600 14th St. NW, Suite 300
Washington, D.C. 20005

March 27, 2024

Department of Commerce
National Telecommunications and Information Administration
1401 Constitution Ave, NW
Washington, DC 20230

**Subject: NTIA–2023–0009 - Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights**

Dear Administrator Davidson,

On behalf of International Business Machines Corporation (IBM), we welcome the opportunity to respond to the National Telecommunications and Information Administration's (NTIA) request for comment on dual use foundation artificial intelligence models with widely available model weights. The topic is a critical one, and NTIA's recommendations will have an impact on the U.S.'s continued leadership in this emerging technology. As the agency evaluates these challenging questions, we recommend that NTIA:

- Ensure that any regulations designed to mitigate AI safety risks posed by widely available model weights follow the science, rather than pre-empt it;
- Work with NIST and other stakeholders to co-develop and test precise regulatory interventions that mitigate specific AI safety risks; and
- Preserve and prioritize the critical benefits of open innovation ecosystems for AI for increasing AI safety, advancing national competitiveness, and promoting democratization and transparency of this technology.

IBM commends NTIA for its thorough approach to investigating these challenging questions around the potential risks and benefits of dual-use foundation models with widely available model weights. IBM strongly believes that while not every AI model can or will be "open," protecting an open innovation ecosystem for AI in which model weights are widely available should be policymakers' highest priority for advancing U.S. leadership in AI. Additionally, IBM recommends NTIA acknowledge the lack of robust scientific evidence that would be sufficient to justify any restrictions on an open innovation ecosystem for AI.

IBM has been a global leader in investigating and advancing the potential of AI since the term "artificial intelligence" was coined at the seminal conference at Dartmouth University which IBM cosponsored in 1956.[1] We look forward to sharing our expertise with policymakers to maximize the benefits AI will offer, so long as it is deployed in safe and trustworthy manner.

Respectfully,

Darío Gil
Senior Vice President and Director, IBM Research

[1] https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

IBM urges policymakers and all AI stakeholders to embrace a degree of epistemic humility when evaluating the potential risks of dual-use foundation models with widely evaluable model weights. Policymakers are right to want to minimize the risk of harm that AI could pose. However, there is a glaring absence of evidence suggesting that many AI safety risks – particularly existential risks – are imminent or even likely, or that widely available model weights could increase them. For example, a recent study led by researchers at Stanford University and Princeton University found little to no evidence supporting the argument that widely available model weights posed any marginal increase in societal risk across key categories, such as biosecurity and disinformation.[2]

This is not to say that safety risks of dual-use foundation models, including those potentially exacerbated by the widely available model weights, will not or cannot emerge as this technology matures. But policymakers should recognize three critical factors that justify a more prudent approach than taking early action to restrict access to model weights:

- The science of evaluating AI models for potential safety risks is extremely immature. The U.S., United Kingdom, and Japan have taken the laudable step to launch AI Safety Institutes to specifically develop "science-based and empirically backed guidelines and standards for AI measurement and policy, laying the foundation for AI safety across the world."[3] **Until this science is more mature and such guidelines actually exist, decisions about how to address potential risks will be necessarily under-informed.**
- **Open access to model weights will be a key driver of mitigations to AI safety risks.** Undermining an open ecosystem around AI, in which a wide and diverse group of stakeholders has the freedom to scrutinize, break down, and improve upon these technologies, would significantly increase the U.S.' risk exposure to AI by sacrificing one of our most effective resources for combatting risk.
- The broad economic and social benefits of openness are overwhelming. **Any restrictions on access to model weights will have significant downstream effects that would erode open ecosystems in the US and globally.** Such restrictions would hurt U.S. national security and economic leadership, enable anticompetitive market dynamics, and create significant barriers to the democratization and transparency of this critical technology.

NTIA also asks for input on potential regulatory models that could increase the benefits and/or decrease the risk of dual use foundation models with widely available model weights. One model worth exploring is partnering with the NIST AI Safety Institute (AISI) to develop a "regulatory sandbox for regulation," rather than a sandbox for technology development. Similar initiatives have proven successful in other countries, such as the IMDA AI Verify Foundation in Singapore, which houses AI Verify – "an AI governance testing framework and software toolkit that validates the performance of AI systems against a set of internationally recognised principles through standardised tests."[4] Efforts such as AI Verify and other regulatory sandboxes are vital tools for promoting iterative developments to regulatory

[2] https://crfm.stanford.edu/open-fms/
[3] https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute
[4] https://aiverifyfoundation.sg/what-is-ai-verify/

frameworks, and offer the significant benefit of being flexible, adaptive, and growth- and evolution-oriented.

Novel regulatory approaches may be warranted at some point to address potential risks of foundation models, but the consequences of an imprecise approach to regulation are too great to risk untested solutions. As NIST works to develop the scientific understanding of identifying and evaluating AI safety risks, an NTIA-led pilot that works with NIST to test and iterate on regulatory responses to novel risks not already addressed by existing regulatory frameworks would be extremely beneficial.

Another regulatory model worth exploring as part of this sandbox approach is creating a toolbox of regulatory solutions that could be leveraged if and when certain milestones in the science of AI safety evaluations are achieved. This "conditional regulation" approach should focus on domains of AI that many believe to be intrinsically higher risk, such as bioengineering.[5] While it is intuitive that an AI model capable of generating novel molecular structures for pharmaceutical research might be able to also design dangerous biological agents, there are no widely-agreed upon, standardized processes for empirically detecting this capability in a model, nor evaluating whether access to such a model could increase marginal societal risk of harm. As NIST's AI Safety Institute develops these processes and evaluations, NTIA could propose relevant regulatory agencies work with the Institute to co-develop and test regulatory interventions precisely targeted to mitigate such risks that would only go into effect after this scientifically-grounded understanding is established.

Just as IBM is working with NIST through the AI Safety Institute Consortium to advance the science of evaluating and mitigating AI safety risks, IBM is eager to work with NTIA and other policymakers and stakeholders to apply that science to risk- and evidence-based regulatory interventions.

**What, if any, are the risks associated with widely available model weights?**

The main risk commonly associated with widely available model weights is that it can enable bad actors to modify or remove built in guardrails intended to prevent misuse. This concern does not acknowledge the fact that future innovations – themselves enabled by access to widely available model weights – could reduce this risk.

It is also important to recognize that these risks do not exist in a vacuum. While a bad actor could theoretically modify an open source model to perform malicious activity, such as generating malware that can steal confidential information, such activities are still criminal offenses, regardless of whether the malware was AI-generated or not. Of course, malicious hacking still occurs even though it is illegal. Nonetheless, legal and regulatory frameworks that impose penalties for harmful applications of a technology can still address both the risks and the harms associated with such activities, regardless of the tools bad actors choose to employ in such criminal enterprises.

[5] https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html#scope-of-the-ai-safety-summit-what-is-frontier-ai

Critically, the NIST AISI is tasked with identifying these risks. While responses to this Request for Comment may help shape and advance this work, NTIA should defer to NIST's determinations when completed to ensure that any potential downstream regulation is truly evidence-based.

**How do these risks compare to those associated with closed models?**

It is naïve to assume that bad actors – state actors or otherwise – will be unable to leverage AI to cause harm just because they do not have access to model weights. Restricting access to model weights by requiring all interaction with an AI model to go through an API, for example, has not proven effective at reducing risk. A harassment campaign devoted to generating and disseminating deepfake pornography of Taylor Swift and other celebrities leveraged consumer-facing image generation tools that processed user requests through an API to generate this imagery.[6] While closed models may sometimes introduce friction into the process of using AI for undesirable purposes, it has not proven to be nearly as durable or robust an approach to preventing AI safety risks as many advocates of restricting access to model weights claim it to be.

Additionally, as Arvind Narayanan and Sayash Kapoor of Princeton University note, "AI safety" is not a property of a model, but a contextually dependent evaluation based on multiple factors.[7] How a model is released, including whether model weights are made widely available, can be a relevant factor in evaluating safety, but there is simply not sufficient evidence to indicate that open release strategies marginally increase risks.[8] Thus, it is difficult to draw any meaningful contrast between simply "open" and "closed" models in terms of safety risk.

This highlights the significant need for additional clarity and nuance in how stakeholders talk about "AI safety" and "AI risk." Just as "safety" is not a model property, many of the risks of AI are not unique to AI, yet are discussed as AI-specific challenges that can be addressed at the technology layer, rather than the application layer, with AI-specific regulatory solutions.[9] We know this is not the case for many of the most prominent risks of AI, such as malicious use of deepfakes – while it is true generative AI can exacerbate these harms, many solutions are straightforward and are technology neutral, centered on holding bad actors liable for the harm they cause.[10] The NIST AI Safety Institute is tasked with addressing many of these questions surrounding taxonomy, and we believe that other agencies and stakeholders can and should play a meaningful role in that process.

**What, if any, risks could result from differences in access to widely available models across different jurisdictions?**

Restrictions on access to model weights, which if imposed will vary across jurisdictions, are unlikely to meaningfully address AI safety risks and could pose significant unintended consequences.

---

[6] https://www.nytimes.com/2024/02/05/business/media/taylor-swift-ai-fake-images.html
[7] https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property
[8] https://crfm.stanford.edu/open-fms/
[9] https://newsroom.ibm.com/Whitepaper-A-Policymakers-Guide-to-Foundation-Models
[10] https://newsroom.ibm.com/Blog-Heres-What-Policymakers-Can-Do-About-Deepfakes

First, model weights are easily portable digital files that can be disseminated undetected with minimal cost on consumer hardware and communication platforms. Thus, a bad actor could easily disseminate model weights across jurisdictions.

Second, restrictions on access to model weights in a particular jurisdiction will only deter good actors from accessing them in that jurisdiction. As described in response to other questions, these "good actors" in open ecosystems are critical drivers of innovation, using AI to deliver greater economic and social benefits and advancing AI safety research. Thus, one of the most likely outcomes of jurisdictional variations in access to model weights is that the most restrictive jurisdictions will also be the most limited in their ability to benefit from AI.

**What benefits do open model weights offer for competition and innovation, both in the AI marketplace and in other areas of the economy? In what ways can open dual-use foundation models enable or enhance scientific research, as well as education/training in computer science and related fields?**

Open innovation ecosystems in which model weights are widely available can offer significant economic and social benefits for the AI ecosystem and beyond.[11]

The most obvious benefit of an open ecosystem is that it lowers the barrier to entry for competition and innovation. By making many of the technical resources necessary to develop and deploy AI more readily available, including model weights, open ecosystems enable small and large firms alike, as well as research institutions, to develop new and competitive products and services without steep, and potentially prohibitive, upfront costs. And there is tremendous demand from businesses for these kinds of resources. Not only does IBM rely on collaborative technologies like open source, we also make them available through IBM's watsonx platform, as our clients value access to a variety of powerful tools and models from the open innovation community like Meta's Llama2-chat 70 billion parameter model alongside our proprietary offerings for specialized enterprise use cases.[12]

Openness and innovation go hand-in-hand, and open ecosystems can drive AI advancements in three key ways.

First, more competition means everyone must raise the bar. Companies should compete based on how well they can tailor and deploy AI in valuable ways, rather than on how effectively they can hoard the resources and enact barriers of entry to develop AI.

Second, more access to AI means more stakeholders can identify opportunities to improve AI technologies and more easily pursue novel and valuable applications for AI. Combined, an open AI ecosystem is dramatically more innovative, inclusive, and competitive than a closed one.

---

[11] https://www.ibm.com/policy/why-we-must-protect-an-open-innovation-ecosystem-for-ai/

[12] https://developer.ibm.com/articles/awb-the-open-source-ecosystem-of-watsonx/; https://newsroom.ibm.com/2023-08-09-IBM-Plans-to-Make-Llama-2-Available-within-its-Watsonx-AI-and-Data-Platform

Third, just as openness drives competition, it also drives democratization**.** Open ecosystems mean more opportunities for anyone to explore, test, modify, study, and deploy AI, which can dramatically lower the bar to deploying AI for socially beneficial applications. This is why IBM partnered with NASA to develop a geospatial foundation model and make it available under an open source license.[13] The model can analyze NASA's geospatial data up to four times faster than state-of-the-art deep-learning models, with half as much labeled data, making it a powerful tool to accelerate climate-related discoveries. Since it is openly available, anyone – including non-governmental organizations, governments, and other stakeholders – can leverage this model freely to drive efforts to build resilience to climate change.

It is much easier to learn about a subject when you can access the materials for free. An open innovation ecosystem unleashes a significantly broader pool of AI talent, as students, academics, and existing members of the workforce can more easily access the resources necessary to acquire AI skills. This is part of the reason IBM contributes hundreds of open models and datasets to Hugging Face, the leading open source collaboration platform for the machine learning community building the future of AI.[14] IBM has also committed to training 2 million people in AI by 2026, with a focus on underrepresented communities, by partnering with universities around the world to provide AI training resources and by making AI coursework available for free on our SkillsBuild platform.[15]

**How can making model weights widely available improve the safety, security, and trustworthiness of AI and the robustness of public preparedness against potential AI risks?**

As with open source software, open models can enable a higher level of scrutiny from the community, greatly increasing the likelihood that vulnerabilities are identified and patched. This means that open AI models can be as trusted, secure, and fit for use in different contexts as proprietary models.

Openness can also drive innovations that can improve safety. For example, Stanford researchers developed a novel technique called "meta-learned adversarial censoring (MLAC)" that can greatly impede efforts to manipulate a model to perform harmful activity, even with access to model weights.[16] These researchers themselves relied on open model weights for a language model to test and evaluate their technique – a feat that would not have been possible were model weights restricted.

In some contexts, AI safety can also depend on the ability for diverse stakeholders to scrutinize and evaluate models to identify any vulnerabilities, identify undesirable behaviors, and ensure they are functioning properly. However, without "deep access," which includes access to model weights, these evaluations will be severely limited in their effectiveness.[17]

At a high level, wide access to model weights means more good actors, including those in governments, industry, civil society, academia, and the general public, can leverage AI to mitigate the potential impacts of harmful applications of AI. Restrictions on access to open model weights will mean sufficiently-

[13] https://research.ibm.com/blog/nasa-hugging-face-ibm
[14] https://huggingface.co/ibm
[15] https://newsroom.ibm.com/2023-09-18-IBM-Commits-to-Train-2-Million-in-Artificial-Intelligence-in-Three-Years,-with-a-Focus-on-Underrepresented-Communities
[16] https://arxiv.org/pdf/2211.14946.pdf
[17] https://bpb-us-e1.wpmucdn.com/sites.mit.edu/dist/6/336/files/2024/03/Safe-Harbor-0e192065dccf6d83.pdf

motivated bad actors will still be able to leverage AI to cause harm, whereas good actors, who respect the rule of law, will not be able to sufficiently build resilience to these harms.

**Are there effective ways to create safeguards around foundation models, either to ensure that model weights do not become available, or to protect system integrity or human well-being (including privacy) and reduce security risks in those cases where weights are widely available?**

There are emerging techniques to create safeguards for foundation models when weights are widely available. In addition to the Stanford University researchers' MLAC technique described earlier, Meta has developed a watermarking technique called Stable Signature that embeds a watermark in generated images.[18] Watermarking is currently an unreliable method for detecting synthetic content as watermarks can typically be easily edited out by end users. Stable Signature applies a watermark, which could indicate the model version and user, at the model level in a way that is not affected by additional fine-tuning or image manipulation.[19]

**What are the prospects for developing effective safeguards in the future?**

Foundation models are still an emerging technology and there will be enormous amounts of innovation over the coming decade. While it is impossible to predict exactly what kinds of technologies will be developed, novel safeguards, new types of algorithm architectures, methods for governance, watermarking and detection techniques, and more will emerge that will have significant positive impact for safety.

Fostering the development of these safeguards should be a top priority for the NIST AISI. And critically, the AISI and the entire AI ecosystem will be significantly more successful in developing these safeguards if model weights are widely available.

**In which ways is open-source software policy analogous (or not) to the availability of model weights? Are there lessons we can learn from the history and ecosystem of open-source software, open data, and other "open" initiatives for open foundation models, particularly the availability of model weights?**

There are many lessons to be learned from past debates about the safety and security of open-source software (OSS) that can add clarity to debates around the perceived safety risks of open foundation models. Of these, the most important is that openness has always been a force for improving safety and security. For decades, increased transparency and access to source code has enabled a broad and diverse stakeholder community to identify and fix vulnerabilities, increase performance and resilience, and raise the bar for security overall.

For example, in 2000, Red Hat and the U.S. National Security Agency (NSA) developed Security-Enhanced Linux (SELinux), a Linux kernel module that can modify Linux distributions to provide robust security features.[20] SELinux is both made available as OSS and is built on top of existing OSS – representative of the feedback loop for improving security that openness can enable.

---

[18] https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/
[19] https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/
[20] https://www.redhat.com/en/topics/linux/what-is-selinux

Present-day arguments that access to model weights increase safety risks share parallels with debunked past arguments that providing broad access to source code could make it easier for bad actors to exploit it. While policymakers are right to be asking the question of whether potentially dangerous misuse of AI will be enabled by access to model weights, they should be careful to acknowledge that similar concerns about OSS have routinely been proven wrong, and that openness was a key driver in mitigating safety and security concerns.

**What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?**

There are unlikely any legal measures that can reliably or sustainably limit access to model weights for highly capable models on a global scale. Two of the most commonly proposed solutions – government-issued licenses to develop or operate a highly-capable model and export controls on model weights -- would both come with extreme consequences for U.S. competitiveness. And as described elsewhere, not only would such efforts do little to advance AI safety in the short term, but they risk undermining it in the longer term by limiting the ability for good actors to contribute to AI safety solutions and to take advantage of AI for beneficial applications.

**What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights? Should other government or non-government bodies, currently existing or not, support the government in this role? Should this vary by sector?**

Standards-setting and scientific agencies, particularly NIST through its AI Safety Institute, have a critically important role to play in developing methods to evaluate AI models, measure risk, and develop safeguards. This work should be conducted in close collaboration with other agencies and stakeholders outside of government. The AI Safety Institute Consortium is a worthwhile model which, while still new, shows promise to ensure that these standards are informed by robust scientific evidence and community consensus.

U.S. agencies should work with their counterparts around the world to share best practices, avoid pursuing redundant lines of research, and facilitate the development and mutual adoption of consensus standards around AI risk. Governments should not create a new international body, as some have proposed, to oversee and steer this work globally. Proposed models such as the International Atomic Energy Agency (IAEA), which promotes responsible use of nuclear technology globally, are simply not useful to adopt for AI, and an international body would quickly become heavily politicized.

**Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 10^26 integer or floating-point operations used in the Executive Order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?**

Compute requirements are not a useful metric for identifying safety risks. First, as described above, AI safety is a context-dependent evaluation of multiple factors, not a distinct property of a model. Second, compute requirements do not necessarily indicate dangerous capabilities. As the field matures, greater levels of performance are being achieved with smaller and smaller models. Any compute threshold

intended to differentiate between low and high capability models will quickly become obsolete for this reason.

**Conclusion**

NTIA should not consider any restrictions on how model weights can be distributed or accessed until it has the necessary scientific evidence to understand if such restrictions would be necessary, effective, and worth the tradeoffs that would come with them. NTIA should acknowledge that it does not yet have this evidence and should work with the NIST AISI and other stakeholders to help develop it. At present, any restrictions on the open innovation ecosystem for AI would significantly hinder U.S. leadership in AI without meaningfully addressing AI safety risks.