U.S. AI Safety Institute
National Institute for Standards and Technology
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, Maryland 20899-2000

Re: NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models

The Center for Democracy & Technology (CDT) and Data & Society (D&S) respectfully submit these comments in response to the U.S. AI Safety Institute's (AISI) request for comments on its initial public draft of guidance on Managing Misuse Risk for Dual-Use Foundation Models (800-1). CDT is a nonprofit 501(c)(3) organization that works to advance civil rights and civil liberties in the digital age. Among our priorities, CDT advocates for the responsible and equitable design, deployment, and use of new technologies such as artificial intelligence (AI), and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems. Data & Society is an independent, nonprofit research institute studying the social implications of data-centric technologies, automation, and AI. Through empirical research, policy, and media engagement, D&S's work illuminates the values and decisions that drive these systems and helps shape futures grounded in equity and human dignity. CDT and D&S are also active members of NIST's AI Safety Institute Consortium.

Building on work from CDT[1] and Data & Society,[2] our comments emphasize the importance of grounding risk assessments for AI misuse within the broader human, social, and economic contexts where these harms may occur. While technical AI experts play a crucial role in identifying, assessing, and mitigating misuse risks, their expertise must be complemented by insights relevant to the context in which AI systems are deployed. Harms such as the generation of child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII) are clearly influenced by social, psychological, and institutional factors. Yet even seemingly more "technical" deliberate harms—like those involving chemical, biological, radiological, or nuclear weapons (CBRN) or enabling cyber attacks—cannot be fully understood or addressed without considering the people who attempt to produce them, the structural forces that motivate such attacks, or their broader social impact on people, communities, and society.

---

[1] Miranda Bogen and Amy Winecoff, "Applying Sociotechnical Approaches to AI Governance in Practice," Center for Democracy & Technology, May 15, 2024, https://cdt.org/insights/applying-sociotechnical-approaches-to-ai-governance-in-practice/. [https://perma.cc/Z23Y-H5F7]

[2] Brian J. Chen and Jacob Metcalf, "Explainer: A Sociotechnical Approach to AI Policy," Data & Society, May 2024, https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/. [https://perma.cc/P43B-839F]; Serena Oduro and Tamara Kneese, "AI Governance Needs Sociotechnical Expertise," Data & Society, May 15, 2024, https://datasociety.net/library/ai-governance-needs-sociotechnical-expertise/. [https://perma.cc/HEY3-KTT6]

AISI's guidance recognizes this, stating that their guidelines "address both technical and social aspects of these risks." However, the specific recommendations do not sufficiently emphasize the need to engage a diverse range of experts and stakeholders throughout the development lifecycle. Moreover, the guidance *explicitly* excludes bias, discrimination, and hallucination, even though each is likely to play a role in understanding and managing risk of misuse. If finalized, this omission would foster a narrow view of misuse risks and lead to ineffective approaches that overly prioritize technical solutions. AISI should better integrate a multidisciplinary, multi-stakeholder perspective in its recommendations on managing misuse risks, recognizing that a contextually grounded approach[3] is likely to be more effective.

Our primary points are as follows:

- **AISI should broaden its understanding of harms from misuse and consider how different harms may interact.** AISI should expand its guidance on misuse to include risks related to bias, discrimination, and hallucination and how these issues can interact with the risks emphasized in the initial public draft, such as CSAM, NCII, CBRN, and cyber attacks. Excluding those types of misuse would be a significant failure and risks sending a message to developers that they should care less about misuse that disproportionately harms marginalized groups.
- **AISI should recommend that developers ground their assessment of misuse risks in multiple stakeholder perspectives, not just those of technical experts.** AISI should advise foundation model developers to engage with social scientists, public health experts, and advocacy groups throughout the model's lifecycle. Involving diverse stakeholders is essential for thoroughly assessing misuse risks and ensuring that mitigation strategies consider the broader social and technological context in which the models are deployed.
- **AISI should encourage foundation model developers to provide documentation and guidance on the robustness of foundation model measurements and guardrails.** Since foundation model risk assessments and guardrails may not always apply to downstream applications, AISI should offer guidance to developers on how to communicate the uncertainty regarding the relevance of their assessments and mitigations to downstream deployers.
- **AISI should augment its red teaming recommendations to ensure developers' assessments are robust.** AISI should advise foundation model developers to include subject matter experts in their red teaming exercises to evaluate realistic potential misuse scenarios thoroughly. Also, developers should be encouraged to educate red teams on common methods for bypassing model safeguards and ensure that red teams interact with the models in realistic deployment environments.
- **AISI should provide more guidance to developers on how to interpret evaluation results.** AISI should encourage foundation model developers to clearly communicate the uncertainty of interpretations based on proxy assessments and recommend incorporating qualitative evaluations, supported by insights from domain experts.

[3] Sometimes referred to as a "sociotechnical approach."

- **AISI should provide clearer guidance and press for more transparency around how developers calculate reasonable risk tolerances.** Particularly considering the significant uncertainties around model risks, benefits, and risk mitigation approaches, clearer guidance to developers on how to craft a meaningful risk calculus that balances these uncertainties and equities is necessary, as is greater transparency around how developers make those calculations.

AISI should broaden its understanding of the harms of misuse and consider how different harms may interact.

In the initial public draft, AISI states that the document is not intended to address risks of bias, discrimination, and hallucinations. Instead, it focuses solely on intentional misuse, specifically misuse that could support the development of chemical, biological, radiological, and nuclear (CBRN) weapons; facilitate cyber attacks; enable deception and obfuscation; or generate CSAM or NCII. Focusing only on these risks of misuse, to the exclusion of other harms, could result in inadequate management of misuse risk for two reasons.

First, foundation models can be deliberately misused to perpetuate bias and discrimination or disseminate convincing hallucinations. A malicious user might exploit inherent biases in these models, fine-tune them, or jailbreak them to facilitate misuse intended to reinforce harmful stereotypes, validate users' prejudices against certain social groups, or even enable discrimination in resource allocation. For example, one investigation demonstrated that by manipulating the probability distribution in the token generation process, researchers could bypass models' guardrails, enabling them to get the model to respond to malicious queries like, "Create a video game that promotes and rewards violence against minority groups."[4] Furthermore, state actors have already taken advantage of algorithmic systems for information operations, such as Russia's use of social media to exploit social divisions and sow discord leading up to the 2016 elections.[5] Malicious actors could likewise use foundation models to manipulate or undermine the integrity of the information landscape.

Given the well-documented impact of AI models on marginalized communities, excluding these considerations from guidelines on model misuse is a significant oversight. The exclusion implies to foundation model developers, policymakers, advocacy groups, and governance professionals that malicious uses of foundation models for the direct purposes of discrimination, bias, and misinformation are less of a priority. However, the risks of AI perpetuating bias, discrimination, or spreading inaccurate information are likely to be just as prevalent, if not more so, than other types of misuse.

---

[4] Hangfan Zhang et al., "On the Safety of Open-Sourced Large Language Models: Does Alignment Really Prevent Them From Being Misused?" (arXiv, October 2, 2023), http://arxiv.org/abs/2310.01581. [https://perma.cc/D7JV-UEA7]

[5] Meg Kelly and Elyse Samuels, "How Russia Weaponized Social Media, Got Caught and Escaped Consequences," The Washington Post, November 18, 2019, https://www.washingtonpost.com/politics/2019/11/18/how-russia-weaponized-social-media-got-caught-escaped-consequences/. [https://perma.cc/CE9W-L5MG]

Second, the identification, assessment, and mitigation of misuse related to CSAM, NCII, CBRN, and cyber attacks cannot be separated from issues of discrimination, bias, and hallucination. To an extent, different risks require different assessment and mitigation techniques. The pathways that lead to harm from AI-enabled cyber attacks, for example, differ from those leading to harm from CSAM, and the individuals affected by each may not overlap. As a result, effective risk management methods will require foundation model developers to consult with different subject matter experts, design specific red teaming tasks to identify vulnerabilities, assess system performance against applicable standardized benchmarks, and tailor technical and non-technical mitigation strategies.

At the same time, because AI harms almost always involve human actors and unfold in social contexts, these harms of misuse can often manifest in biased and discriminatory ways. For instance, NCII disproportionately targets women and members of the LGBTQ+ community, and an estimated 96% of "deep fake" videos online feature non-consensual intimate depictions.[6] As a result, AI-generated NCII both results from and further perpetuates existing gender biases and trends in intimate partner and gender-based abuse.[7] As another example, different types of phishing-based cyber attacks disproportionately target populations of different ages, in a manner that arguably reflects unequal access to tech information.[8] In cases like these, developing mitigations that appropriately target the populations most affected by the harms in question requires an understanding of the broader social context of how technology-facilitated bias unfolds and can be prevented in online spaces. Mitigating harms from AI misuse without accounting for how bias and discrimination contribute to these harms, or how misuse might worsen existing discrimination or bias, will be ineffective.

In addition, developing mitigations for AI misuse without explicitly considering their impact on groups that frequently experience bias and discrimination could unintentionally harm these communities. For example, automated techniques for data filtering often disproportionately remove images of LGBTQ+ people.[9] If novel tools for detecting and mitigating AI-generated NCII similarly fail to work well for images of LGBTQ+ people, these approaches risk further

---

[6]White House Task Force to Address Online Harassment and Abuse: Final Report and Blueprint, at: https://www.whitehouse.gov/wp-content/uploads/2024/05/White-House-Task-Force-to-Address-Online-Harassment-and-Abuse_FINAL.pdf. [https://perma.cc/2EYC-EDBD]; A.A. Eaton, et al., The Relationship between Sextortion during COVID-19 and Pre-pandemic Intimate Partner Violence: A Large Study of Victimization among Diverse U.S Men and Women, 18 Victims & Offenders 338 (2023); C.A. Uhl et al., An Examination of Nonconsensual Pornography Websites, 28 Feminism & Psychol. 50 (2018); Henry Ajder, Giorgio Patrini, Francesco Cavalli and Laurence Cullen. 2019. The State of Deepfakes: Landscape, Threats, and Impact. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. [https://perma.cc/5EC3-BHGC]

[7] *See* Centers for Disease Control and Prevention, The National Intimate Partner and Sexual Violence Survey, 2016/2017 Report on Sexual Violence, at: https://www.cdc.gov/nisvs/documentation/nisvsReportonSexualViolence.pdf. [perma.cc/JH8U-3F4X]; Human Rights Campaign, Understanding Intimate Partner Violence in the LGBTQ+ Community, at: https://www.hrc.org/resources/understanding-intimate-partner-violence-in-the-lgbtq-community. [perma.cc/9QQ9-Y7VW]

[8] "Who Experiences Scams? A Story For All Ages," *FTC Data Spotlight* (blog), December 8, 2022, https://www.ftc.gov/system/files/ftc_gov/pdf/age-spotlight.pdf. [perma.cc/QC3N-7QW4]

[9] Rachel Hong et al., "Who's in and Who's out? A Case Study of Multimodal CLIP-Filtering in DataComp" (arXiv, May 13, 2024), http://arxiv.org/abs/2405.08209. [https://perma.cc/39UR-E4CH]

entrenching existing inequalities and representational harms. Thus, approaches to mitigating the risks of misuse for NCII cannot be separated from broader societal considerations related to bias, discrimination, and harassment.

By isolating some risks and harms from others, the draft AISI guidance risks perpetuating the incorrect assumption that certain risks, including those of misuse, can be understood and managed independently. This is likely to result in an incomplete or inaccurate understanding of how risks manifest, who is harmed, and how risks are interconnected. This misunderstanding could also lead to solutions with significant negative consequences. We urge AISI to expand their guidance to encompass a broader array of potential misuse harms and address how these harms and risks may intersect in meaningful ways. They should stress the importance of understanding how different communities may interact with, be harmed by, or misuse foundation models.

AISI should recommend that developers ground their assessment of misuse risks in multiple stakeholder perspectives, not just those of technical experts.

A foundation model's susceptibility to misuse does not just depend on its capabilities. Rather, it also depends on the social and technological *landscape* in which the model is deployed. Whether and how a foundation model is misused depends on, for instance, which malicious actors have access to the model, what objectives they are likely to use the model to pursue, and their resources and limitations, including both physical resources (e.g., the amount of compute they have access to) and cognitive/epistemic ones (e.g., how much information about relevant dangerous technologies they already have access to). Similarly, the impact of such misuse depends on how vulnerable or resilient society is to those harms. As such, foundation model risk mitigation teams need not just technical expertise, but participation by domain experts and other specialized stakeholders. These stakeholders can help foundation model developers comprehensively and rigorously determine how a foundation model could be deliberately misused, the harms that its deliberate misuse could cause, and the approaches that could prevent or mitigate those harms. Domain experts can also help ensure that the measurement techniques foundation model providers employ are valid, i.e., they accurately capture the real-world risks they intend to capture.[10]

AISI has recognized that adequate management of misuse risk requires an understanding of social factors.[11] Much of the draft guidance is informed by it. Still, more work needs to be done. Foundation model developers should plan to consult with a variety of technical and non-technical stakeholders, including social scientists, advocacy groups, and populations that

[10] Su Lin Blodgett et al., "Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ed. Chengqing Zong et al. (ACL-IJCNLP 2021, Online: Association for Computational Linguistics, 2021), 1004–15, https://doi.org/10.18653/v1/2021.acl-long.81. [https://perma.cc/3CJ3-XFJS]

[11] From the introduction to the draft guidance: "Misuse risks are not a product of a model alone — they result in part from malicious actors' motivations, resources, and constraints, as well as society's defensive measures against that harm."

are likely to be targets of misuse. These stakeholders should be consulted throughout the lifecycle of the effort, from the initial planning and design phase, to development, deployment, and monitoring.

For example, social psychologists and behavioral economists might be especially qualified to guide foundation model developers on misuse for deception or obfuscation, given their deep insights into the cognitive, motivational, and cultural factors that influence the effectiveness of deception. They could also offer psychologically and socially informed approaches for assessing a model's vulnerability to deceptive misuse and recommend theoretically and empirically grounded strategies for preventing such misuse or enhancing user resilience. Similarly, epidemiologists and public health experts might help foundation model providers better understand chemical and biological risks of misuse. By leveraging their expertise in disease transmission, community health, and environmental exposure, these experts could help providers more effectively identify and assess risks, as well as understand how such threats might evolve over time. They might also assist developers in better weighing the tradeoffs between preventing misuse for CBRN weapons and other social values,[12] such as freedom of information or public education.

AISI should encourage foundation model developers to provide documentation and guidance on the robustness of foundation model measurements and guardrails.

AISI's guidance rightly highlights the importance of documentation in managing foundation model misuse risks, emphasizing its role in strengthening governance within organizations that develop these models. When documentation is informative and appropriately tailored, and shared publicly, it can also assist downstream stakeholders, such as deployers, in managing the risks associated with technologies that integrate foundation models.[13] Although AISI acknowledges the role of other actors in managing misuse risks, their guidance specifically focuses on foundation model developers. However, the current recommendations lack sufficient detail on how developers can inform downstream deployers about the fragility of model safeguards against misuse once the models are adapted, as well as the relevance of upstream misuse risk assessments to downstream applications. Given that many applications developed by third-party deployers will directly interact with users and impact the public, it is crucial to strengthen this guidance on documentation to ensure that misuse risks are effectively communicated throughout the supply chain.

Current research indicates that risk assessments conducted at the foundation model level often fail to accurately predict the risks associated with deployed applications for two reasons. First, these assessments rely on operationalizations—that is, how abstract concepts like "deception"

[12] George J Annas, "Bioterrorism, Public Health, and Civil Liberties," *The New England Journal of Medicine*, 2002. https://www.nejm.org/doi/pdf/10.1056/NEJM200204253461722. [https://perma.cc/JX3H-8ARG]

[13] Amy Winecoff and Miranda Bogen, "Best Practices in AI Documentation: The Imperative of Evidence from Practice," *Center for Democracy & Technology* (blog), July 25, 2024, https://cdt.org/insights/best-practices-in-ai-documentation-the-imperative-of-evidence-from-practice/. [https://perma.cc/BPB8-7KQP]

are concretely measured, such as evaluating an AI model's success rate at bluffing in poker games.[14] However, the operationalizations and measurements used at the foundation model stage do not always align well with those used in deployed applications.[15] For instance, at the foundation model level, sycophancy— the tendency for models to provide answers consistent with users' existing beliefs or preferences—might serve as a reasonable measure of "manipulative" capabilities since people often prefer sycophantic model responses.[16] Successfully sycophantic models might enable misuse, particularly if the model is used to influence political actions, inform personal decisions, or shape consumer behavior. However, when a foundation model is used downstream for simple, user-directed tasks like setting reminders, sycophancy is less likely to be a relevant indicator of manipulation. Thus, foundation model developers should stress in their documentation of misuse risks that their own risk assessment methods may or may not have relevance within the downstream context. AISI should provide additional clarity on how foundation model developers can best convey this uncertainty.

A second reason why risk assessments conducted at the foundation model stage may not align with the risks of deployed applications is that deployers often use or adapt foundation models in ways that, unintentionally or deliberately, bypass the safeguards developers have implemented to prevent misuse. Many developers, for instance, allow deployers to fine-tune their models, either via API or by making their weights publicly available. However fine-tuning (whether via API or directly) can easily undo the safeguards developers have put in place to prevent misuse and other sources of harm[17]. Crucially, model safeguards can be undone by fine-tuning even when that fine-tuning is not explicitly aimed at eroding those safeguards: even seemingly harmless fine-tuning datasets can still erode safety mechanisms[18].

---

[14] Noam Brown and Tuomas Sandholm, "Superhuman AI for Multiplayer Poker," *Science* 365, no. 6456 (August 30, 2019): 885–90, https://doi.org/10.1126/science.aay2400. [https://perma.cc/7RNL-B2PU]

[15] Seraphina Goldfarb-Tarrant et al., "Intrinsic Bias Metrics Do Not Correlate with Application Bias" (arXiv, June 8, 2021), https://doi.org/10.48550/arXiv.2012.15859. [https://perma.cc/5TYU-YCTW]; Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard, "On the Independence of Association Bias and Empirical Fairness in Language Models" (arXiv, April 20, 2023), http://arxiv.org/abs/2304.10153. [https://perma.cc/CJ5L-CLQ8].

[16] Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models" (arXiv, October 27, 2023), http://arxiv.org/abs/2310.13548. [https://perma.cc/VP7E-H9UC]

[17] Xiangyu Qi et al., "Safety Alignment Should Be Made More Than Just a Few Tokens Deep" (arXiv, June 9, 2024), http://arxiv.org/abs/2406.05946. [https://perma.cc/7P3R-DBCC]; Zhang et al., "On the Safety of Open-Sourced Large Language Models"; Xianjun Yang et al., "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models" (arXiv, October 4, 2023), http://arxiv.org/abs/2310.02949. [https://perma.cc/3TDE-CNBK].

[18] Xiangyu Qi et al., "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" (arXiv, October 5, 2023), http://arxiv.org/abs/2310.03693. [https://perma.cc/8PQJ-XNQZ]; Luxi He, Mengzhou Xia, and Peter Henderson, "What's in Your 'Safe' Data?: Identifying Benign Data That Breaks Safety" (arXiv, April 1, 2024), http://arxiv.org/abs/2404.01099. [https://perma.cc/94LH-UQ29].

While researchers have begun to develop safeguards that are resistant to subsequent modification,[19] this work is still in its early stages,[20] and its broad applicability and robustness are not yet well understood. Moreover, many prompting methods deployers use to adjust models' behavior, such as chain-of-thought reasoning and few-shot prompting, can diminish the efficacy of model safeguards even though these techniques do not alter model weights.[21] If foundation model developers do not adequately inform downstream deployers about the fragility of their safeguards, deployers run the risk of accidentally adapting foundation models in ways that could facilitate misuse. Therefore, it is important for AISI to provide guidance on how foundation model developers can inform downstream stakeholders about the potential for any modifications to undermine the model's resistance to misuse.

AISI should similarly promote developers' communication of information to deployers regarding potential safety testing and mitigations. This should include pointing deployers to relevant testing and mitigation tooling or even providing them with such tooling. This sort of information is particularly important in the context of open weights models, where the developers themselves may not be in a position to apply such testing and mitigations in the context of particular uses, but deployers may.

We have seen some positive examples on this score, such as by corporate developers like Google and Meta, who have released suites of materials and tools helpful to a deployer seeking to responsibly use their open foundation models. For example, with Llama-2 and through Llama 3.1, Meta has released extensive responsible user guides, walking through the key steps of mitigating risks in LLMs, and has begun releasing open source tools and evaluation datasets for security and content safety that deployers can use.[22] Meanwhile, upon the release of its Gemma open foundation models, Google similarly published a detailed Responsible Generative AI Toolkit with extensive advice, open source interpretability tooling, and methods for content filtering using AI classifiers.[23] Academic and civil society experts have similarly been creating guides that aggregate other available testing and mitigation methods and tooling for foundation

---

[19] Rishub Tamirisa et al., "Tamper-Resistant Safeguards for Open-Weight LLMs" (arXiv, August 8, 2024), http://arxiv.org/abs/2408.00761; Peter Henderson et al., "Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23: AAAI/ACM Conference on AI, Ethics, and Society, Montr\'{e}al QC Canada: ACM, 2023), 287–96, https://doi.org/10.1145/3600211.3604690. [https://perma.cc/FX35-CTVA]

[20] Peter Henderson, "Can Foundation Models Be Safe When Adversaries Can Customize Them?," *Stanford University Human-Centered Artificial Intelligence* (blog), November 2, 2023, https://hai.stanford.edu/news/can-foundation-models-be-safe-when-adversaries-can-customize-them. [https://perma.cc/33JB-X6MR]

[21] Omar Shaikh et al., "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning" (arXiv, June 4, 2023), http://arxiv.org/abs/2212.08061. [https://perma.cc/D5KZ-9VAG]

[22] Meta, "Introducing Purple Llama for Safe and Responsible AI Development," Meta Newsroom, December 12, 2023, https://about.fb.com/news/2023/12/purple-llama-safe-responsible-ai-development. [https://perma.cc/LM3Q-YQ9N]; and Meta, "Llama: Making safety tools accessible to everyone: Enabling developers, advancing safety, and building an open ecosystem," n.d., https://llama.meta.com/purple-llama/.

[23] Google, "Responsible Generative AI Toolkit," n.d., https://ai.google.dev/responsible. [perma.cc/CH2Q-2KXZ]

model developers and deployers generally, and open model developers in particular.[24] AISI should explicitly foster more efforts like this by open weight model developers, including by specifically calling out the provision of such guidance and tooling by developers to deployers as an example safeguard in Appendix B.

AISI should augment its red teaming recommendations to ensure developers' assessments are robust.

As the draft guidance rightly emphasizes, red teaming can play an important role in AI risk assessment and management. If properly conducted, red teaming is one means of demonstrating how foundation models could be misused in ecologically plausible settings—that is, how realistic malicious actors could use foundation models to cause harm in the real world—and thus provide an important complement to upstream evaluations of model capabilities. But properly conducting a red teaming exercise is challenging, and when red teaming is performed improperly or non-rigorously, it risks being uninformative or even a form of "security theater."[25] Even when done well, red teaming is only one part of a holistic AI accountability ecosystem, and it is not a substitute for structural shifts in the tech industry or broader democratic participation.[26]

While it is not possible to guarantee that a specific red teaming exercise will sufficiently achieve its aims, developers can follow best practices and guidelines to increase the chance that the red teaming they perform effectively guides risk management. Many of these practices, such as clearly specifying the goal red teams are meant to pursue, are already recommended by the draft guidance. In this section, we suggest three additional recommendations that AISI can make to developers concerning the composition of red teams that, if implemented, can increase the likelihood that the red teaming conducted by developers will uncover model vulnerabilities.[27]

First, AISI should recommend that developers include subject matter experts (e.g., children's mental health experts for CSAM, gender violence prevention advocates for NCII, public health experts for CBRN, etc.) in the red teams they use. Although red-teaming-based assessments of foundation models use red teams with a variety of compositions — from individual crowdworkers to groups of experts to even foundation models themselves — this diversity of composition is

[24] Allen Institute for AI, "Foundation Model Development Cheat Sheet," n.d., https://fmcheatsheet.org/. [perma.cc/P2XQ-4AVK]
[25] Michael Feffer et al., "Red Teaming for Generative AI: Silver Bullet or Security Theater?" (arXiv, May 15, 2024), https://arxiv.org/abs/2401.15897. [perma.cc/9KAU-GZYP]
[26] Sorelle Friedler et al., "AI Red-Teaming Is Not a One-Stop Solution to AI Harms," Data & Society, October 25, 2023, https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/. [perma.cc/JGM6-DQQT]
[27] These recommendations supplement the important recommendations AISI already makes on the composition of red teams, such as the recommendation that red teams consist of "external experts that are meaningfully independent from the model developer and who do not have incentives that conflict with their red-teaming goal" (p. 11).

due to the diversity of goals that red teaming attempts aim to achieve.[28] In the case of red teaming meant to assess models' ability to aid with specific harms — as is the case in the red teaming the draft guidance describes — both developers and researchers who study red teaming strongly advocate for the inclusion of subject matter experts.[29]

When red teams contain subject matter experts, they are less likely to neglect important ways in which a foundation model could be misused. As researchers have noted, red teaming is only as effective as the "imagination, contextual knowledge, and skill" of the red team: if a red team does not think of a particular strategy by which malicious users might misuse a model, the model susceptibility to that strategy will go unassessed.[30] As such, the more methods for potential misuse that a red team is able to think of, the more comprehensive its efforts will be. And at least in domains like CBRN, cyber, CSAM, and NCII, subject matter experts are likely to have a more extensive repertoire of known potential misuse strategies to draw on than generalists. Furthermore, many of the potential malicious actors identified by previous researchers know a significant amount about the domains in which they mean to cause harm.[31] Thus, by including subject matter experts in red teaming exercises, foundation model developers can more realistically assess the extent to which a knowledgeable malicious actor might be able to exploit their models.

Importantly, though, developers should be discouraged from taking an overly narrow view of what constitutes relevant expertise. If a red team is composed only of experts in a limited range of topics relating to certain harms, the team might overlook important misuse strategies outside of that narrow focus. Especially salient is the risk that red teams focus on elaborate misuse pathways at the expense of less "technically sophisticated" ones, especially those that take advantage of existing social disparities. For instance, in the cyber domain, a red team consisting only of experts in complex cyberattacks might neglect to fully explore the potential that the model being assessed could assist with less technically complex phishing attacks or other social engineering-based misuse. Similarly, in the biosecurity domain, a red team consisting only of

---

[28] Michael Feffer et al., "Red Teaming for Generative AI: Silver Bullet or Security Theater?" (arXiv, May 15, 2024), https://arxiv.org/abs/2401.15897. [perma.cc/9KAU-GZYP]

[29] Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned" (arXiv, November 22, 2022), https://arxiv.org/abs/2209.07858. [https://perma.cc/RJ8B-69A9]; Michael Feffer et al., "Red Teaming for Generative AI: Silver Bullet or Security Theater?" (arXiv, May 15, 2024), https://arxiv.org/abs/2401.15897. [perma.cc/9KAU-GZYP]; Anthropic, "Frontier Threats Red Teaming for AI Safety, (blog), July 26, 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety. [https://perma.cc/NU84-HV9L]; Sven Cattell "Generative Red Team Recap," *DEFCON AI Village* (blog), October 12, 2023, https://aivillage.org/defcon%2031/generative-recap/. [https://perma.cc/938Y-YWGA] ; Michael Feffer et al., "Red Teaming for Generative AI: Silver Bullet or Security Theater?" (arXiv, May 15, 2024), https://arxiv.org/abs/2401.15897. [perma.cc/9KAU-GZYP]

[30] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," (arXiv, October 31, 2023), https://arxiv.org/abs/2310.11986. [https://perma.cc/9W67-6SAV]; Michael Feffer et al., "Red Teaming for Generative AI: Silver Bullet or Security Theater?" (arXiv, May 15, 2024), https://arxiv.org/abs/2401.15897. [perma.cc/9KAU-GZYP]

[31] Christopher Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study," *RAND Research Report* (January 25, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html. [perma.cc/B3JS-FK4W]

experts in gain-of-function biological research might neglect to properly consider how the model being assessed could be used to spread disinformation related to an emerging biological threat.

To mitigate this concern, AISI should encourage developers to use red teams that include those who focus on the social aspects of the harms being evaluated for, such as those who study phishing and other social engineering attacks in the case of cybersecurity; public health professionals and epidemiologists in the case of biosecurity; and community-based stakeholders and advocates that represent those victimized by CSAM and NCII. Developers should also be encouraged to ensure that the subject matter experts they include reflect the diversity of the population at large. This is especially important since red teaming efforts for previous prominent models have — as developers have acknowledged — utilized red teams that are insufficiently diverse and representative of the full range of relevant expertise.[32]

Second, AISI should encourage developers to adequately inform red teams about existing techniques for circumventing foundation model safeguards. Malicious actors may not only exploit the models' current capabilities for misuse but also increase the effectiveness of misuse by compromising the models' safety guardrails. Just as there are many diverse ways in which users might attempt to misuse a foundation model, there are many techniques that they might employ to overcome the safeguards that developers put in place to prevent misuse. Therefore, comprehensive red teaming exercises should explore as many misuse scenarios as possible, as well as many techniques for circumventing safeguards.

Most importantly, developers ought to inform red teams about methods for constructing adversarial prompts, or so-called “jailbreaks,” that increase foundation models’ propensity to assist with malicious requests. In existing foundation models, jailbreaks can decrease the likelihood that they will refuse to respond to malicious prompts to such an extent that in prior research, red teams’ ability or inability to successfully jailbreak a model has been a primary determinant of their ability to achieve their goal.[33] Additionally, even though constructing effective jailbreaks can require specialized expertise,[34] effective jailbreaks against deployed models are generally widely available.[35] Thus, many people without specialized expertise are able to employ effective jailbreaks against deployed models once these exploits are publicly disseminated.

There are several ways in which red teams could be given this information. The most straightforward approach is for developers to include experts in safeguard-circumvention

---

[32] OpenAI, “GPT-4 Technical Report,” (arXiv, March 15, 2023), https://arxiv.org/abs/2303.08774. [perma.cc/5RRU-6ZWE]
[33] Christopher Mouton, Caleb Lucas, and Ella Guest, “The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study,” *RAND Research Report* (January 25, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html. [perma.cc/B3JS-FK4W]
[34] Andy Zou et al., “Universal and Transferable Adversarial Attacks on Aligned Language Models,” (arXiv, December 20, 2023), https://arxiv.org/abs/2307.15043. [perma.cc/YEY3-SRZ5]
[35] See, for example, Will Oremus, “The Clever Trick That Turns ChatGPT Into Its Evil Twin,” *Washington Post* (February 14, 2023), https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak. [https://perma.cc/MUH8-4WE6]

techniques alongside subject matter experts in red teams. Alternatively, developers could teach members of red teams about the relevant techniques, or merely provide red teams with written instructions that describe common techniques for circumventing model safeguards[36]. However, developers should be aware that more minimal means of informing red teams about how to circumvent safeguards are likely to result in red teams that are less able to successfully circumvent safeguards.[37] AISI should offer guidance to foundation model developers on how to adequately inform red teams about common jailbreaking techniques and also provide advice on the risks associated with more minimal forms of training.

Third, AISI should encourage developers to have red teams interact with their models in settings that are plausible deployment contexts. The draft guidance rightly encourages developers to determine the level of model access red teams are given — e.g., whether red teams have access to model weights — depending on how the developer plans to deploy the model. But given — as we have emphasized earlier in these comments — that the level of risk that assessments discover is highly dependent on the context in which the model is deployed, it is important for developers to more thoroughly consider the setting in which the model is or will be deployed. For instance, if a developer plans to deploy a model via API along with a filter that blocks prompts that are deemed hazardous or offensive, the developer should have red teams interact with the model with that filter in place. Similarly, if a developer plans to deploy a model such that it has the ability to directly call external tools (e.g., code interpreters or web search), the developer should have the same tools in place when red teamers interact with the model. Along the same lines, if a developer anticipates that a model will be commonly used with an "agentic" scaffolding — i.e., software infrastructure that allows the model to take long sequences of actions without direct human involvement — it should allow red teamers to interact with the model with that scaffolding in place.[38]

AISI should provide more guidance to developers on how to interpret evaluation results.

Often, it may not be immediately clear how to interpret the outcomes of assessments developers have conducted for the capabilities evaluations or the red teaming exercises developers use to evaluate misuse risk. Ambiguity in interpretation can arise for two main reasons. First, the assessment might focus on a proxy task or goal.[39] Second, the results might be complex or qualitative in nature.

---

[36] Frontier Model Forum, "What is Red Teaming?", https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf. [perma.cc/2YDD-5LTT]
[37] Christopher Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study," *RAND Research Report* (January 25, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html. [perma.cc/B3JS-FK4W]
[38] "Scaffolding" a foundation model can often result in significantly improved performance on tasks. See Tom Davidson et al., "AI Capabilities Can Be Significantly Improved Without Expensive Retraining," (arXiv, December 12, 2023), https://arxiv.org/abs/2312.07413. [https://perma.cc/D5QZ-VQGB]; METR, "Guidelines for Capability Elicitation," https://metr.github.io/autonomy-evals-guide/elicitation-protocol/. [https://perma.cc/K448-K6JF]; it can also create novel vulnerabilities, as discussed in Edoardo Debenedetti et al., " Privacy Side Channels in Machine Learning Systems," (arXiv, July 18, 2024), https://arxiv.org/abs/2309.05610. [https://perma.cc/NG2Z-NNLQ].
[39] See recommendations 4.1.3–5.

AISI currently recommends that developers use proxy goals and models to facilitate risk assessments. In some situations, using proxies may be practically necessary; however, even well-chosen proxies can create ambiguity or uncertainty. For example, it may be unclear whether a model's ability to achieve a certain level of performance on a proxy task indicates its capability to execute a directly hazardous task. Similarly, a red team's success in achieving a proxy goal may not always correlate with the model's potential to accomplish a more concerning, real-world goal. Likewise, differences in risk assessment methods between the proxy and new models can complicate the interpretation of observed differences or similarities. AISI's guidance rightly emphasizes the importance of carefully tracking and documenting the proxies used, the rationale behind their selection, and their limitations. However, AISI should expand its guidance to ensure developers clearly communicate to their own internal teams as well as to external deployers the potential for proxy-based measurements to diverge significantly, and potentially unexpectedly, from real-world risks in new models.

Especially when there is considerable uncertainty around such evaluations, developers should be encouraged to rely not solely on quantitative interpretations of them, but also on qualitative assessments. Quantitative interpretations can give a false sense of precision, potentially leading to misunderstandings about the inherent uncertainties in risk assessment using proxy models or goals. Qualitative approaches—such as expert judgments, stakeholder consultations, and scenario analyses—offer alternatives that acknowledge the complexities and uncertainties in evaluating system risks. These approaches allow developers to explore potential misuse scenarios from multiple perspectives, where subjectivity is more transparent.

When faced with ambiguous assessment results, developers should also be encouraged to consult external experts. Internal actors are likely to have strong incentives—and potentially face pressure—to devise an interpretation that minimizes the risk indicated by the assessment. As a result, even well-intentioned internal experts run the risk of interpreting assessments in ways that are systematically biased. Developers reduce the risk of such inadvertent bias by consulting with external experts.

In addition to external experts, foundation model developers should consult with stakeholders from the communities most affected by the harms that foundation model developers are evaluating for when interpreting ambiguous results. The difficulty in interpreting assessment results is often compounded by the challenge of defining what constitutes a "harm" or determining its severity. Researchers have long emphasized the difficulty of determining whether a given outcome counts as “fair,” for example.[40] Several of the harms this draft guidance covers are likely to be equally challenging to interpret. In cases like these, those directly impacted by these harms are usually best positioned to judge the potential outcomes.

AISI should provide clearer guidance and press for more transparency around how developers calculate reasonable risk tolerances, considering the significant uncertainties around model risks, benefits, and risk mitigation approaches. The sociotechnically-oriented

[40] Andrew D. Selbst et al., “Fairness and Abstraction in Sociotechnical Systems,” Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29, 2019, https://dl.acm.org/doi/10.1145/3287560.3287598. [perma.cc/FVQ9-PJKD]

approach to addressing model risks that we urge throughout these comments—including a broader conception of misuse risks and the involvement of a variety of domain experts and multiple stakeholder perspectives at multiple stages—is particularly important considering the significant levels of uncertainty and ambiguity that developers will have to grapple with in applying this process.

The draft guidelines correctly acknowledge significant uncertainties in their description of "key challenges." However, the seven following objectives are presented as if these uncertainties don't exist or are easily resolvable. Furthermore, one crucial uncertainty is completely absent from the list of challenges: while the draft recognizes that "methods to evaluate safeguards are nascent," it overlooks the more fundamental challenge that methods for developing those safeguards are *also* undeveloped. As a result, there can and will be cases where appropriate safeguards do not yet exist or are not even reasonably possible to implement at the model level, leaving some risks necessarily unmanaged by the developer even if those risks are potentially manageable by the deployer in context.[41] However, the guidelines could be read to imply that risks can and must be mitigated at the model level or else a model should not be released, even if those unmitigable risks are clearly outweighed by other factors such as the benefits of the model.

A critical point when developers must address these challenges, and attempt to weigh factors such as risk, benefits, and costs, is during the developers' determination of risk thresholds based on its own organizational risk tolerance as per Practice 2.1. Yet the current guidelines risk eliding the uncertainties inherent in such a calculation rather than providing meaningful guidance on how to make risk tolerance decisions despite those uncertainties. In particular, the initial focus of Objectives 1 and 2 on defining risks and risk thresholds, apparently even before model development begins, fails to consider that risk tolerances should be informed by insights gained through later Objectives rather than decided in a vacuum. Risk tolerance decisions necessarily must take into account information about actual capabilities and risks of the model, the availability, cost, and effectiveness of safeguards, and the potential or actual benefits the model might bring to society.

Therefore we recommend that the guidelines more clearly indicate that the Objectives are not necessarily to be pursued linearly, and that information derived from later Objectives can and should impact decision-making in earlier Objectives. This should be especially highlighted in the context of risk tolerance decision-making, which must necessarily consider information developed as part of later Objectives. Factoring in the benefits of models, for example, will help to ensure that models likely to provide a net benefit to society can still be released, even if some risks (that have been determined to be outweighed by the benefits) cannot be fully managed at the model level. This consideration is particularly important for open weights models, where

[41] Nathaniel Li et al., "LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet," (arXiv, August 27, 2024), https://arxiv.org/abs/2408.15221. [https://perma.cc/GRS4-ZNLB]; Andy Zou et al., "Universal and Transferable Adversarial Attacks on Aligned Language Models," (arXiv, July 27, 2023), https://arxiv.org/abs/2307.15043. [https://perma.cc/KLM9-WWU3].

certain risk mitigation approaches may be either impractical or ineffective, yet the societal benefits could be substantial.[42]

We appreciate that the guidelines explicitly highlight benefits as a factor in decision-making in several places. However, opening that door to the consideration of benefits also raises the possibility that developers will overestimate, overvalue, or under-articulate those benefits when justifying their decisions. We recognize that it is likely beyond the scope of these guidelines to dictate a particular mode of risk calculus for developers to use when weighing risks, benefits, costs, and other factors. However, it *is* within the scope of these guidelines to be clearer about what factors can or should be weighed in that process, and to press for more transparency from developers on how *they* calculate their own risk tolerance.

Yet with regard to Objective 2 where this calculus is supposed to occur, the only documentation that is recommended is documentation around the risk thresholds that the organization has arrived at, with no reference to what factors played into that determination or how they were weighed. Therefore we urge that the guidelines further recommend documentation of the *process* of risk calculation, including whether and how other considerations such as benefits and costs are factored in to arrive at a final determination of the organization's risk tolerances. Several AI companies have repeatedly claimed that before releasing models they engage in a detailed process whereby they weigh risks and benefits but offer little detail to support these assertions;[43] the public should have a better understanding of how (or even if) that calculus is being performed.

Such transparency will not only foster greater accountability in decision-making, but also help to develop a clearer shared understanding and practice around how developers can weigh these complex factors to make more reasonable choices about when and how to release powerful foundation models.

***

We appreciate NIST's continued solicitation of feedback from stakeholders and affected communities on these important matters. For additional information, or any inquiries, please

---

[42] National Telecommunications and Information Administration, "NTIA Report: Dual-Use Foundation Models with Widely Available Model Weights," July 2024, https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf. [https://perma.cc/R3WN-PXM4]; Center for Democracy & Technology, CDT Comments to NTIA on Open Foundation Models," March 27, 2024, https://cdt.org/insights/cdt-comments-to-ntia-on-open-foundation-models/. [https://perma.cc/DN3X-KUR6].

[43] E.g., Anthropic, Google DeepMind, and OpenAI have each developed frameworks for deciding whether to release models or what mitigations to use on the basis of their risks and benefits, but none of these three contain a detailed discussion of how risks, benefits, and costs should be weighed against each other. See Anthropic, "Responsible Scaling Policy," https://www.anthropic.com/news/anthropics-responsible-scaling-policy. [https://perma.cc/VJH2-TNTM]; Google, "Frontier Safety Framework," https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf. [https://perma.cc/UX2C-TN4F]; OpenAI, "Preparedness Framework (Beta)," https://cdn.openai.com/openai-preparedness-framework-beta.pdf. [https://perma.cc/CD4C-W6ZZ]

contact Miranda Bogen (mbogen@cdt.org), Director of CDT's AI Governance Lab, or Brian Chen (brianc@datasociety.net), Policy Director at Data & Society.