ABUNDANCE INSTITUTE

Request for Comments on the U.S. Artificial Intelligence Safety Institute's Draft Document: Managing Misuse Risk for Dual-Use Foundation Models

→ Neil Chilson, Head of AI Policy

REGULATIONS.GOV DOCKET NO. #240802-0209

SUBMITTED | September 9, 2024

The Abundance Institute is a mission-driven non-profit dedicated to creating the policy and cultural environment where emerging technologies can develop and thrive in order to perpetually expand widespread human prosperity. This comment is designed to assist the agency as it explores these issues. The views expressed in this comment are those of the author(s) and do not necessarily reflect the views of the Abundance Institute.

Introduction

We appreciate the opportunity to comment on NIST's draft guidance on managing misuse risk for dual use foundation models.[1] The suggested practices for model developers demonstrate thoughtful consideration. However, the draft guidance gives the incorrect impression that model developers are always best suited to prevent model misuse; is insufficiently clear about when to consider the benefits of model deployment; and lacks any guidance around the impact on user free expression. We propose three corresponding categories of changes to improve the guidance:

I. Clarify the role of model developers in reducing AI model misuse;
II. Encourage consideration of potential benefits when implementing safeguards; and
III. Encourage developers to protect freedom of expression.

I. Clarify the Role of Model Developers in Reducing AI Model Misuse

NIST should clarify that model developers are not the only stakeholders that can mitigate model misuse risk. The current draft is misleading on this point. The title and introductory framing of the draft guidance indicates a comprehensive approach to "misuse risk," defined as "[a] risk that an AI model will be deliberately misused to cause harm."[2] Such risks "result in part

1 NIST, Managing Misuse Risk for Dual-Use Foundation Models ("Draft Misuse Risk Guidance").

2 Id. at 18. Framing this as "model misuse risk" is itself somewhat convoluted. It would be awkward to say that a kitchen knife has a "murder risk" because a knife might be used in a stabbing. "Prevent murder" is a clear goal, but "minimize the murder risk of every pointy object" obfuscates that goal. Framing the effort to prevent and deter AI model misuse as "managing misuse risk" focuses attention away from the bad actor and toward the model developer.

from malicious actors' motivations, resources, and constraints, as well as society's defensive measures against that harm."[3] "As a result," the introduction promises, "the guidelines provided here address both technical and social aspects of these risks."[4] This would suggest that the guidelines intend to offer a comprehensive exploration of how to reduce AI model misuse.

Yet, despite that holistic framing, the draft guidance focuses narrowly on model developers. Indeed, despite acknowledging that "actors across the lifecycle of a model all play a role in managing misuse risks," the document suggests that the model developer bears primary responsibility for mitigating that risk.[5] It claims that model developers have a "central role" and "contribute most to determining … safeguards against [models'] misuse."[6] This framing unnecessarily discounts other potentially effective approaches to reducing misuse.

We have little evidence that model developers are uniformly and permanently the best-situated parties to prevent or deter misuse. Common sense (and common law) suggest that restricting or otherwise deterring those who would misuse a tool can also constrain misuse.[7] Preventing misuse in other industries typically involves a mix of product design, market mechanisms, social norms, tort law, civil and criminal penalties for bad actors, and regulation. How this mix of governance develops for AI models will depend on

3 *Id.* at 1. (Both the Introduction and the Scope sections of the draft guidance are on "page 1" according to the table of contents, although the Introduction page is unnumbered.)

4 *Id.*

5 *Id.* at 0.

6 *Id.*

7 See ICLE Comments on Managing Misuse Risk for Dual-Use Foundation Models at 3 (Sept. 9, 2024), https://laweconcenter.org/resources/icle-comments-on-managing-misuse-risk-for-dual-use-foundation-models/ (explaining the concept of "least-cost avoider" in tort law).

how feasible and effective it is for developers (as compared to others in the ecosystem) to predict potential misuse and limit it without constraining proper and intended uses of the tools.

Attempting to mitigate misuse risk during the training or deployment may make these models less powerful, more expensive, or less accessible for intended uses. It may be more efficient and effective to mitigate misuse risk at later steps in the supply chain for AI-powered tools. The best point in the supply chain could also differ based on the type of model and the developer's interactions with the end user.

Developers obviously have an important role in preventing misuse, but they are not the only ones, and perhaps not even the best suited ones. We are not suggesting that this guidance document ought to elaborate on all the potential actors and every method to mitigate misuse risk. But NIST should correct the current draft's overall impression that model developers are solely responsible for misuse. NIST can do so by:

→ clarifying that developer efforts to mitigate risk do not make developers liable;

→ emphasizing that the guidance does not absolve those who misuse tools; and

→ acknowledging that other approaches may more effectively mitigate misuse risk while facilitating intended uses.

II. Encourage Consideration of Benefits

NIST should explicitly encourage developers to consider the potential benefits of deploying AI models and how various practices and safeguards could affect those benefits. The draft guidance

recommends considering the benefits of releasing an AI model, but only in the context of setting acceptable risk thresholds under Objective 2.[8] Perhaps this is intended to suggest that developers assess all recommended practices for their potential effects on the benefits of releasing a model. In any case, NIST should expressly recommend that developers evaluate every recommended objective and practice to determine whether and how applying it will improve, preserve, or foreclose benefits from the AI model under consideration.

More specifically, benefits should be considered as a core part of the Objectives 2, 3, and 5. For example, any roadmap developed under Practice 2.2 ought to incorporate careful consideration of potential benefits of the AI model and what benefits may be foregone if one path is chosen over another. In addition, Practice 5.2 should recommend "[I]mplementing safeguards proportionate to the model's misuse risk" ***and its potential beneficial uses.*** Adopting safeguards with no consideration of their effects on beneficial uses could be counterproductive.

Developers of open source and open weight models have unique benefits that must be weighed when considering tactics to mitigate misuse risk. Transparency, collaboration, faster experimentation, security, robustness, and educational value are just some of these benefits. Without considering the benefits of model release, much of the draft guidance would disfavor the release of open model weights. For example, draft Objective 3 is in tension with the release of open model weights, unless one considers the potential benefits of release. We do not believe the draft guidance was intended to discount or ignore the beneficial uses of models. Still, NIST should revise it to make that perfectly clear.

8 Draft Misuse Risk Guidance at 7.

III. Encourage Developers to Protect Freedom of Expression

NIST should encourage developers to protect the freedom of expression of future model users. Generative AI models are powerful speech tools. Creating content is one of their core intended uses. Mitigating misuse should not come at the expense of user free expression. Thus, NIST should encourage developers to carefully evaluate how any misuse risk safeguards could hamper user speech.

Some of the example safeguards could potentially impact free expressions. For example, detecting and blocking attempted misuse, performed with a recommended "margin of safety," could directly halt speech that does not constitute misuse.[9]

Free expression considerations could build on the draft guidance's recommendations to consider the privacy impacts of various risk mitigation practices. For example, Practice 6.1.1 encourages model developers to "[m]onitor APIs, websites, and other distribution channels for misuse while maintaining privacy of users."[10] Elsewhere, Practice 6.1.5 encourages developers to consider using "tiered methods of detection when doing so helps … improve privacy …"[11] Such measures, without careful application, could not only threaten privacy but also chill user speech. They should be modified to better protect user speech interests.

9 *See id.* at 19 tbl.1 (describing the "detect and block attempted misuse" example safeguard) *and id.* at 13 (Practice 5.3(2) recommending "leaving a margin of safety between the estimated level of risk at the point of deployment and the organization's risk tolerance.").

10 *Id.* at 14.

11 *Id.*

IV. Conclusion

Addressing these three areas will significantly improve the draft guidance, providing a more balanced and clear approach to managing misuse risk in dual-use foundation models. We appreciate NIST's efforts in this important area and hope these suggestions contribute to more effective final guidance.