Response to OSTP “National Priorities for Artificial Intelligence Request for Information”

The Center for a New American Security (CNAS) welcomes the opportunity to provide comments in response to OSTP’s “National Priorities for Artificial Intelligence Request for Information”. CNAS is an independent, bipartisan organization dedicated to developing bold, pragmatic, and principled national security solutions. The CNAS AI Safety & Stability project is a multi-year, multi-program effort that addresses the established and emerging risks associated with artificial intelligence.

Authors:
Timothy Fist, Fellow, Technology and National Security Program (tfist@cnas.org)
Michael Depp, Research Associate, AI Safety and Stability Project (mdepp@cnas.org)
Caleb Withers, Research Assistant, Technology and National Security Program (cwithers@cnas.org)

This document reflects the personal views of the authors alone. As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.
Contents

1. Introduction .......................................................................................................................... 3
2. Summary of recommendations ............................................................................................... 3
3. Defining frontier AI and its risks ............................................................................................ 4
   3.1. Emerging Risks from AI ............................................................................................... 5
       3.1.1. Weaponizable scientific research and manufacturing ........................................... 5
       3.1.2. Cyber operations .................................................................................................. 5
       3.1.3. Deception/persuasion operations ....................................................................... 6
4. Protecting rights, safety, and national security across the AI lifecycle .................................. 6
   4.1. Laying the foundations: technical research into safeguards .......................................... 6
   4.2. Safeguards in the hardware and computation supply chain ........................................... 7
   4.3. Safeguards in the model development process ............................................................. 8
       4.3.1. Pre-training risk assessments .............................................................................. 8
       4.3.2. Model evaluations .............................................................................................. 8
   4.4. Safeguards for model deployments ................................................................................. 9
       4.4.1. Pre-deployment risk assessment ...................................................................... 9
       4.4.2. Deployment strategy ........................................................................................... 9
   4.5. Further considerations for guardrails throughout the AI lifecycle ................................ 10
       4.5.1. Accounting for iteration in the AI lifecycle ....................................................... 10
       4.5.2. Internal audit functions .................................................................................... 10
       4.5.3. Cybersecurity .................................................................................................. 10
       4.5.4. Accounting for risks from transparency ........................................................... 11
5. Regulatory infrastructure ........................................................................................................ 11
   5.1. Potential roles for federal agencies .............................................................................. 11
   5.2. Licensing ..................................................................................................................... 12
   5.3. Liability ....................................................................................................................... 12
   5.4. Government procurement ............................................................................................ 12
6. Notes and references ............................................................................................................. 13
1. Introduction

This document outlines a set of policy measures for harnessing the benefits and mitigating the risks of AI. We focus on questions within the category “Protecting rights, safety, and national security”. We focus on “frontier” AI models—general-purpose “foundation” models at the frontier of research and development (R&D)—for several reasons. First, these models are likely to represent the forefront of AI capabilities and risks across many domains. Second, the development and deployment of these models is currently largely unregulated. Lastly, these models present a special set of challenges for policymakers that calls for dedicated guardrails on top of existing sector-specific regulation: they develop new capabilities in an unpredictable way, are hard to make reliably safe, and are likely to proliferate rapidly due to their multitude of possible uses. While the risks and challenges posed by frontier AI will often be novel, there are lessons to be learned from other sectors. For example, risk management for frontier AI will likely have a similar character to risk management in the cyber domain, which generally involves:

- A ‘security mindset’, expecting to confront unintended vulnerabilities and adversaries;
- Establishing organization-wide governance, risk management, and standards;
- Incident response capabilities;
- Integrating safeguards into the design process rather than implementing them as an afterthought;
- Promoting information sharing and collaboration to learn lessons and improve safety practices.

Our recommendations draw on the emerging literature of best practices and predicted risks for developing and deploying frontier AI models.

2. Summary of recommendations

Question 1: What specific measures—such as standards, regulations, investments, and improved trust and safety practices—are needed to ensure that AI systems are designed, developed, and deployed in a manner that protects people’s rights and safety? Which specific entities should develop and implement these measures?

As a foundation for safe and trustworthy AI, government bodies such as the National Science Foundation should provide active support and substantial funding for research into relevant technical solutions and evaluations—in particular, research specifically applicable to frontier systems and focused on ensuring their trustworthiness with high levels of confidence (as in the National Science Foundation’s recent Safe Learning-Enabled Systems solicitation). (see section 4.1)

Given the difficulty of preventing dangerous models from proliferating once they have been developed and/or released, the federal government should establish appropriate thresholds for “frontier models” and work with data center operators to ensure a transparent chain of frontier model provenance. This would involve tracking data centers capable of efficiently producing frontier models, requiring these data centers to implement a minimum set of cybersecurity standards, and requiring ‘Know Your Customer’-style diligence on users who are accessing large amounts of compute. (see section 4.2)

Question 3: Are there forms of voluntary or mandatory oversight of AI systems that would help mitigate risk? Can inspiration be drawn from analogous or instructive models of risk management in other sectors, such as laws and policies that promote oversight through registration, incentives, certification, or licensing?

Frontier AI labs should adopt the following best practices to mitigate risks throughout the frontier AI development and deployment life cycle:

- Pre-training and pre-deployment risk assessments and model evaluations; (see sections 4.3 and 4.4)
- Accounting for the risk profiles of different deployment strategies (e.g. staged release and structured access); (see section 4.4.2)
- Strong, independent internal audit functions; (see section 4.5.2)
- Strong cybersecurity practices, with regard to relevant standards. (see section 4.5.3)

Many of these practices can, and have been, voluntarily adopted by frontier labs. The federal government should encourage their adoption by requiring AI and cloud computing companies to adopt these best practices for frontier models as a condition of government contracts. (see section 5.4)

Congress should establish a regulatory body specifically focused on frontier AI. This body should have the authority to enforce minimum standards for the development and deployment of frontier AI systems. (see sections 5.1 and 5.2)

Congress should also consider whether current liability laws adequately account for the challenges of frontier AI systems—such as opaque internal logic, growing autonomy, the wide diffusion of powerful capabilities, and a yet-to-emerge consensus on what constitutes reasonable care in their development and deployment. (see section 5.3)

**Question 7: What are the national security risks associated with AI? What can be done to mitigate these risks?**

The government should focus attention on safety measures for highly capable, general-purpose foundation models and develop technical thresholds to categorize such systems based on whether they could pose meaningful risks to public safety. Three key areas where such systems are likely to pose national security risks in the future are dual-use science, cyber operations, and deception/persuasion operations. (see section 3.1)

**3. Defining frontier AI and its risks**

We use the term “Frontier AI” to refer to foundation models at the frontier of capabilities that are powerful enough to have a meaningful chance of causing severe harm to public safety and national security. Recent years have seen remarkable improvements in the capabilities of foundation models, following empirically derived ‘scaling laws’ which show that models steadily improve as they are made larger and trained using more computation. However, specific qualitative capabilities can emerge suddenly or improve dramatically as systems are scaled up, making it difficult to predict the exact nature and timeline of future benefits and dangers from AI systems. This necessitates proactive policymaking to prepare for potential capability enhancements in the most critical domains. We highlight three domains below where results from today’s most powerful AI models imply that future generations of these models could pose serious risks. However, the dual-use nature of such models makes for a thorny definitional challenge: how can regulators demarcate frontier models from those that clearly do not pose meaningful risks to public safety and national security?

At present, there does not exist a method for reliably establishing whether a planned model (based on its architecture, training data, training task, computational inputs, etc.) will have a given set of dangerous capabilities. One straightforward approach might be to use computational inputs (measured in floating point operations, or FLOP), which has historically been a decent proxy for the breadth and depth of model capabilities. This approach has the advantage of being determinable ahead of time based on a model’s planned inputs. Establishing such a threshold would also allow for the identification of data centers capable of efficiently training frontier models (i.e. data centers capable of producing the threshold amount of FLOP within a certain period of time or within a certain budget), which would provide policymakers with visibility of where high-risk models are likely to be produced, and establish more transparency over their provenance. This approach also has a key downside: research progress in efficient algorithms for training new AI models will decrease the computational inputs required to build models with dangerous capabilities as time passes.

This highlights three priorities for the federal government. First, to develop a near-term set of technical thresholds for categorizing highly-capable models that are liable to possess dangerous capabilities, in consultation
with technical AI experts developing these models. Second, to keep these thresholds flexible based on progress in AI R&D and the development of risk mitigations. Lastly, given the inevitable downsides of any threshold based on a technical proxy for capabilities, build a system of regulatory guardrails for developing and deploying powerful AI systems that use model-specific risk assessments and evaluations.

3.1. Emerging Risks from AI

Current AI systems demonstrate emerging capabilities that could have destabilizing effects on international relations and lower the barrier to entry for non-state actors to cause harm.

3.1.1. Weaponizable scientific research and manufacturing

Certain dual-use scientific research directions present both benefits and national security risks—for example, the development of chemical and biological compounds. The U.S. government accounts for these risks in policies and regulations, such as in Life Sciences Dual Use Research of Concern or ITAR restrictions on toxicological agents (along with algorithms and models supporting their design or deployment). However, these policies generally limit their focus to tools specifically designed for these domains, which could prove problematic as generative AI models increasingly demonstrate relevant capabilities. To date, key examples of how such tools could be used to cause harm include the following:

- In March 2022, researchers took MegaSyn, a generative AI tool for discovering therapeutic molecules, and ‘inverted’ it, tasking it with finding molecules that harm the nervous system. It found tens of thousands of candidate molecules, including known chemical weapons and novel molecules predicted to be as or more deadly.
- In April 2023, researchers gave a GPT-4 powered system access to the internet, code execution, hardware documentation, and remote control of an automated ‘cloud’ laboratory. It was “capable of autonomously designing, planning, and executing complex scientific experiments”, and in some cases, was willing to outline and execute viable methods for synthesizing illegal drugs and chemical weapons.

3.1.2. Cyber operations

Most software has dormant security vulnerabilities; finding them for offensive or defensive purposes is a matter of resourcing, time, effort, and skill. AI can automate relevant activities: detecting and responding to threats, finding and fixing vulnerabilities, and launching offensive attacks. These capabilities can support cyber defense and national security when used by the U.S. government and other responsible actors, but they can also cause significant harm when used by malicious actors. Large foundation models are already highly capable at coding and excel in data-rich domains—attributes which support their effectiveness in cyberspace. They can also be deployed at scale to target many systems at once and can be flexible, re-writing code on the fly to avoid detection and look for sensitive data. So far, there are several public examples of how these capabilities can be used:

- ‘Polymorphic’ malware is a type of malicious software that can change its code to avoid being detected. BlackMamba is a polymorphic keystroke logger that avoids detection by not having any hardcoded keylogging capabilities. Instead of having a fixed method of stealing information from users’ keyboards, it requests new keylogging code from ChatGPT every time it runs. This makes it harder to be stopped by security programs that look for specific segments or patterns of malicious code.
- DarwinGPT is a proof-of-concept virus that uses ChatGPT to modify its code and act autonomously. It has the goal of spreading itself to other systems. It can create new functions to help it achieve this goal without any specific instructions.
- Similarly, PentestGPT uses ChatGPT to automate penetration testing—the process of testing a system’s security by simulating an attack.
- ChatGPT can be used to generate plausible “spear phishing” emails, which target specific individuals with personalized messages that lure them to click on malicious links.
- GPT-4 can help identify vulnerabilities in computer code.
3.1.3. Deception/persuasion operations

Information operations, propaganda, and manipulation—and defense against them—have long played a critical role in international relations and national security, as state and non-state actors look to influence perceptions, behaviors, and decision-making processes at various levels. In the contemporary digital landscape, these operations can now be conducted at unprecedented scales and speeds. With AI systems increasingly able to generate human-like text and realistic media, this trend will be accelerated, as adversaries look to employ AI’s potential for increasingly bespoke and adaptable operations while retaining scale. Generative AI has demonstrated capabilities for deception and persuasion:

- Meta’s Cicero AI achieved human-level performance in Diplomacy, a strategy game centered around natural language negotiation and alliance building.\(^\text{12}\)
- An initial version of GPT-4 was tasked with getting a human worker on TaskRabbit to solve a CAPTCHA (an online test used to distinguish between humans and automated systems) for it. When challenged on whether it was a robot, it recognized it needed to come up with an excuse and pretended to be a visually impaired human.\(^\text{20}\)
- GPT-3 performed at human levels in crafting persuasive messages on contentious political issues.\(^\text{21}\)

Widely-available generative AI can also be used to fabricate multimedia such as deepfake images\(^\text{22}\) or speech mimicking someone’s voice\(^\text{24}\). The most capable models have advantages in generating compelling and personalized content:

- For image generation, the largest and most capable models perform better at incorporating “world knowledge, specific perspectives, or writing and symbol rendering”.\(^\text{24}\)
- Larger models are better able to model and tailor their answers to users’ preferences. Larger language models, especially when ‘fine-tuned’ to provide answers that humans prefer, tend to be more ‘sycophantic’, forgoing moderation and consistency to give answers that correspond to the user’s apparent political views.\(^\text{25}\)

4. Protecting rights, safety, and national security across the AI lifecycle

The U.S. government must work with industry to establish safeguards across the entire frontier AI lifecycle, from gathering data through to releasing a trained model. Safeguards are needed well before frontier models are released: if a trained model is released for download, it becomes highly difficult to track and regulate. A trained model is an easy-to-copy piece of software, with capable models presenting an attractive target for theft. Additionally, once someone has access to a downloaded model, it is possible to ‘fine-tune’ the model to elicit dangerous new capabilities with minimal additional computation relative to what was required for initial training. This poses a problem for oversight, as the fine-tuning process can involve the elicitation of dangerous capabilities that were previously not present—for example, if a conversational AI were fine-tuned to specialize in generating cyber-offensive code.

4.1. Laying the foundations: technical research into safeguards

As frontier AI models continue to rapidly advance, the engineering of reliable safeguards remains an open challenge: frontier models can currently be coaxed into providing dangerous information\(^\text{26}\), have largely inscrutable internal reasoning,\(^\text{28}\) and are prone to “hallucination” (making up facts).\(^\text{28}\) The *National AI R&D Strategic Plan* highlights the need for “further research… to enhance the validity[,] reliability[,] security and resilience of these large models”, and articulates the challenge of determining “what level of testing is sufficient to ensure the safety and security of non-deterministic and/or not fully explainable systems”.

Government bodies such as the National Science Foundation should continue to scale up their support for research into technical solutions and evaluations to ensure the trustworthiness of frontier AI systems. Given the unique challenges in ensuring the trustworthiness of frontier models\(^\text{29}\) and the inevitability that their weaknesses will be probed by adversaries, particular priority should be given to research that is specifically focused on
ensuring these models’ trustworthiness with high levels of confidence. This focus is exemplified by the National Science Foundation's recent Safe Learning-Enabled Systems solicitation.

Frontier AI labs themselves are major sources of relevant research and leading technical experts. In designing research solicitations, particular consideration should be given to how to incentivize labs to publish or conduct additional research into safeguards that they would not otherwise engage in—avoiding fungibility with already-commercially-incentivized research or the crowding out of existing safeguards research.

4.2. Safeguards in the hardware and computation supply chain

Frontier AI models are trained in large AI data centers, with current techniques requiring the use of thousands of specialized AI chips simultaneously, together with advanced networking equipment (switches and optical links) and dedicated power and cooling infrastructure. As long as this continues to be the method via which frontier models are produced, the federal government should treat specialized AI hardware and the data centers that house it as key targets for regulation. Oversight at this early state of the AI lifecycle is needed given the difficulty of preventing AI systems from proliferating once they have been developed and/or released.

Adherence to safety and security standards should be required of operators of data centers (typically owned by either AI developers themselves, or more commonly, cloud service providers) that are capable of training frontier models within competitive time-based thresholds for training. Specifically, categorization could work as follows:

1. Define a lower bound on the amount of training compute required to train a frontier AI model, measured in FLOP (floating point operations). Then, divide this number by an appropriate fraction that takes into account practically achievable amounts of parallelization among data centers. 32
2. Define a maximum training time to target. For example, if a model takes 3 years to train, it is unlikely to be at the frontier of capabilities by the time the training run has finished.
3. Treat all data centers that can provide the amount of FLOP outputted from (1) within the timeframe in (2) as targets for specific regulation.

Operators running such data centers should be required to provide information that can be used to track the provenance of AI systems that exceed yet-to-be-established “dangerous capability” thresholds (see section 3). 33 Such measures should include ‘Know Your Customer’-style diligence requirements for data center operators who make high-end computing for efficiently training frontier models available on the cloud and a minimum set of cybersecurity standards to prevent trained models from being stolen. 34

U.S. AI developers are not the only ones who will have the ability to develop potentially dangerous AI models in the future. If the U.S. passes domestic AI regulation, some fear that this will dampen U.S. competitiveness in AI while also not addressing another set of important risks: those emerging from the development and misuse of dangerous AI models by foreign, illicit actors or hostile militaries. Export controls could play an important complementary role to domestic regulations in supporting the implementation of AI safety and security measures while also making it difficult for the aforementioned actors to develop and/or deploy dangerous AI models.

Specifically, the federal government should consider applying export controls to cloud computing services and trained models that meet certain thresholds for frontier AI development. As in the case of encryption software, 35 export license approvals could be tied on a case-by-case basis to specific technical criteria, such as passing mandatory risk assessments and model evaluations. This would help prevent illicit actors or foreign militaries from accessing these capabilities, while also promoting model safety and security requirements. To make these measures more effective, the federal government should work with allied countries to multilateralize controls on cloud computing and trained models, and place end-use controls on advanced chips such that they cannot be exported to data center operators who do not comply with a minimum set of safety and security standards.
4.3. Safeguards in the model development process

4.3.1. Pre-training risk assessments

AI developers should undertake risk assessment and planning before training a new frontier model, because:

- Risks and mitigations around the partially trained model leaking need to be accounted for, as do risks from access to the model that will be provided during training (such as to crowdworkers to enable training on human feedback);
- Early decisions around training procedures, datasets, and technical safeguards can have long-lasting impacts on the final model that are hard or impossible to reverse;
- This allows developers to monitor how well they can predict the model’s performance and behavior in advance and ensure appropriate records are kept of model behavior and performance throughout the training process;
- This supports a proactive risk management culture, with the possibility of committing to certain mitigations in advance (such as pausing training in the case of novel, unexpected dangerous behaviors or capabilities), and reducing decision-making pressure from ‘sunk costs’.

At a minimum, these efforts should reflect the NIST AI Risk Management Framework and associated resources.\(^1\)

Pre-training risk assessments should also be informed by evaluations of similar, weaker models;\(^3\) OpenAI undertook such experimentation and forecasting in advance of training GPT-4. It is particularly important to understand how model capabilities and behaviors evolve as they approach the size of a frontier training run. Undesirable characteristics can sometimes worsen as models scale, a phenomenon known as ‘inverse scaling’. These trends are sometimes relatively consistent across model sizes, but can also exhibit non-linearities, such as occurring more drastically within certain model size ranges or within “U-shaped” scaling patterns that trend in different directions at different scales.\(^3\)

4.3.2. Model evaluations

After training but before frontier models are deployed, developers should comprehensively evaluate the model to help ensure that they are not inadvertently deploying dangerous capabilities. In addition to more general evaluations (such as for accuracy and bias), developers should evaluate frontier models’ ability to cause large-scale harm (such as through cyber-offense, weapons development, or deception), as well as their propensity to do so. Ideally, model evaluations would be comprehensive, interpretable and safe, per the below table.

<table>
<thead>
<tr>
<th>Comprehensive:</th>
<th>Interpretable:</th>
<th>Safe:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Cover as many plausible extreme threat models as possible.</td>
<td>- Include evaluations that present risks in an accessible way.</td>
<td>- Ensure evaluations are safe to implement: not introducing unacceptable levels of risk themselves.</td>
</tr>
<tr>
<td>- Take advantage of automated and human-assisted evaluations.</td>
<td>- Cover wide ranges of difficulty so trends can be tracked over time.</td>
<td></td>
</tr>
<tr>
<td>- Look at both a model’s behavior and how it produced that behavior.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Use adversarial testing to purposefully search for cases where models produce concerning results.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Pursue robustness against deliberate model deception to pass evaluations.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Surface latent capabilities through practices such as prompt engineering and fine-tuning.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Conduct evaluations throughout the model lifecycle.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Study models both with and without relevant system integrations (such as external tools or classifiers).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model evaluation should involve scientific and policy experts in domains such as biosecurity and cyber operations, to help contextualize the risk posed to society by given capabilities. To this end, federal agencies with relevant expertise should consider collaborating with frontier AI evaluators and standards bodies—for example, CISA and/or NSA could offer guidance on the types of cyber capabilities that could pose serious risks to national security if available in released models.

**Mitigating evaluation gaming**

In implementing and interpreting model evaluations, risks from evaluations being gamed should be accounted for (this could be intentional or inadvertent on the part of the developer). If labs optimize for demonstrating safety through particular evaluations, these evaluations may lose validity or usefulness as indicators. Measures to reduce this risk can include:

- Avoiding directly training models to pass evaluations;
  - Using *canary strings*: some benchmarking and evaluation projects include unique identifying text called ‘canary strings’ in associated documents—these identifiers help developers exclude these documents from training datasets (and help evaluators detect if they have been included);
- Keeping some details of some evaluations non-public;
- Research into methods for identifying if a model is ‘gaming’ evaluations and tests.

**Involving external evaluators**

Given AI labs’ conflict of interest in evaluating their own models, third parties should also be involved. Evidence from other domains suggests that third-party evaluators should:

- have well-scoped objectives;
- be professionalized—for example through training, standardized methodologies, and third-party accreditation;
- have a high degree of access to models—direct access to model outputs at a minimum, and potentially information about training data and model architecture, and the ability to fine-tune models;
- maintain independence—for example, avoiding compensation schemes that are tied to evaluation results, or cross-selling of other services.

Involving a more diverse range of evaluators can also help to identify risks more comprehensively. Frontier AI labs can demonstrate responsibility by taking these factors into account as they determine the involvement of external evaluators. Since the frontier model evaluation ecosystem is still emerging, the federal government should consider supporting the development of relevant standards and an accreditation body.

### 4.4. Safeguards for model deployments

#### 4.4.1. Pre-deployment risk assessment

Frontier AI developers should undertake risk assessments before deployment. In addition to supporting internal decision-making, these could be shared with external evaluators and researchers, regulators, or publicly. Risk assessments should include evaluation results, alongside the developer’s justification for why the deployment is safe given those results.” At a minimum, these assessments should also reflect the NIST *AI Risk Management Framework* and associated resources. Pre-deployment risk assessments should also account for the capabilities of systems that are already publicly available (including the cost and ease of employing these capabilities): if systems are already widely available with a particular capability/risk profile, the risk from additional systems with this same profile could be marginal.

#### 4.4.2. Deployment strategy

Frontier AI labs should account for the risk profile of their deployment strategy. In particular, they can reduce risks by providing users with *structured access* to models through web interfaces and APIs, instead of allowing them to download the trained AI models directly. This:

- allows labs to implement ‘Know Your Customer’ policies;
● allows labs to monitor for misuse of their models;
● allows restriction of access to the models if concerning vulnerabilities, capabilities, or synergies with external tools are discovered post-deployment;
● prevents modification of the underlying model, such as through fine-tuning, to remove safeguards or introduce new dangerous capabilities;
● helps protect intellectual property, such as innovations in model architecture, which could otherwise help strategic competitors of the US to ‘catch up’ in frontier AI capabilities.

Risks can also be reduced by staging the release of models: first testing them with relatively small and trusted groups of users before making them available more widely. This can help reduce the chance that unanticipated vulnerabilities are exploited by bad actors or cause harm at scale.

Where frontier labs do pursue more open release strategies, this should be accounted for in their risk assessments. For instance, risks from malicious actors become greater if a model is widely released, and risks from unattributable use or fine-tuning become greater if a model is open-sourced.

4.5. Further considerations for guardrails throughout the AI lifecycle

4.5.1. Accounting for iteration in the AI lifecycle

Though this document includes references to discrete phases such as training and deployment, in practice the lifecycle of frontier AI systems will often be fluid, iterative, or overlapping. For example, the latter stages of training frontier models will often involve some external availability, like making models available to crowdworkers to enable training on human feedback. Additionally, systems may be updated frequently or in real-time based on user interaction following their deployment. For AI systems that are subject to continuous refinement, new rounds of risk assessments could be triggered following significant additional training computation and/or improvement on performance benchmarks.

Additionally, risk assessments may need to be revisited after models have been deployed more widely: users may uncover new capabilities or vulnerabilities, or new tools and datasets may become available to the system. Just as the FDA conducts postmarketing surveillance of medical products after they have been released on the market, frontier AI models will need similar efforts to track and mitigate risks that become apparent after deployment.

4.5.2. Internal audit functions

Frontier AI labs should have internal audit functions that are specific to the risks from these systems and accounting for the full AI lifecycle. The value of internal audit has been recognized in, for example, requirements for public companies through the Sarbanes-Oxley Act or federal agencies per the Government Accountability Office’s ‘Green Book’. Functioning with a degree of independence, internal audit functions should evaluate an organization’s risk profile and report to the executive team. They should understand risk management literature and field-specific standards, especially regarding AI. The mandate for these boards should ensure they have sufficient authority, including access to models and associated documents, to conduct their evaluations effectively.

4.5.3. Cybersecurity

Frontier labs should put significant effort into cybersecurity. As economically valuable pieces of software that can be easily shared and replicated, frontier AI models are likely to be targets for theft and hacking.

At the very least, labs developing frontier models should adhere to basic cybersecurity standards such as the NIST Cybersecurity Framework or ISO 27001. Cybersecurity should also be considered in the implementation of efforts to mitigate risks from frontier models: for example, where regulators or evaluators require access to frontier models, secure onsite access will generally be preferable to distributing models over the internet.
4.5.4. Accounting for risks from transparency

In general, transparency from frontier AI labs around their models can help support accountability and trust. However, it should be noted that:

- Full transparency around data sources, training methods, and model architecture (let alone model weights themselves) would often involve the disclosure of proprietary secrets that could help strategic competitors of the US to ‘catch up’ in frontier AI capabilities;
- Full transparency around model evaluations and capabilities could draw attention to and support the development of dangerous capabilities by bad actors and strategic competitors, or intensify security dilemma dynamics;
- Full transparency around how model evaluations are conducted could allow them to be ‘gamed’ by developers.\textsuperscript{22}

In some cases, only regulators or external evaluators should have full access to these details (with appropriate commitments to confidentiality), with the public instead receiving high-level conclusions or summaries.

5. Regulatory infrastructure

5.1. Potential roles for federal agencies

The regulatory challenge posed by frontier AI models would benefit from a dedicated regulatory body: to help ensure clear accountability, responsiveness to the rapid pace of technological advancements, a critical mass of technical expertise, and holistic oversight of the entire AI lifecycle for these models. Congress should establish such a body.

The most important reason for a dedicated regulatory approach is the general-purpose nature of frontier AI: any sector-specific regulatory bodies would inevitably overlap significantly with each other, given they would need to oversee the development and deployment processes of the same general-purpose models. A dedicated regulatory body should have sufficient resourcing, expertise and information access powers to oversee labs developing frontier AI systems and the associated accountability ecosystem, and have the remit to focus especially on societal-scale risks. Although this body would have overarching responsibility, it should not preclude interagency collaboration, or the possibility of assigning specific support or enforcement roles to other organizations.

In considering agency roles around regulating frontier AI models, several factors should be taken into account. These may include, but are not limited to:

- The overlap between regulating advanced AI models and the Department of Commerce’s role in controlling the exports of high-tech AI chips used for their training;
- NIST’s role in developing and promoting relevant standards;
- The Department of Energy’s experience in handling scientific advancements with national security implications, such as advanced computing;
- The support the Intelligence Community can provide in assessing whether releasing model architectures or weights would aid the AI initiatives of potential rivals, and in identifying misuse of frontier AI systems.

Enhanced regulatory focus on frontier AI models should not be at the expense of existing regulators overseeing relevant use of AI systems—frontier or otherwise. As emphasized in a recent statement by the Consumer Financial Protection Bureau, Justice Department’s Civil Rights Division, Equal Employment Opportunity Commission, and Federal Trade Commission, “Existing legal authorities apply to the use of automated systems and innovative new technologies just as they apply to other practices.”\textsuperscript{31} The National AI Advisory Committee’s \textit{Year 1 Report} includes a number of recommendations that would support the U.S. Government’s overall support of AI accountability.
5.2. Licensing

Licensing requirements are common in high-risk industries, such as aviation and drug manufacturing. A federal regulatory body should license the deployment of frontier AI, given the potential future risks that these systems could pose, and the fact that such systems are likely to proliferate widely in a hard-to-control fashion if released for download. Model development would also ideally be included in a licensing regime: “model deployment” is not well-defined in many contexts, and models (as pieces of software) may proliferate through theft or internal leaks once developed.

The most straightforward licensing regime to create and enforce for AI development would require developers planning to use more than a particular amount of compute during a training run to first obtain a license to do so. Technical thresholds (compute-based or otherwise) for requiring a license should be set so as to only apply to systems that could pose meaningful risks to public safety, as described in section 3. The amount of compute triggering licensing requirements could change over time, slowly increasing in line with AI capabilities and safety mitigations, to ensure that it does not stifle innovation from smaller companies or those pursuing capabilities that are less dangerous.

Licenses themselves could be granted based on the developer demonstrating compliance with a particular set of safety standards, such as conducting thorough risk assessments and engaging in external evaluations (see sections 4.3 and 4.4). The establishment of these standards should be initiated and sustained by policymakers, but driven by technical experts, including AI developers, academic AI researchers, and AI safety & ethics experts. NIST’s AI Risk Management Framework is a promising starting point for these efforts.

5.3. Liability

Liability rules can help incentivize those developing and deploying AI systems to voluntarily implement appropriate risk management practices. While mandatory requirements will also be needed to adequately address severe risks and account for the difficulty of comprehensively identifying downstream harms, liability can help ensure AI labs take full advantage of their technical expertise and implement safeguards that may otherwise take time to be accounted for by regulators. Given potential challenges from frontier AI systems—opaque internal logic, an expanding ecosystem of autonomous decision-making by internet-connected models, wide availability to users of varying means to compensate any victims of AI-fueled harms, and yet-to-emerge consensus on what constitutes reasonable care—Congress should consider whether current liability laws are fit-for-purpose when applied to frontier AI models. Liability models that could help address these challenges could include vicarious liability, strict liability, joint and several liability, or industry-funded compensation funds.

5.4. Government procurement

Procurement rules are a promising avenue to encourage the adoption of best practices in frontier AI labs; the federal government should set minimum standards for frontier AI labs receiving government contracts. As a likely significant customer of these advanced AI systems, the U.S. government has a vested interest in ensuring their trustworthiness and mitigating potential risks. Because of the powerful incentive of government contracts, U.S. government requirements are likely to see broad adoption among leading frontier labs, providing a potential middle ground between voluntary and mandatory standards. While specific agencies may have their own unique requirements, the U.S. government should generally strive for standardized rules regarding frontier AI procurement across agencies, such as through Federal Acquisition Regulations.
6. Notes and references

1. Foundation models are models that are “trained on broad data at scale and are adaptable to a wide range of downstream tasks,” as per Bommasani et al., “On the Opportunities and Risks of Foundation Models,” *arXiv* (2021), https://arxiv.org/abs/2108.07258. For further discussion of the regulatory challenge posed by frontier foundation models, see O’Keefe et al., “Frontier AI Regulation: Managing Emerging Risks to Public Safety,” forthcoming.


3. O’Keefe et al., forthcoming

4. The empirical observation that models that leverage the most computation are the most capable is sometimes known as “The Bitter Lesson”, a term popularized by the AI research scientist Rich Sutton: “The Bitter Lesson,” (Incomplete Ideas, March 13 2019, http://www.incompleteideas.net/InCompleteIdeas/BitterLesson.html). This observation has been characterized in “scaling laws”, which describe how model capabilities scale with training inputs: Villalobos, “Scaling Laws Literature Review,” Jan 26 2023, https://epochai.org/blog/scaling-laws-literature-review.


7. As of June 2023, a threshold for capturing the next generation of foundation models would be around 10^27 bit operations (using the definition for operations adopted by the Bureau of Industry and Security in export controls: “Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification; Updates to the Controls To Add Macau.” Federal Register, Jan 18 2023, https://www.federalregister.gov/documents/2023/01/18/2023-00888/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor). Such a threshold at present would capture only a handful of top AI developers with the large budget necessary to train such a model (around $100 million).


For an example, see Bernhard Mueller (@muellerberndt), “I gave #GPT access to a bunch of hacking tools. This is PentestGPT autonomously attacking a Metasploitable VM,” Twitter, April 8, 2023, https://twitter.com/muellerberndt/status/164457189065111425


Technical challenges and associated research directions around safeguards for frontier systems are discussed in Hendrycks et al. 2023


Unique challenges to trustworthiness and safety from frontier systems may include:

- Training on intermediate tasks distinct from models’ ultimate applications, such as large language models trained to predict unstructured text (Bommasani et al., 2021, section 4.4.2);
- Use of systems far beyond their training distribution, given their ability to perform novel tasks with few or zero examples (Bommasani et al., 2021, section 4.8.2);

Burns et al., “Discovering Latent Knowledge in Language Models Without Supervision,” arXiv (2022), https://arxiv.org/abs/2212.03827 is an example of a notable research paper in this vein. In this paper, researchers analyzed how large language models internally represent knowledge as they answer true/false questions. In doing so, the researchers gained insight into what the models seemed to "believe", independently of the models' actual responses. Notably, the researchers were able to elicit higher accuracy on true/false answers in doing so. Further exploring how models represent knowledge internally can help identify when they exhibit inconsistent or misleading behavior. For further commentary on this paper, see Scheffler, “A new AI lie detector reveals their ‘inner thoughts’,” Freethink, March 20 2023, https://www.freethink.com/robots-ai/ai-lie-detector and “How "Discovering Latent Knowledge in Language Models Without Supervision" Fits Into a Broader Alignment Scheme,” Alignment Forum, December 15 2022, https://www.alignmentforum.org/posts/L4anhrxjv8j2yRKKp/how-discovering-latent-knowledge-in-language-models-without.

For an overview of this form of parallelism ("data parallelism"), see Weng, “How to Train Really Large Models on Many GPUs?” Lil’Log, September 24 2021, https://lilianweng.github.io/posts/2021-09-25-train-large/.

One method for tracking model provenance is described in Shavit, “What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring,” arXiv (2023), https://arxiv.org/abs/2303.11341. Though the details of an appropriate scheme are yet to be developed, we believe that technical measures for verifying model provenance using computing resources are a promising form of assurance for addressing dangerous AI proliferation.

‘Know Your Customer’ requirements have been passed in Executive Order 13984, which seeks to protect the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities, Federal Register, 25 Jan. 2021, https://www.federalregister.gov/documents/2021/01/25/2021-01714/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious/. We would propose a less ambitious version of the same approach, focused instead on the high-end compute resources used in frontier AI development. Such an approach would apply to only a tiny proportion (far less than 1%) of cloud users.

For example, a threshold for the control of trained models could be based on training compute (FLOP), and a threshold for cloud computing could be based on whether a configuration of chips were being made available that exceeded a given FLOP per second (FLOP/s) threshold for large model training.


Ngo, Chan & Mindermann, 2022


ARC Evals, “Response to Request for Comments on AI Accountability Policy” (2023), https://www.regulations.gov/comment/NTIA-2023-0005-1442


Shevlane et al., 2023

For proposals on how the Framework could be more tailored and actionable for high-consequence systems, see Barrett et al., 2023.


Current model safeguards can generally be fine-tuned away if someone has access to an AI model’s weights. For instance, Meta released the weights of their “LLaMA” family of large language models; users have started fine-tuning LLaMA models to emulate datasets of chatbots following instructions, except with any examples of chatbots not complying due to safeguards excluded (Eric Hartford, “WizardLM_alpaca_evol_instruct_70k_unfiltered,” Hugging Face (2023), https://huggingface.co/datasets/ehartford/WizardLM_alpaca_evol_instruct_70k_unfiltered). In doing so, they have managed to create so-called ‘uncensored models’ (we verified that, when prompted, these models would generate ideas for terrorist attacks) with similar capabilities to their original counterparts. The cost of such fine-tuning can be as low as hundreds of dollars (Taori et al. “Alpaca: A Strong, Replicable Instruction-Following Model,” Stanford Center for Research on Foundation Models, 2023 https://crfm.stanford.edu/2023/03/13/alpaca.html).


Shevlane et al., 2023


Shevlane et al., 2023


For example, models could be used “behind the scenes” at companies to generate new products or research.
