TECHFREEDOM
LAW FOR A DYNAMIC FUTURE

Comments of

TechFreedom

Andy Jung[i]

In the Matter of

Managing Misuse Risk for Dual-Use Foundation Models

NIST AI 800-1

September 9, 2024

[i] Andy Jung is Associate Counsel at TechFreedom, a nonprofit, nonpartisan technology policy think tank. He can be reached at ajung@techfreedom.org.

TABLE OF CONTENTS

INTRODUCTION

On October 30, 2023, President Biden signed Executive Order 14110 on *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*.[1] The Order directed the Department of Commerce, acting through the National Institute of Standards and Technology (NIST), to develop guidelines and best practices for AI safety and security.[2] The Order also directed the Department of Commerce to use the National Telecommunications and Information Administration (NTIA) to seek public comment "on the potential risks, benefits, other implications, and appropriate policy and regulatory approaches related to" open-source AI models.[3]

On July 26, 2024, NIST released a draft guidance document on *Managing Misuse Risk for Dual-Use Foundation Models*, NIST AI 800-1.[4] The guidance is "intended to help software developers mitigate the risks stemming from generative AI and dual-use foundation models—AI systems that can be used for either beneficial or harmful purposes."[5] NIST AI 800-1 focuses on the "initial developers" of AI models and "identifies best practices to map, measure, manage, and govern misuse risks."[6] NIST seeks public comment on the guidance, including how to "better address the ways in which misuse risks differ based on deployment (*e.g.*, how a foundation model is released)?"[7]

On July 30, 2024, NTIA released a *Report on Dual-Use Foundation Models with Widely Available Model Weights*, which concludes: "The current evidence base of the marginal risks and benefits of open-weight foundation models is not sufficient either to definitively conclude that restrictions on such open-weight models are warranted, or that restrictions will never be appropriate in the future."[8] To assess the overall risk of open-sourcing an AI model, NTIA recommends "a marginal risk and benefit analysis framework" that compares

---

[1] Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191 (2023), https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.
[2] *Id.* at 75196, § 4.1.
[3] *Id.* at 75203, § 4.6(a).
[4] U.S. AI SAFETY INST., NIST AI 800-1 (2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf.
[5] *Department of Commerce Announces New Guidance, Tools 270 Days Following President Biden's Executive Order on AI*, NIST (July 26, 2024), https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following.
[6] NIST AI 800-1 at Introduction.
[7] Request for Comments on the U.S. Artificial Intelligence Safety Institute's Draft Document: Managing Misuse Risk for Dual-Use Foundation Models, 89 Fed. Reg. 64878 (proposed Aug. 8, 2024), https://www.federalregister.gov/documents/2024/08/08/2024-17614/request-for-comments-on-the-us-artificial-intelligence-safety-institutes-draft-document-managing.
[8] NAT'L TELECOMM. & INFO. ADMIN., DUAL-USE FOUNDATION MODELS WITH WIDELY AVAILABLE MODEL WEIGHTS 47 (2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

"the additional risks and benefits that widely available model weights introduce compared to those that come from non-open foundation models or from other technologies more generally."[9]

NIST AI 800-1 fails to apply the marginal risk and benefit analysis used by NTIA to assess the safety of open-source AI models. Some of the best practices in NIST AI 800-1 arguably do not apply to open-source models at all, and those that do apply ignore the benefits of open source. Instead, NIST AI 800-1's main provisions direct open-source developers to implement safeguards against misuse of their public models regardless of comparable risks posed by other digital technologies, like closed-source models, or the overall benefit of open source. And many of NIST's safeguards would hinder or prevent open sourcing of AI models altogether.[10]

Executive Order 14110 tasked the Department of Commerce with developing standards and guidelines for deploying safe AI technologies. The agency, however, has sent mixed messages to open-source developers. By failing to apply marginal risk and benefit analysis to open-source AI development, NIST AI 800-1 targets open-source models "with restrictions that are unduly stricter than alternative systems that pose a similar balance of benefits and risks."[11]

I. NTIA Uses Marginal Risk and Benefit Analysis to Assess Risks Posed by Open-Source AI Models.

The NTIA report opens by acknowledging the trade-off between benefits and risks inherent to open-source AI models:

> Dual-use foundation models with widely available model weights (referred to in this Report as open foundation models) 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.[12]

---

[9] *See id.* at 10.
[10] *See, e.g.,* NIST AI 800-1 at 19, Appendix B ("Safeguard: Limit access to the model's capabilities.").
[11] *See* NAT'L TELECOMM. & INFO. ADMIN., DUAL-USE FOUNDATION MODELS WITH WIDELY AVAILABLE MODEL WEIGHTS 10 (2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.
[12] *Id.* at 2.

The Report notes, however, that open-source models "could also engender harms and risks to national security, equity, safety, privacy, or civil rights through affirmative misuse, failures of effective oversight, or lack of clear accountability mechanisms."[13] Open-source models create "marginal risks" that are "unique" to open-source "relative to risks from other existing technologies, including closed weight models."[14]

To assess the overall risk of open-source models, the NTIA report uses a "marginal risk and benefit analysis framework" to weigh "the additional risks and benefits" of open models "compared to those that come from" closed models or "other technologies more generally."[15] "Public commenters generally agreed that []marginal risk and benefit analysis [] is appropriate" to understand the implications of open-source models.[16]

NTIA's framework recognizes that, because open-source software has unique risks and benefits, open-source developers must tailor their risk mitigation strategies:[17]

> Risks from open models and closed models should both be managed, though the particular mitigations required may vary. In some cases, managing the risk of open models may pose unique opportunities and challenges to reduce risk while maintaining as many of the benefits of openness as possible.[18]

II. NIST AI 800-1 Ignores Marginal Risk and Benefit Analysis.

NIST AI 800-1 "identifies best practices to map, measure, manage, and govern misuse risks from foundation models, as well how organizations can provide transparency into how they are managing these risks."[19] The best practices, however, do not weigh the marginal risks and benefits of open models relative to closed models or other technologies. Overall, the guidance is inherently hostile to open sourcing AI models.

NIST AI 800-1 applies to "models' initial developers," although the guidance acknowledges that "[o]ther parties also play important roles in managing misuse risks … includ[ing] downstream developers and deployers, third-party evaluators and auditors, civil society

---

[13] *Id.*
[14] *Id.* at 3.
[15] *Id.* at 10.
[16] *Id. See also id.* at 10 n.30 (public comments made to NTIA endorsing marginal risk and benefit analysis for open-source AI models).
[17] *See* Madhulika Srikumar, et al., *Risk Mitigation Strategies for the Open Foundation Model Value Chain*, PARTNERSHIP ON AI (July 11, 2024), https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/.
[18] NAT'L TELECOMM. & INFO. ADMIN., DUAL-USE FOUNDATION MODELS WITH WIDELY AVAILABLE MODEL WEIGHTS 10 (2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.
[19] NIST AI 800-1 at Introduction.

organizations, and government agencies."[20] The guidance does not distinguish between open and closed-source developers: the best practices apply equally to both.

NIST AI 800-1's guidance, however, is ambiguous and inconsistent as applied to open-source models. Some of NIST's best practices, like those designed to prevent model theft,[21] do not apply to open-source at all—because open models are already freely available to the public. It is unclear how, or even whether, open-source developers could comply.

The core best practices in NIST AI 800-1 would apply to open-source models; the practices, however, fail to incorporate the marginal risk and benefit analysis used in the NTIA report. Practice 1.3 and Practice 4.1, for example, incorporate quasi-marginal-risk-analysis by directing developers to estimate and measure the misuse risk of their models by "comparing the characteristics of the upcoming model with existing models."[22] But the practices do not weigh these risks against the corresponding benefits of open-sourcing.

Practices 5.1, 5.2, and 5.3 are flatly hostile to open-source models. NIST AI 800-1 directs open-source developers to "[t]ake actions to increase access to the model (*e.g.*, deploying a model via API or releasing its weights) only when misuse risks are adequately managed, including that they are at minimum within the organization's risk tolerance."[23] The focus on minimizing risk before deploying a model without counterbalancing the benefits of increased access runs counter to NTIA's marginal risk and benefit analysis for open source.

Practice 5.1 directs developers to "[a]ssess the effect of a potential deployment on the model's misuse risk,"[24] and Practice 5.3 allows developers to pursue deployment only "where misuse risk is adequately managed … based on the assessed misuse risk and a consideration of any safeguards."[25] NIST recommends that developers "[i]dentify the level of access that a malicious actor could obtain under each proposed deployment (*e.g.*, would it grant them access to API inference, access to a fine-tuning API, access to model weights) and consider how the deployment may affect misuse risk."[26]

Under an open-source deployment, benign and malicious actors alike have nearly full access to a model, including access to weights and fine-tuning capabilities. The guidance stresses the risks of open-sourcing—"For example, allowing fine-tuning via API can significantly limit options to prevent jailbreaking and sharing the model's weights can significantly limit

[20] *Id.* at 1.
[21] *Id.* at 8 (Practice 3.1), 9 (Practice 3.3).
[22] *Id.* at 6.
[23] *Id.* at 12.
[24] *Id.*
[25] *Id.* at 13.
[26] *Id.* at 12.

options to monitor for misuse (Practice 6.1) and respond to instances of misuse (Practice 6.2)"[27]—without weighing countervailing benefits of openness.

If developers fail to manage these risks, many of which are inherent to open-source models, NIST AI 800-1 suggests that their "deployment should be modified, delayed, or canceled."[28] In order to comply, Practice 5.2 directs developers to "[i]mplement safeguards proportionate to the model's misuse risk,"[29] and Appendix B lists possible safeguards.[30]

Many of NIST's safeguards undermine open source, pressuring developers to limit access and keep models closed. Appendix B, for example, recommends that developers "[l]imit access to the model's capabilities," including "[l]imit[ing] the ability to interact with the model to contexts and users where the misuse risks are lower" and "[r]educ[ing] access to the model reactively when misuse is detected … such as by rolling back a model to a previous version."[31] These vague "safeguards" directly prevent open-sourcing a model: limiting access to capabilities and users is antithetical to the openness inherent in "open source" software, and open-source developers are unable to roll back models once they are publicly distributed.

Appendix B also directs developers to build safeguards to "[e]nsure the level of access to the model's weights is appropriate," noting that "[o]nce a model's weights are made widely available, options to roll back or prevent its further sharing and modification are severely limited" and that allowing users to fine-tune models "can reduce the availability of safeguards."[32] Open weights and the ability for users to fine-tune are unique and beneficial qualities of open-source models. These are exactly the sorts of marginal benefits the NTIA framework balances when assessing the risk of open-source models and considering mitigation strategies.[33] In contrast, NIST AI 800-1 recommends against openness altogether.

Rather than follow NTIA's lead by undertaking marginal risk and benefit analysis for open-source models, NIST AI 800-1 recommends practices and safeguards that directly undermine open-source AI development. The only time NIST AI 800-1 mentions weighing benefits is in a footnote on the last page: "Limiting access to the model weights should be weighed against the potential benefits of access, such as for innovation and research, including research into

---

[27] *Id.*
[28] *Id.* at 13.
[29] *Id.* at 12.
[30] *See id.* at 19.
[31] *Id.*
[32] *Id.* at 19.
[33] *See* NAT'L TELECOMM. & INFO. ADMIN., DUAL-USE FOUNDATION MODELS WITH WIDELY AVAILABLE MODEL WEIGHTS 10 (2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

safety."[34] The entire document obfuscates and undermines the very reasons developers open-source in the first place.

CONCLUSION

The Department of Commerce issued the NTIA report and NIST guidance at the direction of the Executive Order on safe AI. The documents, however, take irreconcilable positions on open-source models. While NTIA "outlines a cautious yet optimistic path" for open-source AI,[35] NIST recommends against quintessential features of openness like sharing weights and fine-tuning.[36]

When releasing guidance on AI, NIST should craft recommendations that are consistent with NTIA and overarching Department of Commerce policies on AI. Here, NIST should release separate guidance documents for open and closed-source models, *e.g.* NIST AI 800-1A-OPEN and NIST AI 800-1B-CLOSED. In the open-source guidance, NIST should explain and apply NTIA's marginal risk and benefit analysis framework. At the least, NIST should reissue the current guidance with amendments clarifying which practices apply to open source versus closed, using the NTIA framework to explain the particular mitigations and safeguards recommended for open models.

Respectfully submitted,

_____/s/_____

Andy Jung
Associate Counsel
TechFreedom
ajung@techfreedom.org
1500 K Street NW
Floor 2
Washington, DC 20005

Date: September 9, 2024

---

[34] NIST AI 800-1 at 22 n.48 ("When a model's weights are available to a threat actor, a range of other safeguards become less effective, particularly those that are implemented at the application level, such as limiting who can use the model and detecting when it is misused. Limiting access to the model weights should be weighed against the potential benefits of access, such as for innovation and research, including research into safety. Limiting access to model weights is also only effective if organizations can prevent model theft.").

[35] *See* NAT'L TELECOMM. & INFO. ADMIN., DUAL-USE FOUNDATION MODELS WITH WIDELY AVAILABLE MODEL WEIGHTS 4 (2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

[36] NIST AI 800-1 at 19.