IST Institute for SECURITY + TECHNOLOGY

IST Leadership

Mike McNerney
Chair, Board of Directors

Philip Reiner
Chief Executive Officer

Megan Stifel
Chief Strategy Officer

Steve Kelly
Chief Trust Officer

Institute for Security and Technology
PO Box 11045
Oakland, CA 64611

January 30, 2024

Ms. Alicia Chambers
Executive Secretariat
U.S. Department of Commerce
National Institute of Standards and Technology
Gaithersburg, MD 20899

Subject: Comments in response to NIST's request for information (RFI) on §§ 4.1, 4.5 and 11 of Executive Order 14110 on *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*; document citation 88 RF 88368; agency docket number 231218-0309.

Dear Ms. Chambers,

The Institute for Security and Technology (IST) appreciates the opportunity to file comments in response to NIST's request for information regarding its assignments under **Executive Order 14110 § 4.1 on "Developing Guidelines, Standards, and Best Practices for AI Safety and Security."** IST submits for consideration elements of its 13 December 2023 report entitled, "*How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access*."[1] The study process leading to the report involved participation from a working group of stakeholders from leading AI labs, industry, academia, and civil society.

Consistent with the AI Risk Management Framework's (RMF's) "map" core function, the report identified six specific risks which we believe will be helpful to NIST's work going forward (*see*, pp. 13-15 of the report). These are:

- Fueling a race to the bottom,
- Malicious use,
- Capability overhang,
- Compliance failure,
- Taking the human out of the loop, and
- Reinforcing bias.

[1] https://securityandtechnology.org/ai-foundation-model-access-initiative/how-does-access-impact-risk/

The report then defines and describes a gradient of access to AI foundation models comprised of six levels (*see*, pp. 16-21 of the report), as follows:

- Fully closed,
- Paper publication,
- Query API access,
- Modular API access,
- Gated downloadable access,
- Non-gated downloadable access, and
- Fully open access.

Consistent with the RMF's "measure" core function, the report describes the severity of risk for each risk category at each level of access, culminating in a novel risk matrix (*see*, pp. 24-36 of the report). In short, the report concludes that the risk profile changes (typically increasing) as access to foundation models increases.

In support of the RMF's "manage" core function, IST wishes to call attention to a concept developed by the working group and documented in the report with regard to the source of risk as being "upstream" (inherent to the model itself) or "downstream" (driven by user/third party interaction with the model). These terms are described in pages 37-38 of the report, then used regularly thereafter. In the most general terms, preventing harm from upstream risks will likely require interventions in model development, with standards and guidelines targeting AI labs and other developer organizations. Preventing harm from downstream risks, however, will require standards and guidelines on a use-case basis, targeting user interaction with models. While the former category will likely require novel interventions tailored to new AI models and developers, existing frameworks and guidelines targeting industries or use cases (including existing cybersecurity best practices and dual use standards) may prove to be a useful foundation on which to manage effective downstream risks.

With respect to the RMF's "govern" core function, we recommend that AI governance strategies take into account not only the models themselves, but also system components upon which they rely and the varying ways in which access is granted to models at a technical level. The various combinations of these factors might necessitate tailored interventions to prevent harm. For example, if a developer organization grants access to a model via an API, they maintain control of the model and its components. Conversely, if a developer organization makes a model available for

download along with the model's weights, it is not possible to walk back the model release, and such a release might therefore require more intensive testing and evaluation to ensure the risk of harm is mitigated. In short, while all models should be subject to rigorous testing, red-teaming, and evaluation, as access to a given model increases, so too should the burden on the developer organization to ensure the technology they release will not result in harm.

We hope these comments are helpful, particularly our considerations regarding generative AI's implications for the core functions of the RMF. Thank you for considering our comments.

Regards,

Steve Kelly
Chief Trust Officer