MERCATUS CENTER
George Mason University

PUBLIC INTEREST COMMENT

Comment on NIST/AISI AI 800-1: Expanding Guidelines for Open-Source AI

DEAN W. BALL
Research Fellow

NIST/AISI AI 800-1 Draft
Agency: National Institute of Standards of Technology
Comment Period Opens: July 26, 2024
Comment Period Closes: September 9, 2024
Comment Submitted: September 9, 2024
Docket No. NIST-2024-0002
XRIN: 0693-XC137

I am a research fellow in the Artificial Intelligence and Progress Project at the Mercatus Center at George Mason University. This comment represents my own views, but does not represent the views of the Mercatus Center, which does not take institutional positions on matters of public policy. Thank you for taking the time to consider my feedback to the joint US Artificial Intelligence Safety Institute (AISI)/National Institute of Safety Institute (NIST) AI 800-1 draft, "Managing Misuse Risk for Dual-Use Foundation Models." I will keep my thoughts brief.

I was disappointed to see open-source AI treated as something of an inconvenient afterthought in the AI 800-1 draft. While the draft does not explicitly disfavor open-source AI, it does, in my view, make a series of recommendations that would be impossible for an open-source AI developer to comply with. Often, it is as though the authors of the guidelines would prefer that the open-source distribution mechanism did not exist.

For example, Objective 3, "Manage the risk of model theft," implicitly assumes that the weights of a given model are a closely guarded secret. Yet public release of model weights is the key characteristic of any open-source AI model. The draft thereby suggests that releasing model weights publicly is a failure of this objective.

In addition, of the 14 individual "example safeguards" listed in Appendix B, nine are fully incompatible with open-source AI and one seems to suggest that open-source models not be used at all.[1] (The remaining four apply only to pre-deployment training steps and are thus possible for any developer to conduct.)

It is undoubtedly the case that open-source AI carries hypothetical security tradeoffs. It is worth noting, however, that powerful open-source foundation models have been available for the entirety of

[1] The example safeguard that seems to caution against open sourcing models: "Consider when it is appropriate to make the model's weights widely available, such as available for download by the public. Once a model's weights are made widely available, options to roll back or prevent its further sharing and modification are severely limited."

For more information, contact
Mercatus Outreach, 703-993-4930, mercatusoutreach@mercatus.gmu.edu
Mercatus Center at George Mason University
3434 Washington Blvd., 4th Floor, Arlington, VA 22201

the AISI's existence as an institution, and not a single major security incident can be attributed to open-source AI as a "but for" cause. Furthermore, there are numerous benefits open-source AI can offer to safety and interpretability research. Indeed, on the day that these draft guidelines were released, NIST also released a model evaluation tool that *only* works on open-source models, underscoring the value that open-source AI has in practice to those doing AI safety and security research.

In a document that is supposed to set guidelines for this nascent field, it is disheartening to see open-source AI so consistently ignored and implicitly disfavored. I hope that future iterations of these guidelines include more specific guidelines for open-source AI developers, such as the use of emerging methods to prevent model safeguards from being easily unwound through fine-tuning.[2] Open-source AI is likely here to stay, and I believe that this is largely for the best. Acknowledging that reality is likely to lead to a document that is helpful to a broader range of industry and academic stakeholders.

[2] E.g., Rishub Tamirisa et al., "Tamper-Resistant Safeguards for Open-Weight LLMs," arXiv, 2024, https://doi.org/10.48550/arXiv.2408.00761.