epic.org

Electronic Privacy Information Center
1519 New Hampshire Avenue NW
Washington, DC 20036, USA

+1 202 483 1140
+1 202 483 1248
@EPICPrivacy
https://epic.org

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER

to the

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

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

No. 2024-17614

September 9, 2024

INTRODUCTION

The Electronic Privacy Information Center (EPIC) submits these comments in response to the National Institute of Standards and Technology's (NIST's) Request for Comments on the U.S. Artificial Intelligence Safety Institute's Draft Document on Managing Misuse Risk for Dual-Use Foundation Models (NIST AI 800-1).[1]

EPIC is a public interest research center in Washington, D.C., established in 1994 to secure the fundamental right to privacy in the digital age for all people through advocacy, research, and litigation.[2] We advocate for a human-rights-based approach to AI policy that ensures new technologies are subject to democratic governance.[3] Over the last decade, EPIC has consistently advocated for the adoption of clear, commonsense, and actionable AI regulations across the country.[4] EPIC has also published extensive research on emerging AI technologies like generative

---

[1] 89 Fed. Reg. 64,878 (Aug. 8, 2024).
[2] *About Us*, EPIC, https://epic.org/about/ (2023).
[3] *See, e.g.*, *AI and Human Rights*, EPIC, https://epic.org/issues/ai/ (2023); *AI and Human Rights: Criminal Legal System*, EPIC, https://epic.org/issues/ai/ai-in-the-criminal-justice-system/ (2023); EPIC, Outsourced & Automated: How AI Companies Have Taken Over Government Decision-Making (2023), https://epic.org/outsourced-automated/ [hereinafter "Outsourced & Automated Report"]; Letter from EPIC to President Biden and Vice President Harris on Ensuring Adequate Federal Workforce and Resources for Effective AI Oversight (Oct. 24, 2023), https://epic.org/wp-content/uploads/2023/10/EPIC-letter-to-White-House-re-AI-workforce-and-resources-Oct-2023.pdf; EPIC, Comments on the NIST Artificial Intelligence Risk Management Framework: Second Draft (Sept. 28, 2022), https://epic.org/wp-content/uploads/2022/09/EPIC-Comments-NIST-RMF-09-28-22.pdf.
[4] *See, e.g.*, Press Release, EPIC, EPIC Urges DC Council to Pass Algorithmic Discrimination Bill (Sept. 23, 2022), https://epic.org/epic-urges-dc-council-to-pass-algorithmic-discrimination-bill/; EPIC, Comments to the

AI,[5] as well as the ways that government agencies develop, procure, and use AI systems around the country.[6] EPIC is a member of NIST's U.S. Artificial Intelligence Safety Institute Consortium (AISIC).

As the U.S. Artificial Intelligence Safety Institute (AISI) considers updates to Draft Document NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models, EPIC reemphasizes our call for NIST and its affiliated entities to implement actionable AI risk mitigation strategies with strong incentive structures and accountability mechanisms—steps that will ensure that AI developers and deployers faithfully adopt the practices and implementation recommendations within NIST AI 800-1.[7] At the same time, EPIC encourages AISI to view the misuse risks of generative AI technologies, including dual-use and multimodal foundation models, as extensions of traditional AI and automated decision-making risks, rather than qualitatively different risks requiring a curated set of objectives and practices. A holistic approach to AI risk management—including risk identification, measurement, mitigation, and transparency through processes like regular risk assessments and red-teaming exercises—is not only a more efficient risk management paradigm, but also a more effective one at mitigating misuse risks. For example, malicious actors can and have exploited implicit biases within foundation models to reconstruct model parameters and training data samples.[8] Additionally, preserving consumer privacy within foundation models is a fundamental for managing the misuse risks of foundation models—not something that AI developers should deprioritize while attempting to safeguard against the misuse of foundation models.[9]

Patent and Trademark Office on Intellectual Property Protection for Artificial Intelligence Innovation (Jan. 10, 2020), https://epic.org/wp-content/uploads/apa/comments/EPIC-USPTO-Jan2020.pdf; EPIC, Comments on the Department of Housing and Urban Development's Implementation of the Fair Housing Act's Disparate Impact Standard (Oct. 18, 2019), https://epic.org/wp-content/uploads/apa/comments/EPIC-HUD-Oct2019.pdf.

[5] EPIC, Generating Harms: Generative AI's Impact & Paths Forward (2023), https://epic.org/gai [hereinafter "EPIC Generative AI Report I"]; EPIC, Generating Harms II: Generative AI's New & Continued Impacts (2024), https://epic.org/wp-content/uploads/2024/05/EPIC-Generative-AI-II-Report-May2024-1.pdf [hereinafter "EPIC Generative AI Report II"].

[6] Outsourced & Automated Report; EPIC, Screened & Scored in the District of Columbia (2022), https://epic.org/wp-content/uploads/2022/11/EPIC-Screened-in-DC-Report.pdf [hereinafter "Screened & Scored Report"].

[7] *See, e.g.*, EPIC, Comments on the NIST Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Feb. 2, 2024), https://epic.org/wp-content/uploads/2024/02/EPIC-Comment-on-NIST-AI-Executive-Order-Mandates-RFI-02.02.24.pdf; EPIC, Comments on the NIST Artificial Intelligence Risk Management Framework: Second Draft (Sept. 28, 2022), https://epic.org/wp-content/uploads/2022/09/EPIC-Comments-NIST-RMF-09-28-22.pdf.

[8] Apostol Vassilev et al., NIST, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2023, at 29–30 (2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf.

[9] *See* NIST AI 800-1 at 19.

I. Sociotechnical AI Risks Are Misuse Risks

Responsive to Questions 1 and 3.

Sociotechnical risks of AI technologies, including bias risks within foundation models, *are* misuse risks. Many of the sociotechnical AI risks that NIST AI 800-1 does not address, including AI biases and hallucinations or confabulations, can directly influence a foundation model's vulnerability to adversarial attacks, model theft, and other cybersecurity threats. Implicit biases can be used to reconstruct key elements of foundation models like parameters and training datasets.[10] Bias injection techniques permit threat actors to deliberately introduce biases into AI models to alter outcomes in ways that can seriously undermine national security, economic security, or public health.[11] For example, a bias injection attack on foundation models used for financial services could alter business and consumer transaction behaviors in ways that destabilize markets. And model bias can hinder misuse risk mitigation directly as well: false positives or negatives in model outputs during testing, evaluation, assessment, or red teaming can undermine an AI developer's ability to identify and manage the misuse risks within a foundation model.[12]

While EPIC appreciates AISI's interest in developing specific guidance for misuse risks and explicit recommendation for AI actors to manage the AI risks of bias, discrimination, and hallucinations "consistent with relevant guidelines," we urge AISI to reconsider its decision to place these risks fully outside the scope of NIST AI 800-1. Because threat actors can and do use model bias and other sociotechnical features to facilitate foundation model misuse, any framework for managing the misuse risks of such AI models will be limited in its effectiveness without any methods for identifying and managing model bias and discrimination.

II. Preserving Consumer Privacy is Crucial to Managing the Misuse Risks of Foundation Models

Responsive to Questions 1, 2, and 4.

EPIC strongly encourages AISI to reconsider its light-touch approach to consumer privacy within NIST AI 800-1, including it suggestion that some safeguards against foundation model misuse may still be viable despite reducing user privacy.[13] Data privacy and security is at the core of responsible foundation model development, with direct impacts on the likelihood of adversarial

---

[10] *See, e.g.*, Vassilev et al., *supra* note 8.
[11] *Id.*; *see also* Xavier Ferrer et al., *Bias and Discrimination AI: A Cross-Disciplinary Perspective*, 40(2) IEEE Tech. & Soc'y Mag. 72, 72–80 (2021), https://ieeexplore.ieee.org/document/9445793.
[12] *Cf.* Vassilev et al., *supra* note 8, at 12; Gauthama Raman M.R. et al., *Machine Learning for Intrusion Detection in Industrial Control Systems: Challenges and Lessons from Experimental Evaluation*, 4(27) Cybersecurity, 2021, at 6, 9, https://cybersecurity.springeropen.com/articles/10.1186/s42400-021-00095-5.
[13] NIST AI 800-1 at 19.

attacks and other forms of model misuse and exploitation. For foundation models to operate, they must first be trained and finetuned on data, often by splitting a dataset into training and test sets.[14] While many AI datasets involve non-human data, today's most popular AI applications are built using data collected indiscriminately via web scraping[15] and purchased from data brokers.[16] In fact, several popular AI applications have trained on user content after deployment as well, meaning that sensitive or personal information provided in user emails, prompts, and other content are incorporated into AI systems after initial model development.[17] Because AI training data can incorporate extensive sensitive or personally identifiable information about people, the ways AI developers collect, use, and secure their training data directly impacts the privacy rights of countless people—and the value of misuse to threat actors, either directly or by facilitating further threats to national security, economic security, or public health and safety.

Unlike traditional software systems, however, multimodal foundation models built atop machine-learning methods and extensive commercial data cannot easily correct or delete personal data used to train the system after the fact. Once a model is trained on data, it effectively memorizes that data and cannot easily unlearn it.[18] Every model output will reflect the training data, and some foundation models, like large-language models, may even leak personal data directly to users.[19] Because of the ways that AI models incorporate training data, adversarial machine-learning techniques have been developed to effectively identify and extract sensitive information in training datasets through an AI model's behavior.[20] Two such techniques are membership inference attacks, which aim to determine whether a certain data sample was included in a model's training data by

---

[14] *See, e.g.*, Hyojin Bahng et al., *Learning De-Biased Representations with Biased Representations*, 37 Proc. Int. Conf. on Mach. Learning 1 (2020), https://proceedings.mlr.press/v119/bahng20a/bahng20a.pdf.
[15] *See* Müge Fazlioglu, *Training AI on Personal Data Scraped from the Web*, IAPP (Nov. 8, 2023), https://iapp.org/news/a/training-ai-on-personal-data-scraped-from-the-web/; Thomas Claburn, *How to Spot OpenAI's Crawler Bot and Stop it Slurping Sites for Training Data*, Register (Aug. 8, 2023), https://www.theregister.com/2023/08/08/openai_scraping_software/; Sara Morrison, *The Tricky Truth About How Generative AI Uses Your Data*, Vox (July 27, 2023).
[16] *See, e.g.*, Evan Weinberger, *Data Brokers Eyed by CFPB for Selling Sensitive Info for Ads, AI*, Bloomberg Law (Aug. 15, 2023), https://news.bloomberglaw.com/banking-law/data-brokers-eyed-by-cfpb-for-selling-sensitive-info-for-ads-ai.
[17] *See* Geoffrey A. Fowler, *Your Instagrams are Training AI. There's Little You Can Do About It.*, Wash. Post (Sept. 27, 2023), https://www.washingtonpost.com/technology/2023/09/08/gmail-instagram-facebook-trains-ai/; Kyle Wiggers, *Addressing Criticism, OpenAI Will No Longer Use Customer Data to Train its Models by Default*, TechCrunch (Mar. 1, 2023), https://techcrunch.com/2023/03/01/addressing-criticism-openai-will-no-longer-use-customer-data-to-train-its-models-by-default/.
[18] *See* Liwei Song & Prateek Mittal, *Systematic Evaluation of Privacy Risks of Machine Learning Models*, 30 Proc. USENIX Sec. Symp. 2615, 2615 (2021). Hurdles to unlearning data are at the core of recent FTC cases requiring AI model deletion. *See* Jevan Hutson & Ben Winters, *America's Next 'Stop Model!': Model Deletion*, 8 Geo. L. Tech. Rev. 125, 128–134 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4225003.
[19] Tiernan Ray, *ChatGPT can Leak Training Data, Violate Privacy, Says Google's DeepMind*, ZDNet (Dec. 4, 2023), https://www.zdnet.com/article/chatgpt-can-leak-source-data-violate-privacy-says-googles-deepmind/.
[20] *See* Song & Mittal, *supra* note 18, at 2629; Michale Backes et al.,

evaluating model outputs,[21] and attribute inference attacks, which aim to impute sensitive training data attributes using partial knowledge of nonsensitive training data attributes and model outputs.[22] Crucially, many adversarial machine-learning techniques have proven effective at identifying sensitive or personally identifiable information in training datasets in *both* closed and open models.[23]

To fortify data privacy in AI models against vulnerabilities and adversarial attacks, some AI researchers have developed differential privacy techniques for AI models.[24] Differential privacy is one of several burgeoning data privacy-preserving mathematical techniques, wherein random noise is injected into different elements of a technical system to prevent the identification of any one individual's data without significantly impacting the accuracy of the system's outputs.[25] As applied to machine-learning models, differentially private noise can be applied to at least four model elements: (1) a model's training data; (2) a model's loss function, which evaluates how well a trained model predicts an expected outcome; (3) a model's gradients, which are commonly used to optimize model training by adjusting model weights to minimize errors; and (4) a model's weights.[26] However, adding noise to any single element of an AI model will not be effective at preserving privacy: adding noise to training data may combat attribute inference attacks but is less effective against membership inference attacks, the inverse is true for adding noise to the loss function or gradients, and adding noise to model weights themselves may resist both membership and attribute inference attacks at the cost of significantly reducing model accuracy.[27] Because of these limitations, data privacy and security will remain a core concern for responsible foundation model development and management, with the specific privacy advantages and disadvantages of different AI models shifting as developers move along the gradient of AI openness.[28]

---

[21] *See, e.g.*, Reza Shokri et al., *Membership Inference Attacks Against Machine Learning Models*, 2017 IEEE Sump. On Sec. & Priv. 3, https://ieeexplore.ieee.org/document/7958568.
[22] *See, e.g.*, Bargav Jayaraman & David Evans, *Are Attribute Inference Attacks Just Imputation?*, arXiv (Sept. 2, 2022), https://arxiv.org/pdf/2209.01292.pdf.
[23] *See* Song & Mittal, *supra* note 18, at 2615 (overview of research studies into membership inference attacks); Reza Shokri et al., *supra* note 21, at 3–18 (closed AI research); *see generally* Milad Nasr et al., *Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks Against Centralized and Federating Learning*, 2019 IEEE Symp. On Sec. & Priv. (open AI research).
[24] *See* Tianqing Zhu et al., *More than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence*, 344 IEEE Transactions on Knowledge & Data Eng'g 2824, 2830–36 (June 2022), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9158374.
[25] Ctr. for Info. Pol'y Leadership, Hunton Andrews Kurth LLP, Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age 36–41 (2023), https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-understanding-pets-and-ppts-dec2023.pdf.
[26] Zhu et al., *supra* note 24, at 2830.
[27] *Id.*
[28] *See* Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, arXiv (Feb. 5, 2023), https://arxiv.org/pdf/2302.04844.pdf.

To address the direct impact that data privacy vulnerabilities have on the risk of foundation model misuse, EPIC recommends incorporating specific references to and recommendations about privacy and data vulnerabilities across all seven objectives within NIST AI 800-1. These references could include, *inter alia*:

1. **Practice 1.1**: Highlighting data breaches and bias injection attacks as known misuse risks within threat profiles.

2. **Practice 1.2**: Recommending the inclusion of data privacy considerations within threat profile impact assessments.

3. **Practice 2.2**: Expanding discussion of any security practices proposed as part of an AI developer's roadmap to manage misuse risks to include foundation model data security practices and the implementation of privacy-enhancing technologies.

4. **Objective 3**: Expanding discussion of model theft to also include risks of data breaches, bias injection, or prompt injection. Ultimately, these risks share the core harms of model theft: a threat actor may gain access to information about a model or its training process such that the threat actor can endanger national security, financial security, or public health and safety.

5. **Practice 4.2**: Ensuring that red teams understand and test for data privacy vulnerabilities within dual-use foundation models.

CONCLUSION

EPIC welcomes NIST's and AISI's efforts to spur on responsible foundation model development in NIST AI 800-1 and similar guidance documents. The risks of foundation model misuse are real and significant, and EPIC supports a more responsible and transparent approach to AI model development and oversight. As AISI finalizes NIST AI 800-1, however, EPIC encourages the Institute to reconsider its dismissal and deprioritization of bias risks and privacy risks, respectively. Without adequate privacy and bias safeguards in place, AI developers cannot effectively identify, assess, and mitigate misuse risks within dual-use foundation models. Specifically, EPIC recommends that NIST and AISI:

1. Incorporate bias and discrimination risks within NIST AI 800-1, as threat actors can and do exploit model biases as part of foundation model misuse;

2. Reprioritize consumer privacy as a core factor to consider when managing the risks of foundation model misuse, including by incorporating explicit recommendations to identify, monitor, and resolve data vulnerabilities within foundation models as part of

an AI developer's general efforts to manage misuse risks; and

3. Detail the relative benefits and limitations of both model openness and privacy-enhancing technologies as features of an AI developer's misuse risk management efforts.

We appreciate this opportunity to reply to NIST's Request for Comments and are willing to engage with NIST further on any of the issues raised within our comment, either directly or through the AISIC. EPIC's recommendations align closely to the goals of Executive Order 14110 and the NIST AI RMF to increase the safety, equity, and reliability of AI technologies, and we strongly believe that incorporating more sociotechnical factors into NIST AI 800-1, such as data privacy vulnerabilities and bias risks, will only improve the effectiveness of techniques to manage the misuse risks of dual-use foundation models.

Respectfully submitted,

/s/ Grant Fergusson
Grant Fergusson
EPIC Counsel

/s/ Calli Schroeder
Calli Schroeder
EPIC Senior Counsel
Global Privacy Counsel

ELECTRONIC PRIVACY
INFORMATION CENTER (EPIC)
1519 New Hampshire Ave. NW
Washington, DC 20036
202-483-1140 (tel)
202-483-1248 (fax)