CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY | GEORGETOWN UNIVERSITY

March 27, 2024

Agency name: National Telecommunications and Information Administration
Federal Register Document Citation: 89 FR 14059
Organization: The Center for Security and Emerging Technology (CSET)
Respondent type: Organization>Academic institution / Think tank
Primary POC: Kyle Miller (kam471@georgetown.edu)

The Center for Security and Emerging Technology (CSET) at Georgetown University offers the following comments in response to NTIA's Request for Comment (89 FR 14059) on Dual Use Foundation Artificial Intelligence Models With Widely Available Model Weights. A policy research organization within Georgetown University, CSET provides decision-makers with data-driven analysis on the security implications of emerging technologies, focusing on artificial intelligence, advanced computing, and biotechnology. We appreciate the opportunity to offer these comments.

Our response pertains to six of the topics from the enumerated list provided in the RFC:

1. Definition of Open Foundation Models
2. Risks
3. Benefits
4. Managing Risks
5. Governance Mechanisms
6. Planning for the Future

Definition of Open Foundation Models

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

An open or widely available foundation model has weights available online that can be readily downloaded, shared, and moved to any computing infrastructure the user sees fit. This includes model weights that require a prospective user to first accept licensing terms and acceptable use policies before being given access to the weights (e.g., LLama 2). **We do not consider a model to be open if the weights are only accessible through an API**, as the ability to obtain, use, share, and control the weights on a separate computing infrastructure is restricted by the original developer. If model weights are available through a gated release, where only a select set of actors have been given direct access, the analysis is more complicated. We propose that model weights should not be considered widely available if it is feasible and likely that the gated release would be enforced up to and including legal action, for example by model developers pursuing claims against unauthorized disclosure to third parties. If model developers are relying on gating contracts that are unenforceable or unlikely to be enforced, the weights should be considered widely available.

We also recommend removing the "tens of billions" parameter threshold from NTIA's current definition of "foundation models." This number is arbitrary, and many foundation models fall below that threshold. Foundation models can have widely varying parameter counts that depend on the task it was developed to perform, the amount of data used to train it, etc. Language models can be in the hundreds of billions of parameters, while computer vision, text-to-image, and biological foundation models (e.g., Evo) can be in the millions to billions of parameters. Moreover, many popular language models on Hugging Face (measured by download count) have fewer than 10 billion parameters.

(1.b) Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?

The best way to gauge such timeframes may be to directly contact organizations designing foundation models and acquire information regarding their model performance and release strategies. This is the most viable way to get these estimations, although these organizations may not have the will or obligation to provide such information.

Organizations developing foundation models decide whether to release weights, and it currently happens to be the case that organizations making the highest-performing models (OpenAI, Anthropic, and Google) decide to not release them. There are no inherent structural differences between open and closed models, only differences in how developers decide to release weights. That being said, companies may have fewer incentives to keep weights closed if their models are small (cheaper to make) and have lower performance. Conversely, if a company develops an expensive, high-performance model, then it will have more incentive to monetize it and keep the weights closed. This is the case for state-of-the-art models developed by Anthropic and OpenAI, all of which are closed. However, some companies may see longer-term opportunities in opening weights, especially if it comes with licenses that permit downstream monetization. For example, according to Meta's LLama 2 license, if an application that uses LLama 2 exceeds 700 million users a month, then "you must request a license from Meta, which Meta may grant to you in its sole discretion."

(1.c) Should "wide availability" of model weights be defined by level of distribution? If so, at what level of distribution (e.g., 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be "widely available"? If not, how should NTIA define "wide availability?"

"Widely available" should be defined by the level of access (i.e., can the weights be downloaded from the Internet), not by the level of distribution (i.e., how many actors have downloaded the weights). Regardless of the platform or medium through which weights can be shared or distributed, a model should be considered "widely available" if the weights can be downloaded from a website and shared without restriction.

(1.d) Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?

Risks and benefits can depend significantly on the accessibility of the weights.

Hosted access. Through hosted access, users can prompt a model and access its outputs via a web-based user interface, but they have no access to the weights or the inner functionality of the model. The developers have full control over the model, and outside users cannot alter or fine-tune it. This can reduce both potential risks and benefits. The developers can oversee and mediate how actors use the model, and safety measures (e.g., RLHF that reduces model biases or restricts responses to certain prompts) can be put in place to reduce misuse. However, the benefits are also reduced, as the ability to scrutinize or alter the models is restricted, and the onus for transparency is solely on the original developers.

API-based access. With API-based access, the original developers still maintain a high degree of control and the weights are closed, but the level of access may be greater than with hosted access, depending on the functionality provided through the API. For example, the GPT-3.5 API allows users to fine-tune models, while the Claude API does not allow fine-tuning. Developers can oversee and mediate how actors use or fine-tune the models, and can take action against actors who use or fine-tune models maliciously (or in ways that infringe on the user agreement stipulated by the developer).

However, benefits are also reduced when weights are only accessible through an API. For example, users can fine-tune GPT-3.5 via OpenAI's API, but this API is highly restricted. Users can only upload training data and set the number of epochs to fine-tune the model. The ability for actors to customize the models, scrutinize their behavior, and alter their core functionality is constrained. This can reduce the benefits to R&D and innovation, and puts the onus for transparency on the original developers. Policymakers should consider how different degrees of API-based access can improve the benefits without increasing risk, such as supporting more research-friendly APIs that provide more access to and customizability of closed models (either permanently closed models, or models that will eventually have open weights).

Downloadable weights. When model weights are downloadable, anyone can use or alter the model on their own computing infrastructure, thereby enabling both local hosting and edge deployment. This can increase both the potential risks and benefits. On the one hand, risks of misuse are greater because the original developers have no means to oversee or mediate how the model is used or altered; the weights can be readily downloaded, fine-tuned, and used by nefarious actors. On the other hand, the benefits to R&D are also greater, as more researchers can assess, customize, and use the models freely and without restriction (depending on the license). Downloadable weights can also lower the barrier to entry in AI R&D, and may reduce the concentration of power in the AI industry because the original developers do not control access to the models. Most open models fall within this 'downloadable' category, and most of these downloadable models fit along a spectrum of openness based on the accessibility of their components and information about their development – all of which is (or is not) provided by the developers.

Fully open models. Models are fully open when all of their components (e.g., weights and training data) and documentation on how they were developed (e.g., how the model was designed and evaluated) are available online. This means the model is replicable because anyone with sufficient resources can access all of the components and information to recreate it from scratch. However, very few open models are fully open and replicable (e.g., BLOOM).

(1.d.i) Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?

Providing third-party researchers with greater structured access to closed models would be beneficial, and is unlikely to entail more risk than providing hosted access. Prior to releasing model weights, organizations could provide structured access via more research-friendly APIs. This could include allowing users to automate sampling from a model, customize the fine-tuning specifications, and inspect the model's internals (parameters, activations, embeddings, etc.). This would allow a larger pool of actors to scrutinize and experiment with the models, as well as crowdsource some of the red teaming process, which could help inform the developers about the potential risks before releasing weights.

However, control is still highly centralized with structure access, whereas open weights enable a greater degree of ‘democratized access’ where anyone can obtain, use, share, and experiment with models as they see fit.

Risks

(2.a) What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

Releasing training data or training code may help facilitate model replication. Replication of a model does not, in itself, pose any additional risk if the replicated model’s weights are already widely available, since any risks would already have been posed by the original model. However, access to training code and data could pose risks if it contains insights and information that are not obvious to malicious actors with the resources to train their own models. While malicious actors do not need that information to simply make use of the open model, it could help them build systems that are more capable.

For example, suppose a model exhibits some level of dangerous capability, C. This may pose a risk. Releasing the training data and code could allow other well-resourced actors to develop their own model with capability level C, which is unlikely to pose additional risks because they could simply have used the original model. However, if the training data and code released are novel to some other actors, it may help them develop a model with a capability level C+1 *higher* than the capability level of any current open model. If capability C+1 is sufficiently risky and is left unmitigated or used by malicious actors, this could pose an additional risk.

(2.d) Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy? How do these risks compare to those associated with closed models?

There are some unique characteristics of open model weights that should be considered.

First, some popular strategies to mitigate misuse cannot be applied to open models. For example, many developers use reinforcement learning from human feedback to prevent models from responding to certain kinds of harmful prompts. These safeguards can be removed if a model is fine-tuned. In principle, it may be possible to prevent this kind of harmful fine-tuning if it is performed through an API through further safeguards, for example by checking fine-tuned models for degradation in behavior before making them available to end users. However, it would be substantially more difficult (and potentially impossible) to implement reliable safeguards that would entirely prevent the capabilities from being elicited in cases where weights are widely available.

Thus, it is possible that mitigation strategies would be sufficient to reduce risk to acceptable levels if weights are not made widely available. However, these mitigation strategies would be insufficient for models with widely available weights.

Second, opening model weights is irreversible. Once model weights are widely distributed, it is not feasible to restrict them or gate their access. This significantly restricts the responses that deployers can take if new risks arise. For example, Microsoft and OpenAI restricted several threat actors from using OpenAI models, but they would not have been able to do so if that model had widely available model weights. Irreversibility is particularly important for the most advanced foundation models, because uncertainty about their capabilities and the effectiveness of their mitigations will persist for some time after their initial development (see above). If dangerous capabilities or gaps in mitigations are found after the wide release of model weights, there are far fewer responses that can be deployed. Risk assessments should take this factor into account.

These aspects of widely available model weights are simply aspects that should be considered when performing risk assessment prior to making model weights widely available. The overall framework for assessing risk from making model weights widely available need not be fundamentally different from the one used for other kinds of release methods.

Below is a non-exhaustive list of potential risks that could arise from present and future foundation models (whether closed or open), which should be considered before the release of model weights.

Disinformation. Foundation models can be used by state or non-state actors to more effectively spread disinformation. There is evidence that contemporary language models can reduce the cost of generating disinformation. In addition, voice clones and deepfakes are already being deployed in wartime and elections. Closed models that are accessible via hosted or API access can also be leveraged in this way, but the developers have more means to identify and disrupt this malicious use.

Non-consensual intimate imagery and CSAM. Foundation models can aid in the production of nonconsensual intimate imagery and child sexual abuse material (CSAM). The use of these models is documented, for example in the production of synthetic CSAM or nonconsensual images of celebrities.

Cyberattacks. Foundation models may aid nefarious actors in conducting offensive cyber operations. This includes using models to support spearphishing, reconnaissance, and social engineering. They may allow attackers to work more quickly, reduce the barrier to entry for various activities, or even detect and exploit vulnerabilities that humans cannot. At present, code generation capabilities of large language models can improve the productivity of developers, including presumably those engaged in cyberattacks. In addition, research suggests that models may have some limited capacity to engage in vulnerability discovery and exploitation autonomously or semi-autonomously. Overall, the current impact of AI on cybersecurity is uneven, and often depends on the intent, sophistication, and capabilities of the actors involved. For more details, see the UK National Cyber Security Center's report on the near-term impact of AI on the cyber threat.

Biological misuse. Foundation models could be employed in certain scenarios of biological misuse. The type of foundation model – whether trained on natural language or biological data – impacts the potential outcome, which actors are most likely to use it, and what safeguards are most likely to be useful. Chatbots trained on natural language data can help describe scientific concepts or provide links or references to additional resources, including those that a malicious actor could use to cause harm. Biological design tools trained on biological data, like

DNA or protein sequences, could help an expert malicious actor to predict modifications to pathogens or toxins that make them more severe, targeted, or otherwise harmful.

Importantly, however, these risks should be clearly defined and assessed within the context of the existing biorisk landscape. AI is not required to cause biological harm, and alternative methods are widely and readily available to malicious actors. For example, chatbots trained on natural language data could help malicious actors find information about how to carry out a biological attack. However, this information can easily be found from other sources, because scientific information is widely available in accordance with principles of scientific openness, rigor, and transparency. Evidence suggests that present-day chatbots are unlikely to significantly assist a malicious actor for this reason. Similarly, an expert malicious actor has access to several traditional experimental methods that make pathogens or toxins more dangerous. Thus, it may be more effective to address existing, foundational biosecurity gaps for which more comprehensive oversight could address both AI-agnostic and AI-enhanced risks.

Improper use. Careless or uninformed actors may use foundation models for purposes that those models are not suited to. For example, LLM-powered systems have been proposed to automatically respond to emails, but if these tools are vulnerable to adversarial attacks, they may exfiltrate private user data. These risks, present with any technology, are especially acute for foundation models since developers often do not specify their use cases or imply that they can be used in all but a specifically enumerated list of disallowed use cases. There are many present-day examples of foundation models failing when used improperly.

International Competition. It will be difficult to limit the dissemination of open models that may be valuable to state adversaries. Adversaries can access, use, and adapt open models released by U.S. companies. However, the extent to which it helps them compete in AI R&D is unclear.

Increased productivity for criminal actors. Foundation models may substitute for human labor in mundane tasks that are not themselves harmful. This could include composing communications, making financial transactions, generating ideas, etc.. While the resulting productivity gains are helpful for benign actors, they also affect criminal actors, especially those that suffer from a shortage of labor for tasks amenable to foundation models. We are not aware of research illuminating the extent to which present-day foundation models are enhancing the productivity of criminal actors.

Uncontrolled AI agents. Developers are aiming to build AI agents that can autonomously take actions such as navigating web browsers, controlling physical equipment, or even executing contracts. Capable variants of such systems could become uncontrolled without better technical mechanisms to align them to the intent of their user and/or respect requests to cease operation, particularly if such systems are capable of deceiving humans or exfiltrating themselves to unmonitored servers. Malicious or uninformed actors could also intentionally create and unleash uncontrolled AI agents. Evidence suggests that present-day foundation models are not capable enough to pose these risks, though they sometimes operate in an uncontrolled manner.

Note that the same model capabilities that could enable the risks above may also enable beneficial activity and defenses against those or other risks. For example, improved AI cyber capabilities may help with patching software and defending networks, and the increased productivity from foundation models may empower law enforcement. These improvements in defensive capabilities should be included as part of any risk mitigation assessment.

Benefits

(3.a) 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 foundation models may provide wide-ranging benefits, including improved AI R&D, market competition, international competition, oversight, and life sciences R&D.

AI R&D and Market Competition. Open sharing has helped drive much AI R&D, and there are several reasons why open weights may help this trend continue and accelerate. First, the barrier to entry in AI R&D is lowered for less well-resourced actors who cannot develop pre-trained models from scratch (e.g., academics and small-to-medium-sized enterprises). This moves control away from a handful of well-resourced companies, and toward a larger pool of actors. Over time, this may help reduce the concentration of power in the AI industry, foster greater market competition in certain areas, and promote a more diverse R&D ecosystem that better reflects the cultural and sociopolitical diversity of different communities (outside of the model’s original developers). However, the degree to which it could foster greater market competition is unclear, and it may be unevenly distributed across the industry. For example, while high-performing open models could allow more actors to alter and use them for various applications, the availability of such models could also disincentive companies from developing new models from scratch.

Second, platforms can emerge and act as open repositories for a global body of AI researchers and developers (e.g., Hugging Face). Communities can form around these platforms, fostering cross-pollination of research and decentralized collaboration.

Third, more actors can customize and fine-tune open models for specific applications and novel research, especially those with access to bespoke datasets that can be used to fine-tune models for unique tasks. This can be done on their own computing hardware and without restrictions or gatekeeping from the original developers, and the financial benefits from this can be distributed amongst a wider pool of developers.

While open models are generally understood to be a net benefit to R&D, it is still unclear the extent to which open models – that all have varying degrees of openness based on the accessibility of their components and documentation on how they were developed – can benefit R&D, lower the barrier entry, and promote competition in the AI industry. More research is needed to determine what types of research are enabled by open weights, and how that may allow more entrants into the market. Many prospective entrants may lack resources, and it is unclear the extent to which resource constraints may limit the benefits of open models to R&D. Actors may lack the data to fine-tune open models, or lack the compute to use or experiment with open models rigorously and at scale (although resources provided through the NAIRR pilot may help alleviate resource constraints).

International Competition. First, much of the open source software ecosystem resides within the U.S. and the territories of its allies, as do many of the organizations engaging in open AI R&D. This may help the U.S. leverage open innovation (to support their security and economic interests) more effectively than adversaries, as well as put it in a better position to shape international norms around AI. Second, an open, U.S.-promoted AI ecosystem may help foster multilateral collaboration on AI policy and security issues, as it may create a common and more transparent baseline from which all states can assess the technology.

Visibility and Oversight. First, many actors are transparent about how they develop and use open models, often more so than developers of closed models. Many support initiatives around transparency and openness (e.g., EleutherAI, Nomic AI, and Petuum), which can help facilitate oversight and enable more robust testing and evaluation. Second, platforms such as Hugging Face and GitHub can help aggregate and track a significant portion of open models, who develops them, how they are developed, and when new fine-tuned versions are released. This overall visibility may foster greater R&D and innovation, as well as greater competition in the AI industry, especially when compared to closed models (whose developers are often not transparent).

Life Sciences R&D. Foundation models for biological research are emerging as promising tools in biomedicine, ecosystem and climate resilience, agriculture, and biomanufacturing, among other fields. These benefits are largely derived from AI's ability to accomplish an important task at the center of scientific research: identifying complex patterns and pulling out the relevant variables from large amounts of information. In particular, purpose-built models loosely termed biological design tools (BDTs) have the potential to unlock new frontiers in basic and applied research. BDTs are trained on biological data, like DNA sequences or protein structures, and are designed to engineer, predict, or simulate biological molecules, processes, or systems. These tools have already shown success in a wide range of biomedical, pharmaceutical, and basic research applications, including helping researchers to design and optimize protein-based therapies, interpreting DNA sequences and their impact on biological systems, and identifying novel disease proteins and generating custom drug molecules to target them.

Most current BDTs are open models developed by academic labs. The life sciences community places a high value on scientific transparency and openness, and tends to favor open sharing of resources. Openness ensures scientific rigor and reproducibility, as demonstrated by high-impact journals like Cell and Nature that require code, algorithms, and software to be peer-reviewed and released upon publication. Shifting away from open sharing of model weights would also require additional resources, as many academic researchers do not have the time, funding, or infrastructure to set up and maintain an API.

(3.b) 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?

More actors can scrutinize open models, identify vulnerabilities and safety issues, and implement patches and safety measures. There is a long history of open-sourcing software to crowdsource vulnerability discovery, which typically makes the software far more secure than if it was only assessed in-house. More actors can also assess open model functionality, limitations, and biases. This improves overall auditability, which may bolster the reliability and trustworthiness of downstream AI applications.

When a model is open and accessible to many actors, it can be more rigorously scrutinized, and issues are more likely to be identified. However, it is unclear the extent to which issues with open models, once identified, can and will be remediated. For example, if a vulnerability in a popular open base model is identified, then that vulnerability may also exist for all fine-tuned versions of that base model. It is unclear if and how downstream developers will address issues that stem from the original base model. In addition, some challenges with foundation models, such as out-of-distribution robustness and adversarial robustness, remain unsolved despite years of research effort, indicating that some vulnerabilities may be very difficult to fix in general.

Managing Risk

(5.a) What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

Potential risks should be assessed in the context of each model: different models may have different capabilities and pose different risks. Model evaluation has been discussed extensively elsewhere, so this section provides only an overview of the assessments that model developers, or others, should conduct to gauge the risks posed by a particular model equipped with a particular set of mitigations. These assessments should be conducted for all models, whether open or closed.

Capability Assessment. It is necessary to first assess a model's capabilities, and then determine how those capabilities may translate to risks at different levels of severity. Assessments should begin with the base models that lack any mitigations or built-in safety features, particularly if the developers will eventually release the weights (where such features can later be removed). These evaluations are not always straightforward, since some questions (e.g., whether a language model could assist with developing a bioweapon) may need to be assessed with controlled experiments involving human subjects, which is costly and takes time.

For models intended to be deployed in certain use cases, models should also be tested for their potential to cause harm through ineffectiveness in that use case (see this report). Effectiveness evaluations include human interaction testing, which focuses on the experience of the human interacting with the system, and system testing, which examines the systems in which the model is embedded and how the model affects (and is affected by) those systems. Moreover, evaluations should consider likely outcomes when the model is used by trained, untrained, or malicious users, as well as outcomes if the model is used in unintended use cases.

Specific evaluations related to risks may include whether models can:

- Produce convincing yet false or illegal content (e.g., this report)
- Assist with or autonomously conduct cyberattacks (e.g., this report)
- Assist with the design or production of chemical, biological, radiological, or nuclear weapons (e.g., this report)
- Deceive human evaluators
- Self-replicate or otherwise evade human control (e.g., this report)

Specific evaluations related to benefits may include:

- Economic productivity studies (e.g., this report)
- Assessments of model performance on important scientific problems (e.g., this report)

Mitigation Assessment. Second, it is necessary to assess any mitigations that are applied to reduce the risk of a model. Mitigations could include *internal* mitigations, such as those designed to alter the model to refuse certain prompts or increase its reliability, such as through reinforcement learning from human feedback. Mitigations could also include *systemic* mitigations external to the model, such as gene synthesis screening for biological risks. The effectiveness of any such mitigations must be rigorously evaluated in both ordinary scenarios, unusual ("out of distribution") scenarios, and scenarios involving concerted adversaries seeking to contravene them.

Uncertainty Assessment. Finally, evaluations of capabilities and associated mitigations are imperfect. In some cases, model capabilities have been discovered (or "elicited") only after a model is developed and deployed. Mitigations can also be contravened through new mechanisms after they are released. The developer's remaining uncertainty after these assessments should be explicitly taken into account when conducting risk analysis. Uncertainty is likely to reduce over time with testing and real-world use, but may initially be high. As we will describe below, assessing uncertainty is particularly important for open models.

For any given model and release method, the core question is 'do the mitigations available for that release method reduce risk to an acceptable level, taking uncertainty into account?' Given the rapid development of foundation model research, these assessments should be conducted with all new systems and release methods, and regularly for those systems thereafter.

Moreover, developers should analyze the marginal risk that foundation models pose relative to the status quo by considering existing risks and defenses absent foundation models.

For models where some degree of deployment (e.g., API-only access) poses an acceptable risk, model developers should then assess additional risks that could arise from making model weights widely available. Such assessments should include:

- Estimates of the reliability of risk assessments. Since it is not possible to reverse the act of opening model weights, uncertainty in overall risk assessments may cause model weight release to be unacceptably risky in some cases. Therefore, model developers may need greater confidence in their assessments for risk to be brought to acceptable levels for models with open weights.
- Estimates of the robustness of risk mitigations for open models. Mitigation measures external to the model, such as DNA synthesis screening for biological risks, are unaffected by the openness of model weights. However, measures internal to the model, such as fine-tuning designed to prevent models from responding to certain questions, can often be much more easily circumvented for models with open weights. If the risk posed by a particular model is acceptable only given some set of mitigations and those mitigations cannot be maintained in models with open weights, then opening the weights may be unacceptably risky.

(5.b and 5.c) 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?

Developers should adopt robust cybersecurity practices to ensure that model weights do not become available. This is similar to defending any other kind of intellectual property, but may also include efforts to identify and defend against methods for obtaining weights through queries to an API-only model (e.g., a recent paper showed that it was possible to obtain the weights of the final layer of LLMs simply by running a large number of specific queries).

To reduce security risks in cases where model weights are widely available, both “internal” and “external” mitigations may help.

First, mitigations internal to models could reduce risk, though it appears unlikely that robust solutions will be developed. For example, the unlearning literature focuses on removing dangerous information from models. Currently, most of this work is not robust to fine-tuning, but future research could potentially make it robust. Other work aims to make it difficult or impossible for malicious actors to fine-tune models for certain use cases, but the methods are currently very limited and come at great expense. Given the degree of control afforded with access to model weights, there may never be practical internal mitigations.

Second, mitigations external to models could reduce risk. For example, if a capable foundation model is effectively used for identifying and fixing software bugs, then the risk of making the weights of similar models widely available would be reduced, as it would become harder for malicious actors to use the model to find unresolved vulnerabilities. Similarly, DNA synthesis screening can help reduce risks from biological threats posed by future open models. Finally, "staged release" can help mitigate risk by providing models to trusted defenders, prior to making the weights available. While these external mitigations are essential and could substantially reduce the risks posed by open models, they may not be sufficient in all cases.

(5.d) Are there ways to regain control over and/or restrict access to and/or limit use of weights of an open foundation model that, either inadvertently or purposely, have already become widely available? What are the approximate costs of these methods today? How reliable are they?

Model weights can be readily shared and distributed on the internet. After weights are opened, it can be very difficult to "regain control," restrict their access, or limit their use. While fully restricting their access and use is largely infeasible, there are ways to partially limit or restrict the ease with which they can be disseminated and used:

- If the goal is to restrict access to a certain open model, then removing the model's weights from platforms like Hugging Face and Github (which requires cooperation from the platforms) can make it more difficult, but by no means impossible, for actors to obtain them. This is likely the most feasible and least costly way to limit the dissemination of weights. However, this strategy has limitations, as the weights can still be shared on a myriad of other websites and platforms.
- If model weights are leaked or otherwise released illegally, then there may be an opportunity for legal action against websites that host them. The prospects for pursuing legal action (or threats of it) to coerce actors to remove leaked weights from their websites depends on the actors. For some, particularly those on the dark web who already engage in illicit activity, it will be very difficult or impossible to get weights removed. Like many other illicit activities on the dark web (e.g., selling malware or ransomware-as-a-service), the threat of legal recourse is not an effective disincentive.
- Many actors use cloud compute services to run or fine-tune models. There may be some opportunity for cloud service providers (CSP) to restrict users from running for fine-tuning certain models on their infrastructure. However, the CSP would have to

identify when restricted models are running on their hardware, which would be very challenging and potentially quite costly. Moreover, if actors alter the functionality and/or surrounding code of an open model, then the model's signature may also change, which would make it more challenging for CSPs to know whether or not the model running on their infrastructure is restricted.

Overall, there is no reliable way to fully restrict access to a model after the weights are open, but there are ways to limit their dissemination and use (some of which can come with high costs).

(5.f) Which components of a foundation model need to be available, and to whom, in order to analyze, evaluate, certify, or red-team the model? To the extent possible, please identify specific evaluations or types of evaluations and the component(s) that need to be available for each.

There are different types and degrees of all of these activities. It is not only a question of what components need to be available, but a question of what components may enable more thorough and robust degrees of analysis, evaluation, certification, and red-teaming.

Models with publicly available weights fit along a spectrum of openness, and where they fit depends on the accessibility of their components (e.g., is the training data or surrounding code used to run or fine-tune the model readily available?), as well as the degree of transparency and documentation on how they were developed (e.g., how was the training data collected and processed? How was the model designed and evaluated?).

All of these variables should be considered, as they can impact the ability to replicate and audit models, explain their outputs, scrutinize their functionality, and assess their capabilities and flaws. For example, if the goal is to scrutinize a model, then access to the training data may help explain its outputs; if the goal is to replicate a model, then access to most all of the components and development processes is typically necessary.

However, more research is needed to gauge how different degrees of access and transparency can impact the ability to scrutinize or evaluate open models. For example, many open models come with documentation and model cards, but the level of detail in these documents can vary dramatically, and they can enable (or not enable) different degrees of evaluation.

Notwithstanding the need for more research in this area, below is a list of model components and information that, if made available, could improve analysis, evaluation, certification, and red-teaming:

- Model Components
    - Model Weights
    - Inference code
    - Fine-tuning code
    - Training code
    - Training data
    - Auxiliary code (data code, custom libraries, etc.)
- Model Information
    - Paper, preprint, blog, and/or model card that describes the model (in varying degrees of detail)
    - Documentation
        - Performance evaluations and benchmarking
        - Risk assessments
        - Data preparation, provenance, and validation
        - Configurations
        - Checkpoints
        - Hyperparameters
        - Logs

Note that even limited access to closed models enables a degree of analysis, evaluation, certification, and red-teaming. This is the case for closed models that are available through hosted or API access. This limited access still allows a user to prompt the model and analyze its outputs, which can help with assessing model performance (e.g., how well the model performs on benchmarks) and capabilities (e.g., can the model complete certain tasks).

However, the ability to assess these capabilities is limited to just prompting the model and examining the outputs. The model may have latent capabilities that researchers fail to identify if they do not prompt the model in a way that causes it to exhibit those capabilities. If there is no documentation on how the model works or was designed, then the ability to scrutinize its functionality is constrained; if there is no information on the training data, then it is more difficult to assess its potential capabilities. If an API allows a user to fine-tune a closed model, then there is more room to evaluate it, as researchers can fine-tune it with more data and assess how the model's behavior, performance, and capabilities change. However, API-based fine-tuning is often quite limited (e.g., the ChatGPT API only allows a user to upload fine-tuning data and set the number of training epochs).

Governance Mechanisms

(7.b) How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?

Open weights can help facilitate government action in AI regulation by providing more access, transparency, and visibility. It can reduce barriers for regulators and researchers to assess and experiment with models, which may improve multilateral collaboration on AI policy, as it creates a more transparent baseline from which all states can assess the technology. However, there are other ways to support these efforts, with or without open weights. Under the White House AI Executive Order, safety testing results must be reported for certain models, and in the UK, many developers provide the AI Safety Institute with access to their models, in some cases to unreleased models.

Open weights can also frustrate government action by making it more difficult to monitor or prohibit misuse. Developers of open models cannot oversee how they are used or fine-tuned, so the onus for transparency will be on downstream actors. Regulations that prohibit certain AI use cases may be more difficult to enforce, given that more actors can use open models in unmonitored settings. Moreover, many open models are developed abroad, which further complicates unilateral AI policies and regulations.

(7.d) 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?

First, assist in developing standards and best practices for determining whether the release of model weights is appropriate, based on risk analysis described in section 5.a. These standards should be developed at the NIST AI Safety Institute in collaboration with its Consortium. Standards should focus on the development of evaluations and mitigations to identify and bound risk at acceptable levels, as well as the establishment of risk tolerance levels to define "acceptable" risk.

Second, assist in developing evaluations for measuring the benefits of open models. For models where risk is low and benefits are significant, the government could consider incentivizing developers to open model weights, such as by conditioning grants from the NSF or NAIRR. This may be especially valuable for models designed for scientific, healthcare, or other beneficial tasks, that have been trained appropriately to minimize risk.

Third, assist in establishing infrastructure for incident reporting (for both closed and open models) and supporting resources that track the interdependencies of foundation models. Incident reporting captures data about situations where AI systems cause harm, which allows researchers to track trends and helps prevent the harms from recurring. Data on AI harm can help policymakers and researchers monitor whether approaches to mitigate risks are effective, and determine where new approaches might be needed. Resources that track the interdependencies among foundation models (e.g., Ecosystem Graphs) can complement incident reporting. If a model is linked to a surge of AI incidents, then other models that share similar components may also produce similar harms. Tracking these interdependencies can help illustrate the degree of exposure to risk that is first identified by incidents. The U.S. government could develop internal incident reporting and interdependency tracking functions, or support the development of these functions by outside entities.

(7.e) What should the role of model hosting services (e.g. HuggingFace, GitHub, etc.) be in making dual-use models with open weights more or less available? Should hosting services host models that do not meet certain safety standards? By whom should those standards be prescribed?

As described in section 5.d, removing weights from hosting services may help reduce the dissemination of models that are deemed too risky. But as described in section 7.d, more work by industry and government is needed to establish standards around model safety, risk assessments (and strategies around releasing weights based on those risk assessments), and acceptable levels of risk. Without such standards, hosting services will have to unilaterally gauge what models they are willing to host, and what levels of risk they are willing to accept.

(7.i) Are there effective mechanisms or procedures that can be used by the government or companies to make decisions regarding an appropriate degree of availability of model weights in a dual-use foundation model or the dual-use foundation model ecosystem? Are there methods for making effective decisions about open AI deployment that balance both benefits and risks? This may include responsible capability scaling policies, preparedness frameworks, et cetera.

Responsible scaling policies and preparedness frameworks have been a promising first step towards risk management for dual-use foundation models, but so far have been applied primarily to closed models. Such policies could be expanded to provide guidance for opening model weights. Preparedness frameworks encourage proactive planning for future capabilities, measurement and assessment of risk, and precommitment to risk mitigation. Such measures could be applied to future models before weights are released.

Planning for the Future

(8.b) 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 1026 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.

It is unclear how reliable compute thresholds will be over time, as well as what unintended consequences may result from them. Below we outline some pros and cons of training compute thresholds (in general), and the specific threshold of 10^26 operations.

Pros

- Compute is a readily quantifiable metric, and scaling laws have shown that large models trained with more compute tend to exhibit better performance than models trained with less compute. Subsequently, compute thresholds can be used by the government to identify and manage 'frontier' AI models. Based on publicly available data, no models have been trained on more than 10^26 operations. If the goal of the threshold is to roughly capture the space of 'uncharted' model training (and potential new risks stemming from models trained with that much compute), then the current threshold may be appropriate. If this is the goal, the threshold would likely need to be increased over time (and perhaps adjusted to include non-compute measures, depending on how the frontier of the field develops).
- The compute that will be expended in training a model can be estimated with fairly high precision prior to training. Organizations that must comply with threshold-based requirements can anticipate which requirements are likely to apply to any given training run.

Cons

- There are diminishing returns and limitations to compute scaling, and much progress in AI has been due to algorithmic improvements (in addition to compute). Research has found that "the compute required to reach a set performance threshold has halved approximately every 8 months." Therefore, it is unclear how much compute will be required to reach certain performance levels, and how reliable compute thresholds will be over time.

- Developers of the highest-performing models (i.e., Claude 3 Opus, GPT-4, and Gemini Pro) have not disclosed the amount of compute required to train their models. It is generally understood that these frontier models were likely trained on more compute than any other models seen to date, but without verified information, it is more difficult to gauge how compute thresholds can be used as a proxy for frontier model performance and capabilities. Some have roughly calculated the training compute of these models, but the calculations can be speculative and may have incorrect assumptions.
- A compute-based threshold may not be appropriate if the goal is to encompass all models that may pose risks. Models that fall below the threshold may still pose risks, and as computational and algorithmic efficiency improves, more models with potentially risky capabilities are likely to fall below the threshold.
- Currently, a 10^26 threshold would not be burdensome to most organizations, given that none have trained a model at or beyond that threshold. However:
    - As compute improves and the cost of training models to 10^26 operations is reduced, more organizations will likely be burdened as they develop models that reach or exceed the threshold.
    - If the threshold were to be lowered, then there is greater potential for undue burden to researchers and industry. Moreover, with lower thresholds comes more disclosures and red-teaming assessments from more organizations, but it is unclear if the government will have the expertise and resources to thoroughly vet the disclosed information.
- Overestimating the utility of a 10^26 threshold may lead to complacency and political inaction in other areas, under the assumption that the threshold will reliably encompass models that pose the most serious potential risks.
- Specific thresholds may incentivize developers to ‘game’ the metric, and train their models to a degree that falls below the threshold, thereby circumventing requirements to disclose information on their models.