IST Institute for SECURITY + TECHNOLOGY

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

March 27, 2024

Mr. Bertram Lee
U.S. Department of Commerce
National Telecommunications and Information Administration
1401 Constitution Ave NW
Washington, DC 20230

**Subject: Openness in AI Request for Comment; document number NTIA–2023–0009.**

Dear Mr. Lee,

The Institute for Security and Technology (IST) appreciates the opportunity to file comments in response to **NTIA's request for information (RFI) on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights.** 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.

While dual use foundation models with widely available weights (referred to here as open foundation models) could play a key role in fostering growth among less resourced actors, increasing access to these models also presents risks. IST's December report identified six specific risks which we believe will be helpful to NTIA's assessment (*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
- Reinforcing bias

The report also identifies the *source* of each risk as upstream (inherent to a model and therefore a byproduct of the developer organization's design,

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

training, or tuning choices), downstream (driven by a user's interaction with a model at a given level of access), or both (*see*, pp. 23 of the report).

Regarding your question, **"[h]ow should NTIA define 'open' or 'widely available' when thinking about foundation models and model weights?**" our report departs from the open/closed binary, and instead 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
- Fully open access

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). Widely available model weights are one tool in a suite of many that allow for greater access and impact the ability of less resourced actors to engage with dual use foundation models. While IST's report assesses access across a range of technical components and interface mechanisms, it concludes that the risk profile changes (typically increasing) as access to foundation models increases.

Further, regarding your question, **"[h]ow do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?**" IST wishes to call attention to the impact of widely available model weights on downstream risks, including malicious use, compliance failure, and taking the human out of the loop.

If a developer organization grants access to a model via an API, for example, 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. Such a release might therefore require more intensive testing and evaluation to ensure the risk of harm, especially resulting from malicious use, compliance failure, and/or taking the human out of the loop, 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 recommend that AI governance strategies take into account not only model weights, 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.

The working group that informed IST's December report has convened to develop a set of recommended technical and policy mitigations for the identified risks and looks forward to publishing a companion report this spring. Thank you for considering our comments.

Regards,

Steve Kelly
Chief Trust Officer