Considerations for Dual-Use Foundation Models with Widely Available Weights

Insights to Inform a Request for Comment by the National Telecommunications and Information Administration

Everett Smith, Gaurav Sett

RAND Global and Emerging Risks Division

PE-A3309-1
March 2024

Published by the RAND Corporation, Santa Monica, Calif.

About This Paper

RAND's Technology and Security Policy Center (TASP) conducts research on dual-use technologies, such as artificial intelligence (AI), and their relevance to the national security of the United States. As part of this work, TASP has investigated the risks and benefits of dual-use AI foundation models and policies that are relevant to models with widely accessible weights. This paper provides insights from RAND experts in response to the National Telecommunications and Information Administration's Request for Comment related to the Biden administration's Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Technology and Security Policy Center

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Technology and Security Policy Center, which explores how high-consequence, dual-use technologies change the global competition and threat environment, then develops policy and technology options to advance the security of the United States, its allies and partners, and the world. For more information, contact tasp@rand.org.

Funding

Funding for this work was provided by gifts from RAND supporters.

Acknowledgments

The authors thank Kristin Leuschner, Karson Elmgren, Ashwin Acharya, Mauricio Baker, Greg Smith, and Robin Meili for their technical insights and assistance with this paper. We are particularly grateful for the valuable insights from our reviewers, Li Ang Zhang and Marjory Blumenthal

Summary

The National Telecommunications and Information Administration (NTIA) has requested comment on the "potential risks, benefits, other implications, and appropriate policy and regulatory approaches to dual-use foundation models for which the model weights are widely available."[1] Drawing on our expertise and ongoing research, we respond to NTIA's request by discussing the benefits and risks of open artificial intelligence (AI) foundation models and highlighting *red lines*—risk thresholds that warrant significant responses if crossed—as a potential approach to balancing the risks and benefits of open foundation models.

Key Takeaways

- **Open foundation models can provide benefits, such as reducing AI market concentration and accelerating AI safety research.** For instance, increased model access is helpful for dangerous capabilities evaluations and AI interpretability research.
- **Current evaluations show no significant biological or cyber risks from existing open foundation models.** However, most evaluations do not use all available techniques to elicit model capabilities, and existing proprietary models show higher capabilities. As foundation model capabilities grow, future models may pose larger risks.
- **There are more-secure methods to capture the benefits of open foundation models than by publishing model weights.** For example, *structured access* to models would allow for greater transparency and independent AI safety research while reducing the ability of malicious actors to misuse model weights if they contain dangerous capabilities.
- **Red lines provide a framework for targeted risk management of foundation models if they pose risks, while exempting less risky models.** In this context, red lines are the point at which the risks of publishing model weights clearly outweigh the benefits, as informed by assessments of the risks, benefits, and alternatives to publishing weights. This approach allows targeted interventions on risky models without stifling innovation from less risky models.
- **The European Union (EU) AI Act already includes features of a red-line approach.** The EU AI Act exempts open foundation models from many requirements, except when they are classified as posing "systemic risks" (e.g., a model is developed using large amounts of computational power [*compute*], on the scale of $50 million), in which case it must be evaluated for dangerous capabilities and steps must be taken to mitigate any risks.
- **Coordination with foreign partners will be critical to manage risks from open foundation models.** Attempts to control any open foundation models should consider which foreign actors could re-create equally capable models, as those operating in

[1] NTIA, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights," p. 14059.

countries with different rules could make their model weights publicly accessible, undercutting U.S. controls. Such forums as the G7, AI Safety Summits, the International Dialogues on AI Safety, and the International Scientific Report on Advanced AI Safety's Expert Advisory Panel could serve to facilitate that coordination.

Contents

Considerations for Dual-Use Foundation Models with Widely Available Weights

In this paper, we provide expert insights in response to the National Telecommunications and Information Administration (NTIA)'s February 2024 Request for Comment related to the Biden administration's Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[2] This paper focuses on *foundation models*, a term that refers to any artificial intelligence (AI) model "that is trained on broad data . . . that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks."[3] We also discuss *open foundation models*, which the NTIA defines as "[d]ual use foundation models with widely available weights."[4]

The introduction of open foundation models has led to discussions about the potential benefits of these models in terms of fostering innovation, transparency, and scientific advances. At the same time, concerns have been raised about the potential for such models to pose "risks to security, equity, civil rights, or other harms due to, for instance, affirmative misuse, failures of effective oversight, or lack of clear accountability mechanisms," particularly if the model's weights are made widely available.[5]

This paper is divided into two sections and addresses the following questions posed in the Request for Comment (shown in the order in which they are addressed in this paper):

- Benefits and Risks of Open Foundation Models
  - Question 3: "What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?"
  - Question 2: "How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?"
  - Question 1.i: "Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?"
- Red Lines for Targeted Risk Management
  - Question 8: "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?"
  - Question 7h: "What insights from other countries or other societal systems are most useful to consider?"

[2] NTIA, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights"; White House, "Executive Order 14110 of October 30, 2023."

[3] Bommasani et al., "On the Opportunities and Risks of Foundation Models," p. 3.

[4] NTIA, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights" p. 14060.

[5] NTIA, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights," p. 14061.

- Question 7g: "What should the U.S. prioritize in working with other countries on this topic, and which countries are most important to work with?"[6]

Throughout the paper, each question is shown in an italicized heading before the discussion of that particular topic.

Benefits and Risks of Open Foundation Models

Question 3: "What are the benefits of foundation models with model weights that are widely available as compared to fully closed models?"

A significant concern with the creation of advanced AI systems and the automation of human labor is the growing concentration of economic power in the hands of a few technology companies.[7] When these companies prevent open access to their AI model weights, users of the model must send and receive messages from an AI model hosted in a data center. As a result, the AI developer (or the data center owner) can monitor and shape the behavior of the AI system by automatically filtering messages between the AI and the user for harmful content, even if users disagree with the AI developer on what constitutes such content. Additionally, smaller AI companies can be prevented from adapting foundation models that are developed by large companies to a specific business case.

In contrast, open foundation models reduce barriers to entry to develop and fine-tune AI systems, providing the following benefits (among others):

- **Open foundation models may reduce market concentration.** When smaller actors can access open foundation models, they can avoid the large expense of developing their own models and can therefore better compete with large tech companies in adapting the foundation model to a particular business context.[8] Whether this access will be enough to maintain a competitive market for foundation model based products or services in general will depend on the price to develop and the performance of open models compared with closed models and on how the economics of fine-tuning, adapting, and serving foundation models differs in a particular business application between large and small companies.[9]
- **Open foundation models can decentralize decisions on acceptable model behavior.** Making model weights widely accessible would allow users, rather than AI developers or data center owners, to determine for themselves what model behavior is acceptable.

---

[6] NTIA, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights," pp. 14062–14063.

[7] Vipra and Korinek, "Market Concentration Implications of Foundation Models."

[8] Miller, "Open Foundation Models."

[9] Widder, West, and Whittaker, "Open (for Business)"; Miller, "Open Foundation Models."

However, setting up their own safeguards on model behavior would be technically challenging for smaller organizations that are deploying open foundation models.[10]

In addition, by providing greater accessibility to a wider variety of actors, open foundation models can enhance the following aspects of AI safety:

- **Publishing foundation model weights can aid in AI safety research**. *Sampling-only* access to foundation models, which is the minimum provided for closed AI models, is insufficient for AI safety research, such as evaluating models for dangerous capabilities or developing tools to interpret AI systems. Providing researchers with the model weights can better facilitate this research.[11] However, there are also more-secure alternatives to providing model weights that can capture many of the same benefits (see question 1.i).
- **Making foundation model weights accessible helps uncover vulnerabilities, biases, and potentially dangerous capabilities**. With a wider set of eyes examining these models, there is a higher likelihood of identifying and addressing issues that might have been overlooked by the original developers, as is the case with open-source software broadly.[12] This scrutiny is useful for developing AI systems that are secure, fair, and aligned with societal values. The detection and mitigation of biases in AI models, for instance, are critical steps toward ensuring that AI technologies do not perpetuate or exacerbate social inequalities.

Question 2: "How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?"

As discussed in the previous section, making model weights widely available has potential benefits for users of these models and for the public as a whole. However, the features of open foundation models might make them attractive to a variety of threat actors, including individuals and states, that seek to use the models for malicious purposes. Such purposes could include spreading disinformation, conducting surveillance, launching cyberattacks, or even developing biological, chemical, or nuclear weapons of mass destruction.[13]

To prevent misuse, AI developers often implement safeguards on their foundation models to limit dangerous capabilities, such as fine-tuning models to refuse to output hazardous content or filtering the inputs and outputs of models.[14] However, because open foundation models can be modified by users, these safeguards can be removed at low cost (e.g., $200).[15] At least some compromised systems might be very hard to detect.[16]

---

[10] Kapoor et al., "On the Societal Impact of Open Foundation Models."

[11] Seger et al., "Open-Sourcing Highly Capable Foundation Models."

[12] Boulanger, "Open-Source Versus Proprietary Software."

[13] Shevlane et al., "Model Evaluation for Extreme Risks."

[14] Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback."

[15] Qi et al., "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend to!"; Gade et al., "BadLlama."

[16] Hubinger et al., "Sleeper Agents."

Current evaluations do not show significant biological or cyber risks from existing open foundation models. The science of dangerous capability evaluations in AI is still evolving. Few dangerous capability evaluations have been conducted, and they have not covered all potential risks (we only cite cyber and biological risks in this paper, given their prevalence in discussions related to the Biden administration's AI executive order, but these risks do not represent our full range of concerns for open foundation models). Additionally, evaluations have not covered all open foundation models (or all currently closed models whose weights could be made widely available), and many evaluations have not made use of common techniques to elicit dangerous capabilities.[17] In light of these limitations, the results of evaluations should be interpreted with caution. However, several evaluations of various models for cyber misuse potential have not shown that existing open models possess significantly dangerous cyber capabilities (though leading closed models, which could become open, show stronger cyber capabilities).[18] Likewise, two evaluations did not find significantly dangerous biological misuse capabilities in current large language models compared with internet search.[19] Evaluations science will continue to evolve, and the capabilities of foundation models broadly may continue to improve at significant and surprising speed. Therefore, it may be warranted to prepare for a world in which evaluations provide evidence that foundation models possess significantly dangerous capabilities.

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

Publicly releasing model weights is an irreversible act. Once the weights are on the internet, they can be copied easily and are thus nearly impossible to remove. However, there are alternatives to widely releasing model weights that can capture many of the benefits while reducing risks. Importantly, these alternatives leave open the option to make model weights widely accessible in the future if there is evidence that doing so would pose only acceptable risks.

Structured access is an alternative approach that can provide users with many of the benefits of making foundation model weights widely accessible while reducing some of the risks. The most common approach to structured access is to create flexible application programming interfaces (APIs) that allow researchers, small businesses, or the public to access the model.[20] Updating APIs to provide greater access could significantly improve the transparency that third-party groups would have for proprietary AI systems, providing many of

---

[17] Model Evaluation and Threat Research, "Guidelines for Capability Elicitation."

[18] Fang et al., "LLM Agents Can Autonomously Hack Websites"; Hazell, "Spear Phishing with Large Language Models"; Shestov et al., "Finetuning Large Language Models for Vulnerability Detection."

[19] Mouton, Lucas, and Guest, *The Operational Risks of AI in Large-Scale Biological Attacks*; Patwardhan et al., "Building an Early Warning System for LLM-Aided Biological Threat Creation."

[20] Shevlane, "Structured Access"; Anderljung et al., "Towards Publicly Accountable Frontier LLMs."

the benefits of possessing model weights directly.[21] At the same time, structured access would prevent researchers and users from copying the model weights onto a new device and proliferating them to malicious actors. Once malicious actors—including individuals, nonstate actors, and nation states—possess the model weights, they could remove safeguards on model behavior and use the models to cause harm.[22]

Market concentration and research independence concerns might emerge if today's large, commercial AI developers are the sole body that decides how to provide structured access to their models. A noncommercial body could be granted the authority to manage or review decisions related to structured access to ensure it is being used equitably. Examples of bodies that could potentially fulfill this role are the National AI Research Resource, in partnership with foundation model developers and providers, and the Cybersecurity and Infrastructure Security Agency.

Red Lines for Targeted Risk Management

Question 8: "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?"

We introduce the concept of red lines as a framework for managing the risks and benefits of open foundation models.

What Are Red Lines?

Red lines are thresholds for dangerous capabilities that warrant a significant response if crossed. In the context of this response, red lines represent dangerous capability thresholds beyond which the risks of making the model weights widely available would significantly outweigh the benefits. Agreements on red lines should necessarily include discussions of the methods used to assess whether the red line has been crossed.[23]

Red lines are often discussed in the context of capabilities that are relevant to enabling chemical, biological, nuclear, or cyber attacks. However, red lines are a sufficiently flexible approach to risk management that various actors could design red lines based on their distinct set of risk concerns. For instance, actors concerned that AI models could be used to generate disinformation to interfere with elections could define red lines based on persuasion capability or hyperrealistic video generation capability.

---

[21] Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits"; Bucknall and Trager, *Structured Access for Third-Party Research on Frontier AI Models*.

[22] Shevlane, "Structured Access."

[23] In many cases, the capability of true concern can be difficult or impossible to measure directly, so, for practical purposes, a proxy must be used.

Example Red Lines

The following are some examples of potential red lines for biological, nuclear, and cyber capabilities:

- **Biological.** A graduate student in biology with access to an academic lab and a given AI system can recreate the smallpox virus in X amount of time for less than $Y.
- **Nuclear.** The AI system can, based on open-source information and its knowledge of nuclear physics and engineering, re-derive a given nuclear secret in less than X amount of time for less than $Y.
- **Cyber.** The AI system can discover at least one novel zero-click vulnerability in common or critical operating systems for less than $Y.[24]

These examples are purely for discussion. We think further work is needed to define functional red lines.

How to Design Red Lines

We outline the following potential steps in designing a red line:

- **Step 1: Brainstorm threats.** Red lines should emerge from a process of threat-modeling and scenario-planning. Brainstorming potential ways in which an AI model could cause harm would obviously benefit from subject-matter expertise, but avoiding groupthink and other blinders is also important. Therefore, threat models should be sourced not only from well-credentialed experts but also from a wide and diverse array of contributors . However, in many instances, threat modelling will involve sensitive information that must be controlled, and researchers will need to exercise appropriate caution when selecting and involving contributors. Striking the appropriate balance will be a challenge.
- **Step 2: Identify bottleneck capabilities.** A small number of AI capabilities will likely be bottlenecks for many threat models, making them promising candidates for red lines. Importantly, these capabilities do not need to appear imminent to be worthy of concern. Historically, progress in AI capabilities has surprised many experts.[25] Future capability surprises should be expected and prepared for.
- **Step 3: Brainstorm mitigations and responses.** A red line should tie a dangerous capability measurement to a response. Example mitigations for dangerous capabilities include pursuing alternative deployment strategies (such as structured access) or employing other safeguards (such as fine-tuning or watermarking). No such response is without flaw, and novel risk mitigation techniques will likely be required for future dangerous capabilities.
- **Step 4: Analyze the risks and benefits of capabilities and mitigations, given the threat models.** A sense of the risks and benefits of the capabilities, as well as the mitigations involved in proposed red lines, will be crucial for analyzing whether the red lines are reasonable.

---

[24] A *zero-click vulnerability* is a particularly powerful kind of vulnerability that requires no user interaction to compromise a device. Numerous such vulnerabilities have been discovered for smartphones. For more information, see Kaspersky, "What Is Zero-Click Malware, and How Do Zero-Click Attacks Work?"

[25] Musser, "How AI Knows Things No One Told It."

- **Step 5: Specify evaluation methods.** Ideally, agreements on red lines would specify, at least at a high level, the evaluation methods to be used to measure whether a red line has been crossed. Agreeing on these methods in advance might help make future discussions of measurement results more objective, but stakeholders also run the risk of the evaluation science progressing in the time between the agreement and the triggering of the red line, so agreements on methodology will need to factor this risk in.

What Are the Alternatives to Red Lines for Risk Management?

Red lines are a flexible approach to risk management and are not mutually exclusive with most other risk management strategies, such as liability, prescribed safety standards, defense in depth, and other strategies with an ex ante or ex post character. In fact, we think red lines complement many of these alternative risk management approaches.

What Are the Alternatives to Red Lines Based on Dangerous Capabilities?

We have outlined red lines as thresholds of *dangerous capabilities*, but one could also set red lines to be thresholds of *risk* (e.g., the probability × severity of harm) or of *training compute* (the amount of computing power used to build a foundation model).

Although risk (probability × severity of harm) thresholds would, in theory, better track the risks posed by foundation models than would dangerous capability thresholds, calculating the probability × severity of harm can be highly subjective in practice. Therefore, probability × severity of harm thresholds are poor tools for coordination because different parties might struggle to arrive at an agreed-on measurement result. Compute thresholds have the opposite problem. While they are much cheaper and more objective to measure than dangerous capabilities, they are also a much worse proxy for the risks posed by foundation models.

To balance these factors, one may both define red lines according to a dangerous capability measurement *and* decide on the capability threshold based on probability and severity of harm considerations. Additionally, using a compute threshold as an initial indicator of whether a foundation model might pose significant risks and should then be evaluated for dangerous capabilities can help make evaluations more targeted.

What Are the Advantages of Red Lines?

There are numerous advantages of red lines, as follows:

- **Proactive risk management.** Red lines enable stakeholders to proactively identify and mitigate potential large risks associated with foundation models. By establishing clear thresholds for dangerous capabilities, red lines serve as an early warning system that triggers necessary actions to prevent harm.
- **Clarity and consensus.** The process of defining red lines can be organized to encourage a comprehensive discussion among stakeholders—including governments, companies, and the scientific community—to foster a consensus on what constitutes unacceptable risks. This collective understanding helps in setting clear standards and expectations, reducing ambiguity in risk management.

- **Flexibility and adaptability.** Given the rapid evolution of AI technologies, red lines offer a flexible framework that can be updated or adjusted in response to new developments and insights. This adaptability ensures that risk management strategies remain relevant and effective over time.
- **Enhanced coordination and collaboration.** Red lines facilitate coordination among actors by providing a shared language and benchmarks for discussing AI risks. This enhances collaborative efforts to address complex challenges and ensures a consistent approach to mitigating potential threats.
- **Public trust and confidence.** By transparently defining and enforcing red lines, policymakers and technology developers can build public trust. Knowing that there are safeguards in place to prevent the misuse or unintended consequences of AI technologies reassures the public and stakeholders about the responsible development and use of foundation models.
- **Strategic guidance for research and development.** Red lines can guide researchers and developers in identifying safe and responsible pathways for AI innovation. By understanding the boundaries of acceptable risks, the AI community can focus on advancements that offer significant benefits while minimizing potential harms.
- **Prevention of regulatory overreach.** Establishing red lines allows for a more nuanced approach to regulation that avoids blanket bans or overly restrictive measures that could stifle innovation. By focusing on specific capabilities or outcomes, red lines ensure that regulations are targeted and proportionate to the actual risks.

What Are the Challenges and Limitations of Red Lines?

There are also numerous challenges and limitations to implementing red lines, including the following:

- **Complexity of AI systems.** The intricate nature of advanced AI systems, characterized by their opaque decision making processes and unpredictable behaviors, poses a significant challenge to defining and enforcing red lines. Unlike more straightforward engineering domains, the emergent properties of AI systems can make it difficult to anticipate all potentially dangerous capabilities.
- **Incentives for evasion and gaming.** The establishment of red lines may create incentives for AI developers to find ways to circumvent thresholds without technically violating them. Volkswagen famously designed a car to produce low emissions only when it was being evaluated by regulators.[26] Similarly, AI developers could design their systems to perform safely *only when being evaluated*, leading to AI systems that, while not crossing red lines *according to evaluations*, still pose significant risks. Those designing or conducting evaluations should aim to prevent such gaming by, for instance, keeping the details of some evaluations secret from AI developers.
- **Disputes of measurement validity.** The integrity of red lines can be significantly challenged by potential discrepancies in measurement methods and the subsequent disputes these differences can incite. Given the reliance on predefined results to signal the crossing of a red line, divergent measurement techniques can yield conflicting outcomes. Such inconsistencies may lead to catastrophic misinterpretations or undermine

[26] Jung and Sharon, "The Volkswagen Emissions Scandal and Its Aftermath."

collaborative efforts to manage risks, as motivated parties could exploit these variances to contest the accuracy of findings. This exploitation risks the erosion of trust and cooperation, particularly when actors, seeking to derive political or economic advantages, challenge measurement results after initially agreeing to red lines.

As they stand, red lines are a largely theoretical framework, and as is the case in many policy frameworks, there are likely to be large gaps between theory and implementation.

Question 7.h: "What insights from other countries or other societal systems are most useful to consider?"

In this section, we only discuss the European Union's approach to risk management for open foundation models. We would have liked to discuss how the People's Republic of China is approaching this issue but were unable to find sufficient literature outlining their open foundation model policies.

On March 13, 2024, the European Parliament passed the EU AI Act, a landmark piece of legislation governing AI systems in the EU.[27] It is widely believed that the EU AI Act will set the tone for much of the world's regulation of advanced AI systems, which will significantly affect AI companies in the United States that do business abroad, for better or worse.[28]

The EU AI Act contains many of the key features of a red-line approach, including governing via defined thresholds of risk, targeted risk management that exempts non-risky models, and a heavy emphasis on evidence from evaluations and risk assessments. We discuss some of these features here.

The EU AI Act categorizes AI model risks, with systemic risk being the highest. Systemic risks involve foundation models being misused for chemical, biological, nuclear, or cyber attacks, as well as models posing threats to democracy, human rights, critical infrastructure, and public safety, among other concerns.[29]

Open foundation models are generally exempt from the Act's requirements unless they are deemed to pose systemic risks.[30] All existing open foundation models are classified as not posing systemic risks. Foundation models (open or otherwise) are presumed to pose systemic risks if they required greater than 10^25 floating point operations to train, given that such large models are at the frontier of dangerous capabilities. However, the AI Office has the flexibility to adjust the threshold up or down, and it can designate specific models as posing systemic risks based on further criteria (such as the model's degree of autonomy).[31]

---

[27] Gilchrist and Iordache, "World's First Major Act to Regulate AI Passed by European Lawmakers."

[28] Siegmann and Anderljung, *The Brussels Effect and Artificial Intelligence*.

[29] Council of the European Union, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts."

[30] Future of Life Institute, "High-Level Summary of the AI Act."

[31] European Commission, "Artificial Intelligence – Questions and Answers."

Providers of foundation models that pose systemic risks must perform risk assessments and mitigate risks. The protocol for determining how evaluations should be performed or which steps to mitigate systemic risks are acceptable is not yet specified.[32]

As it stands, we have not seen how the EU AI Act will influence responsible AI development and innovation, and it contains many provisions not related to a red lines approach to AI risk management. However, observing the effects of the provisions of the EU AI Act related to a red lines approach which we outline here over the coming years may provide valuable insight into the implementation and effects of a red lines approach.

Question 7.G: "What should the U.S. prioritize in working with other countries on this topic, and which countries are most important to work with?"

Restricting the publication of open foundation model weights is an important potential policy lever to address the risk of dangerous capabilities falling into the hands of malicious actors, including states, nonstate actors, and individuals. However, attempts to control any open foundation models should consider which foreign actors could re-create equally capable models. Regardless of other issues, controls are only effective if actors that would misuse the controlled model do not already have access to an equivalent model or the ready ability to create one. Analogously, the 1954 Atomic Energy Act's controls on the proliferation of nuclear weapon information are only coherent because few other actors have that information, and they are similarly unwilling to proliferate it.

Models with the greatest potential for dangerous capabilities happen to be very expensive to develop. Initial evaluations in offensive cyber and deception capabilities show that the largest, most expensive, and generally most capable foundation models also possess the most dangerous capabilities.[33] The most capable models with widely accessible weights cost around $1.6 million to develop (Meta's Llama 2 model); however, this figure is dwarfed in comparison with the frontiers of AI research, where models such as GPT-4 and Gemini each likely took over $50 million to train.[34] Given the increasing scale of funding and talent needed, few countries globally contain actors that will be able to compete at the frontier of AI development (and, consequently, few actors will be able to create the most dangerous models). However, the number of actors globally that *can* compete is not zero, and some actors that could not build a model themselves might be able to steal one from U.S. AI developers.

---

[32] Council of the European Union, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts."

[33] Fang et al., "LLM Agents Can Autonomously Hack Websites"; Phuong et al., "Evaluating Frontier Models for Dangerous Capabilities."

[34] Epoch, "Database of Notable Machine Learning Systems"; Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over."

Coordinating with countries that could re-create (or otherwise acquire) frontier AI models will be critical to regulate the diffusion of dangerous AI capabilities. This coordination would be consistent with U.S. strategy on leading international standards development.[35] Several countries with access to capital, talent, and high-performance compute could create foundation models with equally dangerous capabilities as those in the United States. Additionally, countries with strong offensive cyber programs could steal model weights from U.S. developers. If any of these countries chose to make the weights of models with dangerous capabilities widely accessible (e.g., by publishing them online), it might be extremely difficult or impossible to contain or roll back the diffusion of those dangerous AI capabilities to all kinds of malicious actors. Therefore, coordinating with these countries on any model weight controls (in addition to improving AI company cybersecurity and securing the global high performance compute supply) will be critical to preventing the diffusion of dangerous AI capabilities.

International forums that could be used to facilitate coordination on controlling the diffusion of dangerous AI capabilities include the following:

1. **The G7.** The G7—consisting of Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States—has been meeting to coordinate principles and policies for managing risks from foundation models.[36]
2. **The International Scientific Report on Advanced AI Safety's Expert Advisory Panel.** Comprising scientists from 30 nations—including China and the United States—as well as the United Nations and EU, this panel will author a report summarizing the state of the science on the capabilities and risks of advanced AI systems.[37]
3. **AI Safety Summits.** The first AI Safety Summit was held by the United Kingdom in fall 2023 and resulted in the signing of the Bletchley Declaration by 29 countries, which committed to coordinate further on the safe development and deployment of foundation models.[38] The next AI Safety Summits will be held in South Korea and France in 2024.[39]
4. **The International Dialogues on AI Safety.** These dialogues bring together leading AI scientists from around the world to discuss extreme risks to humanity from AI. In March 2024, these scientists published a joint statement on red lines for dangerous AI capabilities.[40]

---

[35] White House, *United States Government National Standards Strategy for Critical and Emerging Technology*.

[36] White House, "G7 Leaders' Statement on the Hiroshima AI Process."

[37] United Kingdom Department for Science, Innovation, and Technology, "International Scientific Report on Advanced AI Safety: Expert Advisory Panel Members"; United Kingdom Department for Science, Innovation, and Technology, "International Scientific Report on Advanced AI Safety: Principles and Procedures."

[38] Government of the United Kingdom, "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023."

[39] "South Korea and France to Host Next Two AI Safety Summits."

[40] International Dialogues on AI Safety, homepage.

Summary

Although most existing open foundation models are highly beneficial and evaluations indicate that they currently pose no significant biological or cyber risks, future gains in the capabilities of foundation models could change the benefit-risk calculation on publishing model weights. The benefits and risks posed by open foundation models thus call for a nuanced, evidence-based policy approach that can track the actual risks posed by publishing the weights of particular models. Adopting and refining frameworks (such as the red-line approach), being informed by such examples as the EU AI Act, and engaging in international coordination provide promising options to promote the responsible advancement of open foundation models while minimizing their risks.

References

Anderljung, Markus, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury, "Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem Under the ASPIRE Framework," *arXiv*, November 15, 2023.

Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," *arXiv*, April 12, 2022.

Bommasani Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., "On the Opportunities and Risks of Foundation Models," *arXiv*, July 12, 2022.

Boulanger, Alan, "Open-Source Versus Proprietary Software: Is One More Reliable and Secure Than the Other?" *IBM Systems Journal*, Vol. 44, No. 2, 2005.

Bucknall, Benjamin S., and Robert F. Trager, *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements*, University of Oxford, October 2023.

Casper, Stephen, Carson Ezell, Charlotte Siegmann, Naom Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," *arXiv*, January 25, 2024.

Council of the European Union, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," Interinstitutional File 2021/0106(COD), January 26, 2024.

Epoch, "Database of Notable Machine Learning Systems," accessed on March 13, 2024. As of March 13, 2024:
https://epochai.org/data/epochdb/table

European Commission, "Artificial Intelligence—Questions and Answers," updated December 14, 2023. As of March 13, 2024:
https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683

Fang, Richard, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang, "LLM Agents Can Autonomously Hack Websites," *arXiv*, February 16, 2024.

Future of Life Institute, "High-Level Summary of the AI Act," February 27, 2024. As of March 13, 2024:
https://artificialintelligenceact.eu/high-level-summary/

Gade, Pranav, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2–Chat 13B," *arXiv*, October 31, 2023.

Gilchrist, Karen, and Ruxandra Iordache, "World's First Major Act to Regulate AI Passed by European Lawmakers," CNBC, March 13, 2024.

Government of the United Kingdom, "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," November 1, 2023. As of March 13, 2024:
https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Hazell, Julian, "Spear Phishing with Large Language Models," *arXiv*, December 22, 2023.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al., "Sleeper Agents: Training Deceptive LLMs That Persis Through Safety Training," *arXiv*, January 17, 2024.

International Dialogues on AI Safety, homepage, undated. As of March 22, 2024:
https://idais.ai

Jung, Jae C., and Elizabeth Sharon, "The Volkswagen Emissions Scandal and Its Aftermath," *Global Business and Organizational Excellence*, Vol. 38, No. 4, May/June 2019.

Kapoor, Sayash, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al., "On the Societal Impact of Open Foundation Models," *arXiv*, February 27, 2024.

Kaspersky, "What Is Zero-Click Malware, and How Do Zero-Click Attacks Work?" webpage, undated. As of March 25, 2024:
https://www.kaspersky.com/resource-center/definitions/what-is-zero-click-malware

Knight, Will, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, April 17, 2023.

Miller, Kyle, "Open Foundation Models: Implications of Contemporary Artificial Intelligence," Center for Security and Emerging Technology, March 12, 2024.

Model Evaluation and Threat Research, "Guidelines for Capability Elicitation," webpage, undated. As of March 26, 2024:
https://metr.github.io/autonomy-evals-guide/elicitation-protocol/

Mouton, Christopher A., Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*, RAND Corporation, RR-A2977-2, 2024. As of March 13, 2024: https://www.rand.org/pubs/research_reports/RRA2977-2.html

Musser, George, "How AI Knows Things No One Told It," *Scientific American*, May 11, 2023.

National Telecommunications and Information Administration, "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights," *Federal Register*, Vol. 89, No. 38, February 26, 2024.

NTIA—*See* National Telecommunications and Information Administration.

Patwardhan, Tejal, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn Jackson, et al., "Building an Early Warning System for LLM-Aided Biological Threat Creation," OpenAI, January 31, 2024. As of March 13, 2024: https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation

Phuong, Mary, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, et al., "Evaluation Frontier Models for Dangerous Capabilities," *arXiv*, March 20, 2024.

Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson, "Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend to!" *arXiv*, October 5, 2023.

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, et al., "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives," *arXiv*, September 29, 2023.

Shestov, Alexey, Rodion Levichev, Ravil Mussabayev, Evgeny Maslov, Anton Cheshkov, and Pavel Zadorozhny, "Finetuning Large Language Models for Vulnerability Detection," *arXiv*, March 1, 2024.

Shevlane, Toby, "Structured Access: An Emerging Paradigm for Safe AI Deployment," *arXiv*, April 11, 2022.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al., "Model Evaluation for Extreme Risks," *arXiv*, September 22, 2023.

Siegmann, Charlotte, and Markus Anderljung, *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*, Centre for the Governance of AI, August 2022.

"South Korea and France to Host Next Two AI Safety Summits," Reuters, November 1, 2023.

United Kingdom Department for Science, Innovation, and Technology, "International Scientific Report on Advanced AI Safety: Expert Advisory Panel Members," February 1, 2024. As of March 13, 2024: https://www.gov.uk/government/publications/international-scientific-report-on-advanced-ai-safety-expert-advisory-panel-and-principles-and-procedures/international-scientific-report-on-advanced-ai-safety-expert-advisory-panel-members

United Kingdom Department for Science, Innovation, and Technology, "International Scientific Report on Advanced AI Safety: Principles and Procedures," February 1, 2024. As of March 13, 2024: https://www.gov.uk/government/publications/international-scientific-report-on-advanced-ai-safety-expert-advisory-panel-and-principles-and-procedures/international-scientific-report-on-advanced-ai-safety-principles-and-procedures

Vipra, Jai, and Anton Korinek, "Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT," Brookings Institution, September 2023.

White House, "Executive Order 14110 of October 30, 2023: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," *Federal Register*, Vol. 88, No. 210, November 1, 2023.

White House, "G7 Leaders' Statement on the Hiroshima AI Process," October 20, 2023. As of March 13, 2024: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/g7-leaders-statement-on-the-hiroshima-ai-process/

White House, *United States Government National Standards Strategy for Critical and Emerging Technology*, May 2023.

Widder, David Gray, Sarah West, and Meredith Whittaker, "Open (for Business): Big Tech, Concentrated Power, and the Political Economy of Open AI," August 17, 2023.