U.S. Department of Homeland Security
Cybersecurity & Infrastructure Security Agency
Office of the Director
Washington, DC 20528

June 18, 2024

MEMORANDUM

FROM: Jen Easterly
Director

SUBJECT: **CISA Response to NTIA Request for Information on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights**

CISA appreciates the opportunity to respond to the National Telecommunications and Information Administration's (NTIA) Request for Information on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights. While definitions are evolving, we use the term "open foundation models" to refer to foundation models with widely available weights.[1]

CISA, as the operational lead for federal cybersecurity and the national coordinator for critical infrastructure security and resilience, is working to maximize the benefit of and minimize risks to artificial intelligence (AI) models, including open foundation models.[2] In November 2023, CISA published its Roadmap for AI, which describes five lines of effort focused on addressing and managing AI risks. These efforts include but are not limited to, responsibly using AI to support our mission, assuring AI systems, protecting critical infrastructure from malicious use of AI, collaborating and communicating on AI efforts, and expanding AI expertise in our workforce.

CISA's position is that we should continue to promote the responsible development and release of open foundation models while mitigating their potential harms. While AI capabilities introduce new threats into the landscape, leading academics have proposed that the marginal risk from open as opposed to closed source models is low.[3] This is not to imply the newly introduced threats are minimal. At CISA, we see value in open foundation models to help strengthen cybersecurity, increase competition, and promote innovation. At the same time, all foundation models, including open source ones, present real safety concerns that we must work to mitigate. As such, CISA recommends that we (1) work to assess and address harms, and (2) learn from existing software security work.

---

[1] DHS defines a "foundation model" as one that that is trained on a broad set of general domain data for the purpose of using that model as an architecture on which to build multiple specialized AI applications. See Foundation Models at the Department of Homeland Security: Use Cases and Considerations and Bommasani et al., On the Opportunities and Risks of Foundation Models

[2] *Id.*

[3] See, for example, Issue Brief Considerations for Governing Open Foundation Models | Stanford HAI.

1. Work to assess and address potential harms of open foundation models.

Foundation models have at least two classes of potential harms. The first is intentionally leveraged by the deployer or user of the model (e.g., use of the foundational model to aid in gaining unauthorized access to data or systems or to develop and spread non-consensual intimate imagery (NCII)). The second class involves impacts that are undesired by those deploying the models (e.g., a cybersecurity vulnerability in a model deployed by a critical infrastructure entity).

For the first class of harms, we encourage a multipronged risk reduction approach. While additional research and investment to limit abuse of the technology is essential, we should expect that many protections will inevitably be rolled back by malicious deployers, especially in open foundation models. Therefore, non-technological risk mitigations are also needed (e.g., enforcing platform terms of service, mitigating the risk of NCII by requiring resources be provided to support victims of such abuse, supporting strong communities that can provide organic social support for everyone involved, and enforcing relevant laws or passing new ones where necessary).

Similarly, reducing the risk of a cyber incident from adversaries leveraging foundation models is best mitigated by organizations' implementation of existing cybersecurity best practices and software developers' implementation of Secure by Design principles. The latter ensures that the burden of cybersecurity does not fall on the targeted organization alone. Additionally, organizations should ensure the systems they acquire and use are built with security in mind. AI model developers should also build in anti-abuse measures, decline to build models for explicitly malicious purposes, and report incidents and negative impacts where they observe models are being used for explicitly malicious purposes. Platforms should avoid hosting such explicitly malicious models.

Creators and deployers of open foundation models can take steps to mitigate the second class of harms by using a "safe by design" approach and building in protections to their model. This may address cybersecurity vulnerabilities or other forms of harms such as biases. Responsibly-developed open foundation models are likely to be less susceptible to harms and misuse, on the whole, than models that cannot be publicly audited.

2. Learn from existing software security work.

Developers of all AI models, including open foundation models, can learn from existing work to secure software, including CISA's lines of effort around Secure by Design and open source software security.

In particular, we highlight valuable parallels to CISA's and the U.S. government's approach to open source software security. CISA's Secure by Design program and the National Cybersecurity Strategy emphasize that it is the software manufacturers who integrate open source software into their products that are responsible for security, not the individual open source developers.

Deployers have a responsibility to evaluate and responsibly deploy the models they depend on, and sustainably contribute to open model improvements. Much as software manufacturers should responsibly select open source dependencies and sustainably support their ongoing maintenance, open foundation model deployers should ensure that models they use have appropriate security and safety precautions included and enabled by default.

Lastly, a secure by design approach can help protect software systems from all forms of attack, including AI-assisted attacks. We urge technology manufacturers to review CISA's Secure by Design guidance and integrate product security from the start. This ensures that upon launch products will be "secure out of the box." Thank you for the consideration of our response. We look forward to further collaboration.