an on-going process, especially when considering the fast-moving nature of the field.[23] For the same reason, Practice 4.1 could be modified to include a recommendation for model developers to stay up to date with research to keep track of how their released models can be augmented to increase performance on capabilities of interest. Therefore, the guidelines should recommend the *periodic* assessment of risks and mitigations.

Finally, the *pre-deployment* risk management practices should more clearly account for the changing nature of model usage post-deployment. Specifically, guidelines should account for the additional uncertainty resulting from evaluating models in a testing environment. For example, while we agree with the guidance that developers should "assess what a threat actor could achieve given access to the weights of a model and the ability to integrate it with other tools" regarding the assessment of the risk of misuse,[24] we also advocate for this to be broadened to all forms of test-time improvements.

More generally, the evaluation of foundation models suffers from an "elicitation gap": Measured capabilities often do not reflect the best possible enhancements for maximizing capabilities.[25] For example, this elicitation gap is hard to characterize as "methods that elicit improved model capabilities are sometimes discovered only after a model has been deployed."[26] Models often benefit greatly from improvements in prompting that do not require access to weights or tool-use. As a signature example, chain-of-thought prompting considerably improves the reasoning capabilities of models.[27]

---

[23] National Institute of Standards and Technology (NIST), "AI Risk Management Framework: Second Draft," August 18, 2022, nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf, 10.
[24] Practice 4.1, Line 22.
[25] METR's Autonomy Evaluation Resources, "Measuring the Impact of Post-Training Enhancements," https://metr.github.io/autonomy-evals-guide/elicitation-gap/.
[26] Yoshua Bengio et al., *International Scientific Report on the Safety of Advanced AI*, May 2024, https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.
[27] Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" (paper presented at the 36th International Conference on Neural Information Processing Systems, December 6, 2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html, 24824–37.

In addition, incorrect formatting of the model input can have a large impact on model performance. For example, one study found that language model performance can differ by up to 76 accuracy points simply due to changes in prompt formatting.[28] As such, the performance during testing should be treated as a "lower-bound", or potential underestimate, of the actual capability of the model. Specifically, Practices 4.1 and 1.3 could account for the uncertainty resulting from testing conditions that may not reflect real-world usage. More specific recommendations on, for example, experimenting with different prompts could also be helpful in reducing the gap between testing and real-world performance. By incorporating these considerations into pre-deployment practices, developers can better account for potential risks associated with the evolving capabilities and usage of foundation models in the real world.

We thank the Commerce Department, NIST, and AISI for the opportunity to share our views, which are based on our scientific research in these areas. Please email nlprishi@stanford.edu with any comments or questions.

Sincerely,

Rishi Bommasani
Society Lead, Stanford Center for Research on Foundation Models
Ph.D. Candidate, Stanford University

Alexander Wan
Researcher, Stanford Center for Research on Foundation Models
Undergraduate, University of California, Berkeley

Yifan Mai
Research Engineer, Stanford Center for Research on Foundation Models
Institute for Human-Centered Artificial Intelligence (HAI), Stanford University

Percy Liang
Director, Stanford Center for Research on Foundation Models
Associate Professor of Computer Science and (by courtesy) of Statistics, Stanford University

Daniel E. Ho
Director, Stanford Regulation, Evaluation, and Governance Lab
William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science and (by courtesy) Computer Science, Stanford University

[28] Melanie Sclar et al., "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting," July 1, 2024, https://arxiv.org/abs/2310.11324.