Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 706 7329
www.microsoft.com

Microsoft

The Honorable Alan Davidson
National Telecommunications and Information Administration
1401 Constitution Ave NW, Washington, D.C. 20230

Submitted electronically at: www.reginfo.gov/public/do/PRAMain

March 27, 2024

RE: Request for Comment (RFC) on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights (Docket No. NTIA–2023–0009)

Microsoft appreciates the opportunity to respond to the NTIA's request for comment on dual use foundation artificial intelligence models with widely available model weights. This is an important and timely topic as foundation models are becoming more powerful and pervasive, and organizations across the public, private, and non-profit sectors are putting them to work in a wide range of scenarios.

At Microsoft, we are committed to advancing the state of the art in AI and ensuring that our AI products and services are safe, secure, and trustworthy. We believe that foundation models, such as large language models (LLMs), have great potential to enable new capabilities and benefit society broadly. Foundation models with model weights that are widely available can especially drive an open innovation cycle that benefits not only open source projects but also the extensive value chain built on those projects. They can also play a key role in improving AI safety and security, enabling a diverse global community to learn about responsible design, contribute to research, and participate in the development of models used to support safety and security tasks. But, like all technologies, foundation models, including both fully closed models and those with model weights that are widely available, need to be developed and deployed responsibly in order to realize that potential.

In our response, we offer recommendations for how NTIA and other stakeholders can foster an effective and well-calibrated approach to the governance of open foundation models with widely available model weights. At a high level, we recommend:

- Establishing clear and consistent definitions for open source AI models, components of AI models, and distribution methods for AI models. These definitions will allow policymakers to target specific attributes that are introducing risk, such as the breadth of deployment or the ability to fine-tune or otherwise modify a model, increasing the effectiveness of any policy interventions and reducing undue burdens.

- Developing and adopting voluntary risk-based and outcome-focused frameworks and guidance for the responsible release of foundation models and model weights. These frameworks and guidance will set expectations for stakeholders while longer-term international standards are developed.
- Promoting risk and impact assessments that are grounded in the specific attributes of widely available model weights that present risk, the marginal risk of such availability compared to existing systems, and the effectiveness of risk mitigations across end-to-end scenarios of use.
- Cultivating a healthy and responsible open source AI ecosystem and ensuring that policies foster innovation and research. This will be achieved through direct engagement with open source communities to understand the impact of policy interventions on them and, as needed, calibrations to address risks of concern while also minimizing negative impacts on innovation and research.
- Encouraging global research and collaboration on AI fundamentals, AI safety and security, and AI applications. The Federal government can incentivize research on risk evaluations and mitigations in a way that ensures that policies, such as export controls, continue to allow for appropriate global collaboration on research.
- Expanding the technical and organizational capabilities and capacities within the public sector for the evaluation of AI risks and capabilities. The public sector has a unique ability to coordinate and prioritize the evaluation of AI risks and capabilities, but this will require ongoing investment to ensure that emerging risks and capabilities are addressed.

Thank you again for the opportunity to contribute our views. We welcome further dialogue as NTIA develops its report as well as any policy and regulatory recommendations designed to address both risks and opportunities.

Sincerely,

Natasha Crampton

Natasha Crampton
Vice President and Chief Responsible AI Officer
Microsoft Corporation

How should NTIA define "open" or "widely available" when thinking about foundation models and model weights?

Recommendations:

1. *Treat "open source" as a concept distinct from "availability," and target policy recommendations at the appropriate concept for the intended outcomes.*
2. *Avoid "open" as a standalone concept as it can be construed to mean either "open source" or "widely available model weights."*
3. *When well-understood concepts (preferably defined by an international standards body or relevant non-profit foundation) do not exist, refer to specific defining attributes of the concept. For example, a policy could target "models that can be modified or fine-tuned."*

For a policy to be effective, its scope needs to be clearly defined, which requires shared understanding between the author of the policy, the entities to which it applies, and its enforcers. Poorly scoped policies may be inefficient, imposing burdens unnecessarily on low-risk use cases, or ineffective, failing to achieve desired policy outcomes for higher-risk use cases. In the present context, "availability" and "openness," while related, may have vastly different interpretations, risks, and benefits, and the NTIA must tease apart and account for these differences in any policies NTIA chooses to adopt or recommend on this topic.

As described in *Considerations for Governing Open Foundation Models,*[1] the "availability" of an AI model and its components sits on a gradient:

- Widely available weights
  - Weights, data, and code available – without use restrictions
  - Weights, data, and code available – with use restrictions
  - Weights available
- Hosted or API access – accessible via a hosted application or API
- Fully closed – not accessible outside the developer organization

Other variations, beyond this list, are possible. For example, if the data and code is available for a model, but the model developer does not provide pre-computed weights, then the model may still be considered to have widely available weights if typical model consumers (including end users and downstream developers) have sufficient resources and documentation to train the model themselves.

"Openness" is harder to define and depends on the freedoms expected for something to be considered "open." For some, "widely available weights," which would allow fine-tuning of a

[1] https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

model, may be sufficient to consider an AI model "open." For others, unrestricted permission to use, modify, and redistribute the AI model (which requires code, and may require weights and/or data) is needed. The terms under which the weights, and other components, are made available will define its "openness" and suitability for different uses.

Traditional software has confronted similar definitional challenges, and, over time, this has resulted in both formal and informal definitions to help stakeholders communicate effectively, including:

- Open Source Software – software with a license that meets the Open Source Initiative's Open Source Definition[2].
- Free Software – software that meets the Free Software Foundation's Free Software Definition[3].
- Source-Available Software[4] – software for which source code is available, but which has a license that is incompatible with the Free Software and Open Source Definitions.
- Closed-Source Software[5] – software for which source code is not available.

The Open Source Initiative[6] (OSI) has started to address these challenges for AI models by developing the Open Source AI Definition,[7] which is currently in draft. In contrast to "availability," which primarily focuses on distribution, the Open Source AI Definition focuses on the freedoms that the AI model's licensing terms grant to developers to be able to use and improve the model for their use case.

As part of the definition, OSI also identifies the components of an AI model needed to exercise those freedoms. For example, modifying or re-training an AI model requires the training data and code, so those must be provided under an OSI-approved license for an AI model to be considered open source under the definition. This means that "open source" AI models will have minimum levels of "availability," but certain levels of "availability" do not automatically imply that the AI model is "open source" unless the freedoms required by the definition have been satisfied.

When scoping any policies, it is important that the NTIA articulates its goals clearly in order to drive clarity. If NTIA's goal is to address risks based on the potential distribution of a model – for

[2] https://opensource.org/osd
[3] https://www.gnu.org/philosophy/free-sw.html
[4] https://en.wikipedia.org/wiki/Source-available_software
[5] https://en.wikipedia.org/wiki/Proprietary_software
[6] Microsoft is a sponsor of the Open Source Initiative (OSI), a "501(c)3 non-profit raising awareness and adoption of open source software (OSS) through advocacy, education and building bridges between communities."
[7] https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-6

example, preventing proliferation of harmful AI models – then the scoping should be based on "availability." Alternatively, if NTIA's goal is to address models developed and released under terms granting certain freedoms, for example, to promote innovation by excluding those from specific requirements, then scoping should be based on "open source."

How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?

Recommendations:

1. *Promote the responsible release of AI models, including performing risk assessments before release.*
2. *Incentivize the development of, and promote the use of, AI risk management guidance and standards that consider marginal risk and unique risks posed by the distribution method (i.e., "availability").*
3. *National Institute of Standards and Technology (NIST) to develop and Office of Management and Budget (OMB) to adopt guidance on the responsible consumption of third-party AI models (i.e., supply chain management).*

A crucial step in planning the release of any AI model is assessing risk; frameworks, such as NIST's *AI Risk Management Framework*[8] (RMF), or standards, such as *ISO/IEC 42001 – AI Management System*[9] (AIMS), provide implementation-agnostic practices for incorporating risk assessment into the release process. Similarly, model consumers, by which we mean end users and downstream developers, should assess the risks posed by the AI models they are consuming.

In October 2023, the Partnership on AI[10] (PAI) released *Guidance for Safe Foundation Model Deployment,*[11] which provides practical guidance on risk assessment and mitigation for publishers of AI models. The guidance is tailored based on both the AI model's capabilities and availability, which allows it to address the unique risks stemming from each of these dimensions and their intersections.

Each distribution method poses unique risks for both model developers and model consumers, and part of these risk assessments should be to assess those risks in the broader context of the release or use.

[8] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[9] https://www.iso.org/standard/81230.html

[10] Microsoft is a corporate sponsor of the Partnership on AI (PAI), "an independent, nonprofit 501(c)(3) organization and funded by charitable contributions from philanthropy and corporate entities."

[11] https://partnershiponai.org/modeldeployment/

This table shows some of the unique risks to the model developer and model consumers for different distribution methods:

Distribution Method	Risk for Model Developer	Risk for Model Consumers
Weights, data, and code available	• No protection of trade secrets • Unable to monitor for and prevent downstream intentional misuse or unintentional harms	• Assumes responsibility for risks and obligations emerging from the model's use
Weights available	• Limited technological protection of intellectual property • Unable to monitor for and prevent[12] downstream intentional misuse or unintentional harms	• Assumes responsibility for risks and obligations emerging from the model's use • Unable to re-train the model[13]
Hosted or API access	• Shared responsibility[14]	• Shared responsibility[14] • Unable to modify or re-train model, unless explicitly enabled and permitted by the service provider • Reliance on service provider for availability, and possibly other functionality or obligations depending on the service offered[15]
Fully closed – the model developer and model consumer are the same entity and the model is not available to third parties	Assumes responsibility for risks and obligations emerging from the model's use	

Focusing on the specific attributes that contribute to the risk, rather than broad categorizations, enables model developers and model consumers to assess and mitigate (or, in lower risk scenarios, accept) those risks. For example, if the risk is exacerbated by the breadth of distribution, then a model developer may consider phased releases that allow measuring and responding to unexpected risks at a smaller scale and mitigate the risk of being unable to "take

[12] Techniques, such as self-destructing models, can be employed to make removing safety mechanisms more difficult.

[13] Some fine-tuning or modification of the model may still be possible.

[14] https://learn.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility-ai

[15] For example, if the model consumer has an obligation to report malicious use of their AI system, they may be dependent on the service provider providing the information needed to meet this obligation.

back" widely available weights. This allows more direct control of the risks than binary public/non-public decisions that lack nuance and can exaggerate the impact or probability of risks (and benefits).

Part of a broader risk assessment, this evaluation of the impact of specific attributes can also help to assess the marginal risk of different options, including non-AI options. For example, if you know that a specific AI model does not pose a greater risk than information currently available on the internet,[16] then you may determine that the breadth of distribution does not have a meaningful impact on the marginal risk.

For some risks – such as the use of a highly capable model to generate chemical, biological, radiological, and nuclear (CBRN) content in a way that threatens safety and security – the impact of a risk or capability may be so significant that the breadth of distribution (which primarily affects the probability of a risk) does not change the policy interventions that are warranted. In these scenarios, the presence of certain capabilities (as defined by the appropriate regulatory agency, such as the Department of Energy for nuclear), could trigger specific obligations, such as licensing of the model or the environment running it, or additional restrictions, such as export restrictions.

As the NIST AI Safety Institute Consortium[17] (AISIC) develops evaluation criteria, and NIST and OMB develop guidance for Federal government development and use of AI systems, we encourage them to focus on identifying the specific attributes that introduce risk and assessing the impact of those attributes for the specific use of the AI system. This is an approach that Microsoft has adopted as part of our *Responsible AI Standard* and its companion *Responsible AI Impact Assessment.*[18]

[16] https://www.rand.org/pubs/research_reports/RRA2977-2.html
[17] Microsoft is a member of the NIST AI Safety Institute Consortium (AISIC)
[18] https://www.microsoft.com/en-us/ai/tools-practices

What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?

Recommendations:

1. *The Department of Homeland Security, the Department of Energy, the Office of Science and Technology Policy, the National Science Foundation, and other relevant federal departments and agencies to partner with the open source AI community and academia to promote and adopt responsible AI design and deployment guidance.*

Open source software provides immense direct and indirect value to society,[19] and we believe the same will be true for open source AI models. This value comes directly from the reduced costs when you reuse components but is also derived indirectly from the innovation, competition, and collaboration arising from open source development. This open innovation cycle benefits not only open source projects but also the extensive value chain built on those projects.

Open source AI models and software also play a key role in improving AI safety and security. Specially trained AI models are being used for content filtering and to detect malicious use of AI systems, and open source libraries and applications are used to interact with models safely and to automate safety and security testing. Encouraging a diverse global community to participate in the development of these models is beneficial to the overall safety and security of the AI ecosystem.

Global collaboration on academic research is heavily dependent on the free exchange of information. AI code, models, and data are often developed and released "in the open," either under an open source license or licenses tailored for the researcher's objectives. This open collaboration creates transparency, provides opportunities for others to review and reproduce the research, and enables other researchers to build on the research. When released under an open source license, this AI code, models, and data can also contribute to further development of open source and commercial AI models and systems, further accelerating innovation.

Open source projects and open research provide valuable opportunities for students and the workforce to learn from, and to participate in, the development of innovative AI models and systems. These opportunities do not need to be restricted to the technical aspects of AI; they can also be an opportunity to teach responsible AI design and deployment practices. Promoting risk-based and outcome-focused guidance for responsible AI design and deployment to these

[19] https://www.hbs.edu/faculty/Pages/item.aspx?num=65230

communities will increase awareness of the risks and provide practical guidance on how to mitigate them.

Maintaining these benefits is dependent on deeply understanding the impact of policies on open source communities and their projects. This requires recognizing the significant diversity of open source development models and engaging with representative open source communities to seek feedback on policy impacts.[20]

Are there other relevant components of open foundation models that, if simultaneously widely available, would change the risks or benefits presented by widely available model weights? If so, please list them and explain their impact.

Recommendations:

1. *NIST AISIC to develop guidance on which components support which outcomes and considerations when releasing them.*

The OSI's *Open Source AI Definition Draft*[21] includes a representative list of components of an open AI model, including:

- Code
- Model
- Data
- Other

These components are subdivided further into subcomponents, with some required in an open source AI model release because they are essential to exercise the freedoms to use, study, modify, and share.

Each component and subcomponent has its own unique risks, benefits, and costs to release. The distribution method and openness of an AI model are two key factors, but others may include: intended audience (for example, researchers or downstream developers), consumer demand, risk and impact assessments, phase of development, and available resources. As part of planning the release of an AI model, developers should consider the outcomes they intend to achieve and the components necessary to support those outcomes, incorporating that context into their risk assessment for the release.

[20] https://www.stiftung-nv.de/en/publication/fostering-open-source-software-security

[21] https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-6

The following table provides an illustrative example of how intended outcomes can be mapped to required components and release considerations. Only selected intended outcomes are mapped; the table is not exhaustive in listing all intended outcomes.

Intended Outcome	Required Components	Release Considerations
Meet Open Source AI Definition	• Code – data pre-processing • Code – training, validation, and testing • Code – inference code • Code – supporting libraries and tools • Model – architecture • Model – parameters (weights)	• Available under OSI-compliant license
Production use	• Code – inference code • Model – model parameters (weights) • Model – model card • Other – usage documentation	• Completed impact and risk assessment
Benchmark verification	• Model – parameters (weights) • Code – code used to perform inference for benchmark tests • Data – benchmarking data sets • Data – evaluation metrics and results	• Privacy review of benchmarking data sets
Evaluation – WMDP[22]	• Code – inference code • Model – parameters (weights)[23]	• Review evaluation results for indications of harmful capabilities • Consider using CUT to unlearn harmful capabilities

As part of the NIST AISIC's working groups,[24] there is an opportunity to identify the specific components of an AI model that are required for each of the practices and processes that they define. This would create a shared understanding between model developers and model consumers of which practices and processes are possible based on the components that have been released. This shared understanding can also be leveraged across Federal government

[22] https://www.wmdp.ai/

[23] For a hosted or API access model, the WMDP evaluation can be performed without weights.

[24] https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute/aisic-working-groups

workstreams, including agency AI use case inventories, risk assessments, procurement requirements, and evaluation and research processes.

What are the legal or business issues or effects related to open foundation models?

Recommendations:

1. *Encourage global research and collaboration on AI fundamentals, AI safety and security, and AI applications.*
2. *Support the responsible release of AI models under safety- and security-enhancing terms and avoid unintended export control implications.*

Microsoft Research works as part of the global research community to enhance our understanding of artificial general intelligence, create new model architectures with novel abilities, achieve societal benefit through the advancement of AI, transform scientific discovery, and extend human capabilities. We work across disciplines – from social sciences to economics, to mathematics and physics – and collaborate across boundaries between research and engineering and with our peers in academia, industry, and government.

Over the past thirty years, Microsoft's research community and collaborators have published tens of thousands of papers and released hundreds of open-source projects and datasets. Nearly all of the research we produce is shared with the research community. Enabling free and open collaboration in the research community is essential to pursue advances in AI, align AI with human goals, ensure positive impact on jobs and the economy, and ensure equitable and safe employment in key sectors, such as education and healthcare.

The decision about what terms under which to release AI models, or their components, is based on many factors, including the developer's objectives,[25] consumer demand,[26] ecosystem norms,[27,28] and legal obligations[29] or restrictions such as export controls.[30] Even once this decision is made, it may change for future releases of the AI model as the benefits, risks, and other factors that influenced the original decision evolve.

NTIA should engage with interagency and interdepartmental entities, including the Bureau of Industry and Security (BIS), to ensure that the terminology "open source" or "widely available," as applied to foundational AI models, is consistent with the "published" definition under the

[25] https://github.com/todogroup/ospo-career-path/tree/main/OSPO-101/module2
[26] https://www.commerce.gov/about/policies/source-code
[27] https://www.gnu.org/licenses/licenses.html
[28] https://apache.org/licenses/
[29] https://www.gnu.org/licenses/gpl-faq.en.html#combining
[30] https://www.apache.org/licenses/exports/

Export Administration Regulations (EAR). Under Section 734.7(a),[31] to qualify as "published" under the EAR, a model must be freely and publicly available without limitation on distribution. It is crucial that the "open source" or "widely available" terms used by the Federal government are in harmony with the EAR's "published" standards to facilitate the unencumbered distribution of these models, thereby avoiding unintended export control implications.

For instance, many AI models are licensed under the MIT License, endorsing unrestricted use. In contrast, certain AI models are subject to the OpenRAIL License, which prescribes specific restrictions on usage in critical scenarios to guide the safe and responsible use of these models. Given these specific restrictions on usage, models that are "open source" or "widely available" according to NTIA's interpretation may not fulfill BIS's "published" criteria and could consequently be regulated by future export controls, undermining efforts to cultivate a healthy open source AI ecosystem that also advances responsible use and open collaboration on AI research.

There is a systemic relationship between research, open source, commercialization, and consumer adoption. Developers' ability to responsibly release AI models under their preferred licensing terms and distribution method is essential to innovate, compete, and partner in the global marketplace. Furthermore, cultivating a healthy open source ecosystem increases competition and provides more opportunities to develop AI talent.[32] Policies that put undue burden on these decisions can have chilling and unintended consequences on America's innovation and competitiveness.

What are current or potential voluntary, domestic regulatory, and international mechanisms to manage the risks and maximize the benefits of foundation models with widely available weights? What kind of entities should take a leadership role across which features of governance?

Recommendations:

1. *OMB and the General Services Administration (GSA) should use Federal purchasing power to incentivize safety and security in foundation models.*
2. *Extend existing cybersecurity engagements with open source foundations and projects to include open source AI safety and security.*
3. *Expand international collaboration and regulatory consistency through bilateral and multilateral agreements.*

[31] https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-734/section-734.7

[32] https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns

The Federal government can influence the safety and security of dual-use foundation models with widely available weights through its procurement practices. Specifically, the Federal government can impose requirements on AI systems powered by dual-use foundation models that agencies build or procure. This is the approach that Executive Order 14028 – *Improving the Nation's Cybersecurity*[33] used to encourage industry adoption of NIST's *Secure Software Development Framework*[34] (SSDF). This approach creates incentives for commercial providers of dual-use foundation models with widely available weights to evaluate and contribute to the safety and security of those models, benefiting not only Federal government customers but also other consumers of those models.

Over the past few years, the Federal government has been increasing its direct engagement with open source communities, primarily in relation to cybersecurity. These engagements include the White House's establishment of the Open-Source Software Security Initiative[35] (OS3I), an "interagency working group with the goal of identifying policy solutions and channeling government resources to foster greater open-source software security across the ecosystem," and the Cybersecurity and Infrastructure Security Agency's (CISA's) establishment of an Open Source Security Section.[36] These engagements with open source communities can also be opportunities to engage relevant stakeholders on issues related to open source AI safety and security. Additionally, the open source community is engaged through participation in the NIST AISIC.

Directly regulating the development of open source is difficult and likely to hamper open research and innovation, but prohibiting harmful uses, such as the production and distribution of non-consensual intimate imagery (NCII); regulating high-risk uses regardless of the openness of the AI models they use, such as the approach taken with respect to "high-risk AI systems" in the European Union's AI Act; and using voluntary, consensus-based international standards, codes of conduct, and guidance, such as NIST's AI Risk Management Framework, can be effective at minimizing harmful impacts and setting expectations with developers of AI models.

Continuing and extending partnerships with international AI Safety Institutes, policymakers, standards bodies, academia, industry, and civil society to develop and adopt voluntary, consensus-based international standards, codes of conduct, and guidance is needed to create consistency and to ensure that they address emerging risks and societal values.

[33] https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/
[34] https://csrc.nist.gov/Projects/ssdf
[35] https://www.whitehouse.gov/oncd/briefing-room/2024/01/30/fact-sheet-biden-harris-administration-releases-end-of-year-report-on-open-source-software-security-initiative/
[36] https://www.cisa.gov/opensource

In the face of continually changing technology, and given unforeseen risks and benefits, how can governments, companies, and individuals make decisions or plans today about open foundation models that will be useful in the future?

Recommendations:

1. *Promote adoption of risk-based and outcome-focused frameworks and standards that are adaptable to changing technologies and use cases.*
2. *Monitor emerging and unforeseen risks and benefits and assess their impact.*
3. *Plan and implement incident response processes to minimize incident impact.*

Policies that are targeted, that focus on outcomes rather than implementations or technologies, and that impose requirements commensurate with risk are most likely to achieve policy objectives and prevent harm. This approach also typically results in policies that are more resilient to changes in technology and that place the burden on those best able to bear it (e.g., implementers of high-risk use cases).

While the development of consensus-based international standards can be slow, the nature of this process is that it creates risk-based and outcome-focused standards that can be applied to a broad range of organizations, technologies, and applications. Policies grounded in international standards can minimize risk and maximize interoperability while maintaining competitiveness by minimizing costs and duplicative compliance efforts. When coupled with international reciprocal agreements, these benefits can be realized up and down the global supply chains underlying modern systems.

To address the gap while these standards are being developed, policies can use voluntary codes of practice, guidelines, and frameworks, such as the NIST AI Risk Management Framework, and incentivize developers through consumer demand and Federal purchasing power. Evaluations, such as those being developed and adopted by international AI Safety Institutes, and continued research will also drive further improvements to safety and security across the AI ecosystem and provide valuable inputs to the development of standards. The establishment of industry groups focused on frontier model development and deployment, such as the Frontier Model Forum,[37] is a positive step in the right direction by industry and increases the likelihood of voluntary policy actions being implemented and effective.

The progress of AI Safety Institutes and groups like the Frontier Model Forum could also help establish scientifically valid techniques and instruments for determining that a model poses

[37] Microsoft is a founding member of the Frontier Model Forum (FMF)

significant safety or security risks, prompting an urgent need to limit and maintain awareness of any distribution of model weights. As discussed above, where risks are exacerbated by the breadth of distribution, phased releases that allow for measuring and responding to unexpected risks at a smaller scale mitigate the risk of being unable to "take back" widely available weights. NTIA may also consider other risk assessment and mitigation approaches as it confronts this challenge with potentially emergent risks.

Risk assessments and mitigation frameworks will also need to factor in our learned experience that model-level safety interventions are necessary but not sufficient. AI safety and security require additional safeguards to be layered in when building and deploying AI applications, ultimately increasing durability and impact. Arvind Narayanan and Sayash Kapoor recently captured this point well, highlighting that "AI safety is not a model property" and that "[s]afety depends on a large extent on the context and the environment in which the AI model or AI system is deployed."[38]

Moreover, even when model and application developers take all reasonable precautions to assess and mitigate risks, mitigations will fail, unmitigated risks will be realized, and unknown risks will emerge. These risks could range from generating harmful content in response to a malicious prompt to the intentional exfiltration of an advanced AI model by a nation state actor. It is essential that developers are prepared to respond to and mitigate these in a timely manner and, when necessary, partner with their stakeholders up and down the supply chain. Cybersecurity has benefited from the existence of standards (such as, *ISO/IEC 29147 – Vulnerability disclosure*[39] and *ISO/IEC 30111 – Vulnerability handling processes*[40]), programs (such as, *CVE*[41]), systems (such as, the *National Vulnerability Database*[42]), and reporting requirements (such as, *Cyber Incident Reporting for Critical Infrastructure Act of 2022*[43]). While these are imperfect, and continue to evolve, they have created a shared understanding of how to responsibly handle incidents and have been universally adopted – mostly voluntarily – by organizations of all sizes, including open source projects.

[38] AI safety is not a model property (aisnakeoil.com)

[39] https://www.iso.org/standard/72311.html

[40] https://www.iso.org/standard/69725.html

[41] https://www.cve.org/

[42] https://nvd.nist.gov/

[43] https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/cyber-incident-reporting-critical-infrastructure-act-2022-circia

What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights?

Recommendations:

1. *NSF, NIST, and OSTP to fund and prioritize ongoing research, evaluation, and training on AI safety, security, and trustworthiness.*

There is active research into technological protections for AI models, such as differential privacy[44] - which could improve the privacy of training data, and self-destructing models[45] - which could make it more difficult to re-train widely available weights to be unsafe. Through risk assessments, we can identify attributes that contribute meaningfully to risk as well as where risk mitigations are not as effective as desired; we can then use this data to prioritize and fund future research to address these challenges.

The research into and operationalization of model evaluations that is underway across academia, civil society, industry, and governments is important to both assess the risk of dual-use foundation models with widely available weights and, more broadly, to advance AI safety and security. The government-led AI safety institutes, such as NIST's, play a vital role in defining the requirements for those evaluations, identifying gaps in existing evaluations, prioritizing new evaluations, and providing trusted environments for executing and reporting the outcomes from evaluations. Funding of these trusted environments, such as ones provided through NIST's AISIC or through the National AI Research Resource (NAIRR) Operating Entity, will be essential to meet the growing demand for evaluations.

With AI being a field undergoing rapid innovation, it is essential that risk assessments and evaluations are regularly reviewed to ensure they are suitable for the changing landscape. For example, small language models (SLMs) – such as Microsoft's Phi-2 – are demonstrating the ability to outperform significantly larger models,[46] and improvements to training data curation and processes are significantly reducing the resources needed to train large language models.

How AI models are used in AI systems, and the broader architecture and implementation of those AI systems, has a significant impact on the resulting risk. For example, even if an AI model exhibits specific risks in an evaluation, its use in an AI system may not expose those risks, or the AI system may combine multiple AI models and filter inputs and outputs to mitigate those risks. The evaluation of an AI model is an important input to the responsible design and

[44] https://csrc.nist.gov/pubs/sp/800/226/ipd

[45] https://arxiv.org/abs/2211.14946

[46] https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

implementation of an AI system, but it is not a substitute for assessing the risk of the AI system in its entirety, and it is not a substitute for assessing the risk of safety- and security-impacting uses of an AI system when deployed.