PRINCETON UNIVERSITY

CENTER FOR INFORMATION TECHNOLOGY POLICY
CITP.PRINCETON.EDU

Department of Commerce
National Institute of Standards and Technology
Docket number 240802-0209
XRIN 0693-XC137
Request for Comments on the U.S. Artificial Intelligence Safety Institute's Draft Document: Managing Misuse Risk for Dual-Use Foundation Models

September 9, 2024

To Whom It May Concern:

We are researchers from Princeton University's Center for Information Technology Policy (CITP) writing to offer the following submission in response to the Request for Comment (RFC) by the National Institute of Standards and Technology on draft guidelines for managing misuse risk for dual-use foundation models (the "Guidelines").[1]

We commend the U.S. AI Safety Institute for developing this framework to address the critical issue of misuse risks in dual-use foundation models. We particularly appreciate the emphasis on researcher access and transparency, which are crucial for fostering an open and collaborative approach to AI safety.

At the same time, we believe there are several areas where the Guidelines could be strengthened to address the evolving landscape of AI capabilities and potential threats. Our comments focus on three main areas: 1) the risk analysis for model development should include offensive AI agents, 2) supplementing model red teaming with a focus on downstream attack surfaces, 3) the approach to model release and deployment strategies should be revised with a focus on marginal risk.

1. Offensive AI agents are an important category of risk that the Guidelines should address

While the current guidelines provide valuable insight into evaluating and safeguarding individual AI models, we believe this focus does not capture the most likely range of potential misuse risks.

Many of the most significant real-world threats are likely to arise not from standalone models, but from AI systems deployed in agentic settings — i.e., AI agents that can take actions and interact with their environment over time.[2]

[1] In keeping with Princeton's tradition of service, CITP provides nonpartisan research, analysis, and commentary to policy makers, industry participants, journalists, and the public. This response reflects the independent views of the undersigned scholars.
[2] Sayash Kapoor et al., "AI Agents That Matter" (arXiv, July 1, 2024), https://doi.org/10.48550/arXiv.2407.01502.

For example, an AI system designed to autonomously find and exploit software vulnerabilities poses very different risks compared to a language model that can generate text about hacking techniques.[3] The agentic system can potentially discover novel attack vectors, adapt to defenses, and execute multi-step attack chains without human intervention.[4] While independent third parties are best placed for conducting evaluation on agents, there is an important role for model developers to release standardized tooling and evaluations to measure misuse risk.

Therefore, we recommend expanding the scope of the guidelines for model developers to explicitly consider offensive AI agents as part of safety testing and risk assessment.

Recommendations for model developers:

- Develop scenarios and testing frameworks for AI agents in domains like cybersecurity, influence operations, and autonomous weapons systems.
- Assess how foundation models could be leveraged as components in more complex AI systems and agents.
- Release agent testbeds to enable independent research (including comparison with other models, models of lower capability, and open models, to assess the marginal risk of releasing more capable models—see item 3 below for more.)

2. Red-team evaluations of models must be supplemented with efforts to detect misuse at actual attack surfaces

The guidelines rightly emphasize the importance of third-party evaluation and testing. We agree this is important for ensuring unbiased assessments.

However, we believe it's crucial to recognize that many of the most effective defenses against AI misuse will need to be implemented at the "attack surface" — i.e., the downstream sites where malicious actors would actually deploy AI-generated content or execute AI-aided attacks.

Model developers should share early access to models and tools to detect misuse easier to use with downstream attack surfaces. While the document points out the importance of tracking misuse across deployment vectors, we believe coordination with developers of downstream attack surfaces is crucial for improving resilience to AI risk.[5]

Recommendations for model developers:

- Identify key attack surfaces across various domains. For example, the attack surface for disinformation is typically a social media platform—that is where influence operators seek to disseminate disinformation and persuade people. For security vulnerabilities, the

[3] Gelei, Deng et al. "PentestGPT: An LLM-empowered automatic penetration testing tool." (arXiv, August 13, 2023), https://doi.org/10.48550/arXiv.2308.06782
[4] Andy K. Zhang et al., "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models." (arXiv, August 15, 2024), https://doi.org/10.48550/arXiv.2408.08926
[5] 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.

attack surface may be software codebases in critical infrastructure. Model developers should identify key attack surfaces for the main threats they identify.

- Develop best practices for hardening these attack surfaces against AI-enhanced threats. This could involve creating a coordination framework with downstream actors to enable information sharing about capabilities and risks that tilt the offense-defense balance in favor of defenders. For example, model developers could provide early model access to software developers in critical domains to help them find and fix security vulnerabilities. Similarly, even if an API to detect watermarked text outputs from an LLM is not publicly available, model developers can make it available to social media platforms to help detect bots. While more sophisticated threat actors can use open-weight models, this would still help detect bot-generated content from lower-resource actors across social media and other online platforms.

3. Reassess the model deployment decision with a focus on marginal risk

We believe that the guidelines' attention to the mitigation of misuse risk before deployment (objective 5) would be better served by a focus on the marginal risk of deploying models and releasing model weights.[6]

Marginal risk refers to the incremental risk of deploying a model or releasing its weights *over and above* the existing risk of already-released models as well as existing technology. For example, to assess the biosecurity risks of language models, it is essential to compare them against widely available existing technology such as information found via search engines and Wikipedia.[7]

In addition, when deployed as part of offensive agents, gains in model capability might not be required to achieve certain offensive attacks. Building task-specific improvements (such as program verification for coding agents) and scaling inference compute[8] might lead to similar increases in offensive capabilities as a new model generation.[9]

While the Guidelines offer guidance on comparing model capabilities to existing models, we recommend a more nuanced approach based on assessing the marginal risk of deploying a model or releasing its weights.

Recommendations for model developers:

- Develop frameworks to quantify marginal risk in various AI contexts. This would allow developers to focus efforts on scenarios where the marginal risk is demonstrably high.

[6] Sayash Kapoor and Rishi Bommasani et al., "On the Societal Impact of Open Foundation Models," February 27, 2024, https://arxiv.org/pdf/2403.07918v1.
[7] Neel Guha et al., "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," SSRN Scholarly Paper (Rochester, NY, November 15, 2023), https://papers.ssrn.com/abstract=4634443.
[8] Bradley Brown et al., "Large Language Monkeys: Scaling Inference Compute with Repeated Sampling" (arXiv, July 31, 2024), https://doi.org/10.48550/arXiv.2407.21787.
[9] Michael Hassid et al., "The Larger the Better? Improved LLM Code-Generation via Budget Reallocation." (arXiv, 25 July, 2024), https://doi.org/10.48550/arXiv.2404.00725

- For cases where marginal risk of model release is low (for example, if existing closed or open models are similarly risky when deployed as offensive agents or after scaling inference compute), consider alternatives to delaying or halting model release, such as coordinating with downstream attack surfaces and increasing societal resilience.
- Establish clear methodologies for comparing new AI capabilities to existing non-AI methods for achieving similar outcomes. As more capable models are released openly, regularly update marginal risk assessments.

Conclusion

The Guidelines represent an important step toward advancing AI safety that will be further enhanced with our recommendations. We appreciate the opportunity to provide feedback. Please do not hesitate to reach out if you have any questions or would like to discuss these ideas further.

Sincerely,

Sayash Kapoor
Researcher, Princeton Center for Information Technology Policy
Ph.D. Candidate, Princeton University

Mihir Kshirsagar
Technology Policy Clinic Lead, Center for Information Technology Policy

Arvind Narayanan
Professor of Computer Science, Princeton University
Director, Princeton Center for Information Technology Policy

Benedikt Stroebl
Researcher, Princeton Center for Information Technology Policy
Ph.D. Student, Princeton University