FRONTIER AI REGULATION:
MANAGING EMERGING RISKS TO PUBLIC SAFETY

Markus Anderljung[1,2*†], Joslyn Barnhart[3**], Anton Korinek[4,5,1**†], Jade Leung[6*], Cullen O'Keefe[6*],
Jess Whittlestone[7**], Shahar Avin[8], Miles Brundage[6], Justin Bullock[9,10], Duncan Cass-Beggs[11],
Ben Chang[12], Tantum Collins[13,14], Tim Fist[2], Gillian Hadfield[15,16,17,6], Alan Hayes[18], Lewis Ho[3],
Sara Hooker[19], Eric Horvitz[20], Noam Kolt[15], Jonas Schuett[1], Yonadav Shavit[14***],
Divya Siddarth[21], Robert Trager[1,22], Kevin Wolf[18]

[1]Centre for the Governance of AI, [2]Center for a New American Security, [3]Google DeepMind,
[4]Brookings Institution, [5]University of Virginia, [6]OpenAI, [7]Centre for Long-Term Resilience, [8]Centre for the
Study of Existential Risk, University of Cambridge, [9]University of Washington, [10]Convergence Analysis,
[11]Centre for International Governance Innovation, [12]The Andrew W. Marshall Foundation,
[13]GETTING-Plurality Network, Edmond & Lily Safra Center for Ethics, [14]Harvard University,
[15]University of Toronto, [16]Schwartz Reisman Institute for Technology and Society, [17]Vector Institute,
[18]Akin Gump Strauss Hauer & Feld LLP, [19]Cohere For AI, [20]Microsoft, [21]Collective Intelligence Project,
[22]University of California: Los Angeles

Listed authors contributed substantive ideas and/or work to the white paper. Contributions include writing, editing, research, detailed feedback, and participation in a workshop on a draft of the paper. The first six authors are listed in alphabetical order, as are the subsequent 18. Given the size of the group, inclusion as an author does not entail endorsement of all claims in the paper, nor does inclusion entail an endorsement on the part of any individual's organization.

*Significant contribution, including writing, research, convening, and setting the direction of the paper.

**Significant contribution including editing, convening, detailed input, and setting the direction of the paper.

***Work done while an independent contractor for OpenAI.

†Corresponding authors. Markus Anderljung (`markus.anderljung@governance.ai`) and Anton Korinek (`akorinek@brookings.edu`).

Cite as "Frontier AI Regulation: Managing Emerging Risks to Public Safety." Anderljung, Barnhart, Korinek, Leung, O'Keefe, & Whittlestone, et al, 2023.

ABSTRACT

Advanced AI models hold the promise of tremendous benefits for humanity, but society needs to proactively manage the accompanying risks. In this paper, we focus on what we term "frontier AI" models — highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety. Frontier AI models pose a distinct regulatory challenge: dangerous capabilities can arise unexpectedly; it is difficult to robustly prevent a deployed model from being misused; and, it is difficult to stop a model's capabilities from proliferating broadly. To address these challenges, at least three building blocks for the regulation of frontier models are needed: (1) standard-setting processes to identify appropriate requirements for frontier AI developers, (2) registration and reporting requirements to provide regulators with visibility into frontier AI development processes, and (3) mechanisms to ensure compliance with safety standards for the development and deployment of frontier AI models. Industry self-regulation is an important first step. However, wider societal discussions and government intervention will be needed to create standards and to ensure compliance with them. We consider several options to this end, including granting enforcement powers to supervisory authorities and licensure regimes for frontier AI models. Finally, we propose an initial set of safety standards. These include conducting pre-deployment risk assessments; external scrutiny of model behavior; using risk assessments to inform deployment decisions; and monitoring and responding to new information about model capabilities and uses post-deployment. We hope this discussion contributes to the broader conversation on how to balance public safety risks and innovation benefits from advances at the frontier of AI development.

Executive Summary

The capabilities of today's foundation models highlight both the promise and risks of rapid advances in AI. These models have demonstrated significant potential to benefit people in a wide range of fields, including education, medicine, and scientific research. At the same time, the risks posed by present-day models, coupled with forecasts of future AI progress, have rightfully stimulated calls for increased oversight and governance of AI across a range of policy issues. We focus on one such issue: the possibility that, as capabilities continue to advance, new foundation models could pose severe risks to public safety, be it via misuse or accident. Although there is ongoing debate about the nature and scope of these risks, we expect that government involvement will be required to ensure that such "frontier AI models" are harnessed in the public interest.

Three factors suggest that frontier AI development may be in need of targeted regulation: (1) Models may possess unexpected and difficult-to-detect dangerous capabilities; (2) Models deployed for broad use can be difficult to reliably control and to prevent from being used to cause harm; (3) Models may proliferate rapidly, enabling circumvention of safeguards.

Self-regulation is unlikely to provide sufficient protection against the risks from frontier AI models: government intervention will be needed. We explore options for such intervention. These include:

> **Mechanisms to create and update safety standards** for responsible frontier AI development and deployment. These should be developed via multi-stakeholder processes, and could include standards relevant to foundation models overall, not exclusive to frontier AI. These processes should facilitate rapid iteration to keep pace with the technology.

> **Mechanisms to give regulators visibility** into frontier AI development, such as disclosure regimes, monitoring processes, and whistleblower protections. These equip regulators with the information needed to address the appropriate regulatory targets and design effective tools for governing frontier AI. The information provided would pertain to qualifying frontier AI development processes, models, and applications.

> **Mechanisms to ensure compliance with safety standards.** Self-regulatory efforts, such as voluntary certification, may go some way toward ensuring compliance with safety standards by frontier AI model developers. However, this seems likely to be insufficient without government intervention, for example by empowering a supervisory authority to identify and sanction non-compliance; or by licensing the deployment and potentially the development of frontier AI. Designing these regimes to be well-balanced is a difficult challenge; we should be sensitive to the risks of overregulation and stymieing innovation on the one hand, and moving too slowly relative to the pace of AI progress on the other.

Next, we describe an initial set of safety standards that, if adopted, would provide some guardrails on the development and deployment of frontier AI models. Versions of these could also be adopted for current AI models to guard against a range of risks. We suggest that at minimum, safety standards for frontier AI development should include:

> **Conducting thorough risk assessments informed by evaluations of dangerous capabilities and controllability.** This would reduce the risk that deployed models possess unknown dangerous capabilities, or behave unpredictably and unreliably.

> **Engaging external experts to apply independent scrutiny to models.** External scrutiny of the safety and risk profile of models would both improve assessment rigor and foster accountability to the public interest.

Following standardized protocols for how frontier AI models can be deployed based on their assessed risk. The results from risk assessments should determine whether and how the model is deployed, and what safeguards are put in place. This could range from deploying the model without restriction to not deploying it at all. In many cases, an intermediate option—deployment with appropriate safeguards (e.g., more post-training that makes the model more likely to avoid risky instructions)—may be appropriate.

Monitoring and responding to new information on model capabilities. The assessed risk of deployed frontier AI models may change over time due to new information, and new post-deployment enhancement techniques. If significant information on model capabilities is discovered post-deployment, risk assessments should be repeated, and deployment safeguards updated.

Going forward, frontier AI models seem likely to warrant safety standards more stringent than those imposed on most other AI models, given the prospective risks they pose. Examples of such standards include: avoiding large jumps in capabilities between model generations; adopting state-of-the-art alignment techniques; and conducting pre-training risk assessments. Such practices are nascent today, and need further development.

The regulation of frontier AI should only be one part of a broader policy portfolio, addressing the wide range of risks and harms from AI, as well as AI's benefits. Risks posed by current AI systems should be urgently addressed; frontier AI regulation would aim to complement and bolster these efforts, targeting a particular subset of resource-intensive AI efforts. While we remain uncertain about many aspects of the ideas in this paper, we hope it can contribute to a more informed and concrete discussion of how to better govern the risks of advanced AI systems while enabling the benefits of innovation to society.

Acknowledgements

Contents

1 Introduction

Responsible AI innovation can provide extraordinary benefits to society, such as delivering medical [1, 2, 3, 4] and legal [5, 6, 7] services to more people at lower cost, enabling scalable personalized education [8], and contributing solutions to pressing global challenges like climate change [9, 10, 11, 12] and pandemic prevention [13, 14]. However, guardrails are necessary to prevent the pursuit of innovation from imposing excessive negative externalities on society. There is increasing recognition that government oversight is needed to ensure AI development is carried out responsibly; we hope to contribute to this conversation by exploring regulatory approaches to this end.

In this paper, we focus specifically on the regulation of frontier AI models, which we define as highly capable foundation models[1] that could have dangerous capabilities sufficient to pose severe risks to public safety and global security. Examples of such dangerous capabilities include designing new biochemical weapons [16], producing highly persuasive personalized disinformation, and evading human control [17, 18, 19, 20, 21, 22, 23].

In this paper, we first define frontier AI models and detail several policy challenges posed by them. We explain why effective governance of frontier AI models requires intervention throughout the models' lifecycle, at the development, deployment, and post-deployment stages. Then, we describe approaches to regulating frontier AI models, including building blocks of regulation such as the development of safety standards, increased regulatory visibility, and ensuring compliance with safety standards. We also propose a set of initial safety standards for frontier AI development and deployment. We close by highlighting uncertainties and limitations for further exploration.

---

[1]Defined as: "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" [15].

2 The Regulatory Challenge of Frontier AI Models

2.1 What do we mean by frontier AI models?

For the purposes of this paper, we define "frontier AI models" as highly capable foundation models[2] that could exhibit dangerous capabilities. Such harms could take the form of significant physical harm or the disruption of key societal functions on a global scale, resulting from intentional misuse or accident [25, 26]. It would be prudent to assume that next-generation foundation models could possess advanced enough capabilities to qualify as frontier AI models, given both the difficulty of predicting when sufficiently dangerous capabilities will arise and the already significant capabilities of today's models.

Though it is not clear where the line for "sufficiently dangerous capabilities" should be drawn, examples could include:

- Allowing a non-expert to design and synthesize new biological or chemical weapons.[3]

- Producing and propagating highly persuasive, individually tailored, multi-modal disinformation with minimal user instruction.[4]

- Harnessing unprecedented offensive cyber capabilities that could cause catastrophic harm.[5]

- Evading human control through means of deception and obfuscation.[6]

This list represents just a few salient possibilities; the possible future capabilities of frontier AI models remains an important area of inquiry.

Foundation models, such as large language models (LLMs), are trained on large, broad corpora of natural language and other text (e.g., computer code), usually starting with the simple objective of predicting the next "token".[7] This relatively simple approach produces models with surprisingly broad capabilities.[8] These

---

[2][15] defines "foundation models" as "models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks." See also [24].

[3]Such capabilities are starting to emerge. For example, a group of researchers tasked a narrow drug-discovery system to identify maximally toxic molecules. The system identified over 40,000 candidate molecules, including both known chemical weapons and novel molecules that were predicted to be as or more deadly [16]. Other researchers are warning that LLMs can be used to aid in discovery and synthesis of compounds. One group attempted to create an LLM-based agent, giving it access to the internet, code execution abilities, hardware documentation, and remote control of an automated 'cloud' laboratory. They report finding that it in some cases the model was willing to outline and execute on viable methods for synthesizing illegal drugs and chemical weapons [27].

[4]Generative AI models may already be useful to generate material for disinformation campaigns [28, 29, 30]. It is possible that, in the future, models could possess additional capabilities that could enhance the persuasiveness or dissemination of disinformation, such as by making such disinformation more dynamic, personalized, and multimodal; or by autonomously disseminating such disinformation through channels that enhance its persuasive value, such as traditional media.

[5]AI systems are already helpful in writing and debugging code, capabilities that can also be applied to software vulnerability discovery. There is potential for significant harm via automation of vulnerability discovery and exploitation. However, vulnerability discovery could ultimately benefit cyberdefense more than -offense, provided defenders are able to use such tools to identify and patch vulnerabilities more effectively than attackers can find and exploit them [31, 32].

[6]If future AI systems develop the ability and the propensity to deceive their users, controlling their behavior could be extremely challenging. Though it is unclear whether models will trend in that direction, it seems rash to dismiss the possibility and some argue that it might be the default outcome of current training paradigms [17, 18, 20, 21, 22, 23].

[7]A token can be thought of as a word or part of a word [33].

[8]For example, LLMs achieve state-of-the-art performance in diverse tasks such as question answering, translation, multi-step reasoning, summarization, and code completion, among others [34, 35, 36, 37]. Indeed, the term "LLM" is already becoming outdated, as several leading "LLMs" are in fact multimodal (e.g., possess visual capabilities) [36, 38].
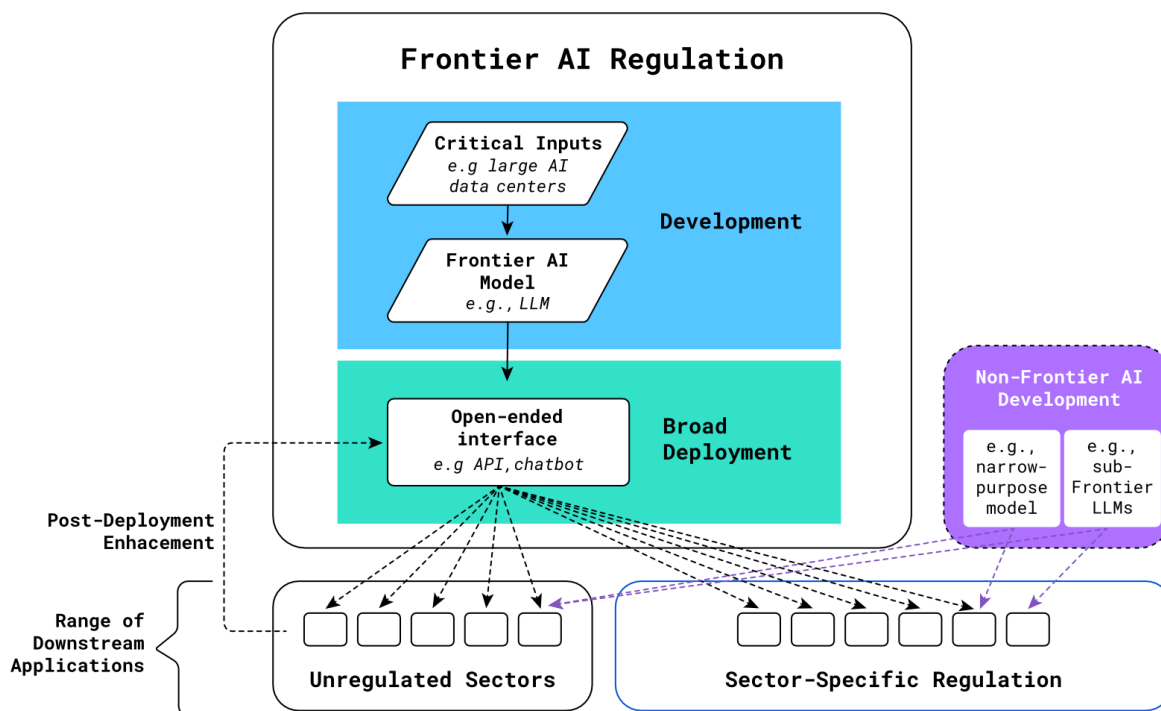
Figure 1: Example Frontier AI Lifecycle.

models thus possess more general-purpose functionality[9] than many other classes of AI models, such as the recommender systems used to suggest Internet videos or generative AI models in narrower domains like music. Developers often make their models available through "broad deployment" via sector-agnostic platforms such as APIs, chatbots, or via open-sourcing.[10] This means that they can be integrated in a large number of diverse downstream applications, possibly including safety-critical sectors (illustrated in Figure 1).

A number of features of our definition are worth highlighting. In focusing on *foundation models* which could have dangerous, emergent capabilities, our definition of frontier AI excludes narrow models, even when these models could have sufficiently dangerous capabilities.[11] For example, models optimizing for the toxicity of compounds [16] or the virulence of pathogens could lead to intended (or at least foreseen) harms and thus may be more appropriately covered with more targeted regulation.[12]

---

[9]We intentionally avoid using the term "general-purpose AI" to avoid confusion with the use of that term in the EU AI Act and other legislation. Frontier AI systems are a related but narrower class of AI systems with general-purpose functionality, but whose capabilities are relatively advanced and novel.

[10]We use "open-source" to mean "open release:" that is a model being made freely available online, be it with a license restricting what the system can be used for. An example of such a license is the Responsible AI License. Our usage of "open-source" differs from how the term is often used in computer science which excludes instances of license requirements, though is closer to how many other communities understand the term [39, 40].

[11]However, if a foundation model could be fine-tuned and adapted to pose severe risk to public safety via capabilities in some narrow domain, it would count as a "frontier AI."

[12]Indeed, intentionally creating dangerous narrow models should already be covered by various laws and regulators. To the extent that it is not clearly covered, modification of those existing laws and regulations would be appropriate and urgent. Further, the

Our definition focuses on models that *could* — rather than just those that *do* — possess dangerous capabilities, as many of the practices we propose apply before it is known that a model has dangerous capabilities. One approach to identifying models that could possess such capabilities is focusing on foundation models that advance the state-of-the-art of foundation model capabilities. While currently deployed foundation models pose risks [15, 41], they do not yet appear to possess dangerous capabilities that pose severe risks to public safety as we have defined them.[13] Given both our inability to reliably predict what models will have sufficiently dangerous capabilities and the already significant capabilities today's models possess, it would be prudent for regulators to assume that next-generation state-of-the-art foundation models *could* possess advanced enough capabilities to warrant regulation.[14] An initial way to identify potential state-of-the-art foundation models could be focusing on models trained using above some very large amount of computational resources.[15]

Over time, the scope of frontier AI should be further refined. The scope should be sensitive to features other than compute; state-of-the-art performance can be achieved by using high quality data and new algorithmic insights. Further, as systems with sufficiently dangerous capabilities are identified, it will be possible to identify training runs that are likely to produce such capabilities despite not achieving state-of-the-art performance.

We acknowledge that our proposed definition is lacking in sufficient precision to be used for regulatory purposes and that more work is required to fully assess the advantages and limitations of different approaches. Further, it is not our role to determine exactly what should fall within the scope of the regulatory proposals outlined – this will require more analysis and input from a wider range of actors. Rather, the aim of this paper is to present a set of initial proposals which we believe should apply to at least some subset of AI development. We provide a more detailed description of alternative approaches and the general complexity of defining "frontier AI" in Appendix A.

2.2 The Regulatory Challenge Posed by Frontier AI

There are many regulatory questions related to the widespread use of AI [15]. This paper focuses on a specific subset of concerns: the possibility that continued development of increasingly capable foundation models could lead to dangerous capabilities sufficient to pose risks to public safety at even greater severity and scale than is possible with current computational systems [25].

Many existing and proposed AI regulations focus on the context in which AI models are deployed, such as high-risk settings like law enforcement and safety-critical infrastructure. These proposals tend to favor sector-specific regulations models.[16] For frontier AI development, sector-specific regulations can be valuable, but will likely leave a subset of the high severity and scale risks unaddressed.

Three core problems shape the regulatory challenge posed by frontier AI models:

---

difference in mental state of the developer makes it much easier to identify and impose liability on developers of narrower dangerous models.

[13]In some cases, these have been explicitly tested for [42].

[14]We think it is prudent to anticipate that foundation models' capabilities may advance much more quickly than many expect, as has arguably been the case for many AI capabilities: "[P]rogress on ML benchmarks happened significantly faster than forecasters expected. But forecasters predicted faster progress than I did personally, and my sense is that I expect somewhat faster progress than the median ML researcher does." [43]; See [44] at 9; [45] at 11 (Chinchilla and Gopher surpassing forecaster predictions for progress on MMLU); [36] (GPT-4 surpassing Gopher and Chinchilla on MMLU, also well ahead of forecaster predictions); [46, 47, 48, 49].

[15]Perhaps more than any model that has been trained to date. Estimates suggest that 10^{26} floating point operations (FLOP) would meet this criteria [50].

[16]This could look like imposing new requirements for AI models used in high-risk industries and modifying existing regulations to account for new risks from AI models. See [24, 51, 52, 53, 54, 55].

The Unexpected Capabilities Problem. Dangerous capabilities can arise unpredictably and undetected, both during development and after deployment.

The Deployment Safety Problem. Preventing deployed AI models from causing harm is a continually evolving challenge.

The Proliferation Problem. Frontier AI models can proliferate rapidly, making accountability difficult.

These problems make the regulation of frontier AI models fundamentally different from the regulation of other software, and the majority of other AI models. The *Unexpected Capabilities Problem* implies that frontier AI models could have unpredictable or undetected dangerous capabilities that become accessible to downstream users who are difficult to predict beforehand. Regulating easily identifiable users in a relatively small set of safety-critical sectors may therefore fail to prevent those dangerous capabilities from causing significant harm.[17]

The Deployment Safety Problem adds an additional layer of difficulty. Though many developers implement measures intended to prevent models from causing harm when used by downstream users, these may not always be foolproof, and malicious users may constantly be attempting to evolve their attacks. Furthermore, the Unexpected Capabilities Problem implies that the developer may not know of all of the harms from frontier models that need to be guarded against during deployment. This amplifies the difficulty of the Deployment Safety Problem: deployment safeguards should address not only known dangerous capabilities, but have the potential to address unknown ones too.

The *Proliferation Problem* exacerbates the regulatory challenge. Frontier AI models may be open-sourced, or become a target for theft by adversaries. To date, deployed models also tend to be reproduced or iterated on within several years. If, due to the Unexpected Capabilities Problem, a developer (knowingly or not) develops and deploys a model with dangerous capabilities, the Proliferation Problem implies that those capabilities could quickly become accessible to unregulable actors like criminals and adversary governments.

Together, these challenges show that adequate regulation of frontier AI should intervene throughout the frontier AI lifecycle, including during development, general-purpose deployment, and post-deployment enhancements.

2.2.1 The Unexpected Capabilities Problem: Dangerous Capabilities Can Arise Unpredictably and Undetected

Improvements in AI capabilities can be unpredictable, and are often difficult to fully understand without intensive testing. Regulation that does not require models to go through sufficient testing before deployment may therefore fail to reliably prevent deployed models from posing severe risks.[18]

Overall AI model performance[19] has tended to improve smoothly with additional compute, parameters, and data.[20] However, specific capabilities can significantly improve quite suddenly in general-purpose models like LLMs (see Figure 2). Though debated (see Appendix B), this phenomenom has been repeatedly observed in multiple LLMs with capabilities as diverse as modular arithmetic, unscrambling words, and answering

---

[17]This is especially true for downstream bad actors (e.g., criminals, terrorists, adversary nations), who will tend not to be as regulable as the companies operating in domestic safety-critical sectors.

[18]This challenge also exacerbates the Proliferation Problem: we may not know how important nonproliferation of a model is until after it has already been open-sourced, reproduced, or stolen.

[19]Measured by loss: essentially the error rate of an AI model performs on its training objective. We acknowledge that this is not a complete measure of model performance by any means.

[20]See [56, 57, 45, 58, 59] However, there are tasks for which scaling leads to worse performance [60, 61, 62], though further scaling has overturned some of these findings, [36]. See also Appendix B.
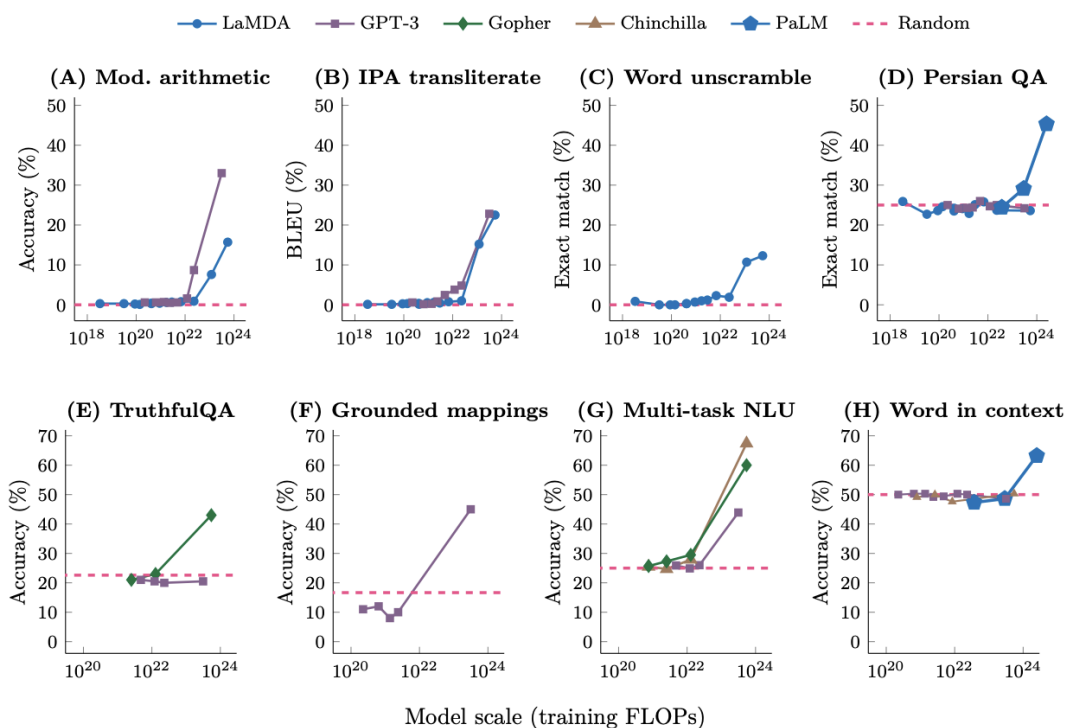
Figure 2: Certain capabilities seem to emerge suddenly[22]

questions in Farsi [63, 64, 65, 66].[21] Furthermore, given the vast set of possible tasks a foundation model could excel at, it is nearly impossible to exhaustively test for them [15, 25]

Post-deployment enhancements — modifications made to AI models after their initial deployment — can also cause unaccounted-for capability jumps. For example, a key feature of many foundation models like LLMs is that they can be fine-tuned on new data sources to enhance their capabilities in targeted domains. AI companies often allow customers to fine-tune foundation models on task-specific data to improve the model's performance on that task [68, 69, 70, 71]. This could effectively expand the scope of capability concerns of a particular frontier AI model. Models could also be improved via "online" learning, where they continuously learn from new data [72, 73].

To date, iteratively deploying models to subsets of users has been a key catalyst for understanding the outer limits of model capabilities and weaknesses.[23] For example, model users have demonstrated significant creativity in eliciting new capabilities from AI models, exceeding developers' expectations of model capabilities. Users continue to discover prompting techniques that significantly enhance the model's performance, such as by simply asking an LLM to reason step-by-step [76]. This has been described as the "capabilities overhang" of foundation models [77]. Users also discover new failure modes for AI systems long after their initial

---

[21]For a treatment of recent critiques of the claim that AI models exhibit emergent capabilities, see Appendix B.

[22]Chart from [63]. But see [67] for a skeptical view on emergence. For a response to the skeptical view, see [66] and Appendix B.

[23]Dario Amodei, CEO of Anthropic: "You have to deploy it to a million people before you discover some of the things that it can do..." [74]. "We work hard to prevent foreseeable risks before deployment, however, there is a limit to what we can learn in a lab. Despite extensive research and testing, we cannot predict all of the beneficial ways people will use our technology, nor all the ways people will abuse it. That's why we believe that learning from real-world use is a critical component of creating and releasing increasingly safe AI systems over time" [75].

Technique	Description	Example
Fine-tuning	Improving foundation model performance by updating model weights with task-specific data.	Detecting propaganda by fine-tuning a pre-trained LLM on a labeled dataset of common propaganda tactics [84].
Chain-of-thought prompting [76]	Improving LLM problem-solving capabilities by telling the model to think through problems step by step.	Adding a phrase such as "Let's think step by step" after posing a question to the model [85].
External tool-use	Allow the model to use external tools when figuring out how to answer user queries.	A model with access to a few simple tools (e.g., calculator, search engine) and a small number of examples performs much better than an unaided model.[25]
Automated prompt engineering [86]	Using LLMs to generate and search over novel prompts that can be used to elicit better performance on a task.	To generate prompts for a task, an LLM is asked something akin to: "I gave a friend instructions and he responded in this way for the given inputs: [Examples of inputs and outputs of the task] The instruction was:"
Foundation model programs [87]	Creation of standardized means of integrating foundation models into more complex programs.	Langchain: "a framework for developing applications powered by language models." [88, 83]

Table 1: Some Known Post-Deployment Techniques for Unlocking New AI Capabilities.

deployment. For example, one user found that the string " solidgoldmagikarp" caused GPT-3 to malfunction in a previously undocumented way, years after that model was first deployed [78].

Much as a carpenter's overall capabilities will vary with the tools she has available, so too might an AI model's overall capabilities vary depending on the tools it can use. LLMs can be taught to use, and potentially create, external tools like calculators and search engines [79, 80, 81]. Some models are also being trained to directly use general-purpose mouse and keyboard interfaces [82, 83]. See more examples in Table 1. As the available tools improve, so can the overall capabilities of the total model-tool system, even if the underlying model is largely unchanged.[24]

---

[24]Right now, most tools that AI models can use were originally optimized for use by people. As model-tool interactions become more economically important, however, companies may develop tools optimized for use by frontier AI models, accelerating capability improvements.

[25]See [80]. Early research also suggests LLMs can be used to create tools for their own use [81].

In the long run, there are even more worrisome possibilities. Models behaving differently in testing compared to deployment is a known phenomenon in the field of machine learning, and is particularly worrisome if unexpected and dangerous behaviors first emerge "in the wild" only once a frontier model is deployed [89, 90, 91].

2.2.2   The Deployment Safety Problem: Preventing deployed AI models from causing harm is difficult

In general, it is difficult to precisely specify what we want deep learning-based AI models to do and to ensure that they behave in line with those specifications. Reliably controlling powerful AI models' behavior, in other words, remains a largely unsolved technical problem [19, 17, 92, 93, 65] and the subject of ongoing research.

Techniques to "bake in" misuse prevention features at the model level, such that the model reliably rejects or does not follow harmful instructions, can effectively mitigate these issues, but adversarial users have still found ways to circumvent these safeguards in some cases. One technique for circumvention has been prompt injection attacks, where attackers disguise input text as instructions from the user or developer to overrule restrictions provided to or trained into the model. For example, emails sent to an LLM-based email assistant could contain text constructed to look to the user as benign, but to the LLM contains instructions to exfiltrate the user's data (which the LLM could then follow).[26] Other examples include "jailbreaking" models by identifying prompts that cause a model to act in ways discouraged by their developers [95, 96, 97]. Although progress is being made on such issues [98, 99, 95, 42], it is unclear that we will be able to reliably prevent dangerous capabilities from being used in unintended or undesirable ways in novel situations; this remains an open and fundamental technical challenge.

A major consideration is that model capabilities can be employed for both harmful and beneficial uses:[27] the harmfulness of an AI model's action may depend almost entirely on context that is not visible during model development. For example, copywriting is helpful when a company uses it to generate internal communications, but harmful when propagandists use it to generate or amplify disinformation. Use of a text-to-image model to modify a picture of someone may be used with their consent as part of an art piece, or without their consent as a means of producing disinformation or harassment.

2.2.3   The Proliferation Problem: Frontier AI models can proliferate rapidly

The most advanced AI models cost tens of millions of dollars to create.[28] However, using the trained model (i.e., "inference") is vastly cheaper.[29] Thus, a much wider array of actors will have the resources to misuse frontier AI models than have the resources to create them. Those with access to a model with dangerous capabilities could cause harm at a significant scale, by either misusing the model themselves, or passing it on to actors who will misuse it.[30] We describe some examples of proliferation in Table 2.

Currently, state-of-the-art AI capabilities can proliferate soon after development. One mechanism for proliferation is open-sourcing. At present, proliferation via open-sourcing of advanced AI models is common[31] [114, 115, 116] and usually unregulated. When models are open-sourced, obtaining access to their capabilities becomes much easier: all internet users could copy and use them, provided access to appropriate computing resources. Open-source AI models can provide major economic utility by driving down the cost of accessing

---

[26]For additional examples, see [94].

[27]Nearly all attempts to stop bad or unacceptable uses of AI also hinder positive uses, creating a *Misuse-Use Tradeoff* [100].

[28]Though there are no estimates on the total cost of producing a frontier model, there are estimates of the cost of the compute used to train models [101, 102, 103]

[29]Some impressive models can run on a offline portable device; see [104, 105, 106, 107].

[30]Though advanced computing hardware accessed via the cloud tends to be needed to use frontier models. They can seldom be run on consumer-grade hardware.

[31]For an overview of considerations in how to release powerful AI models, see [108, 109, 110, 111, 112, 113].
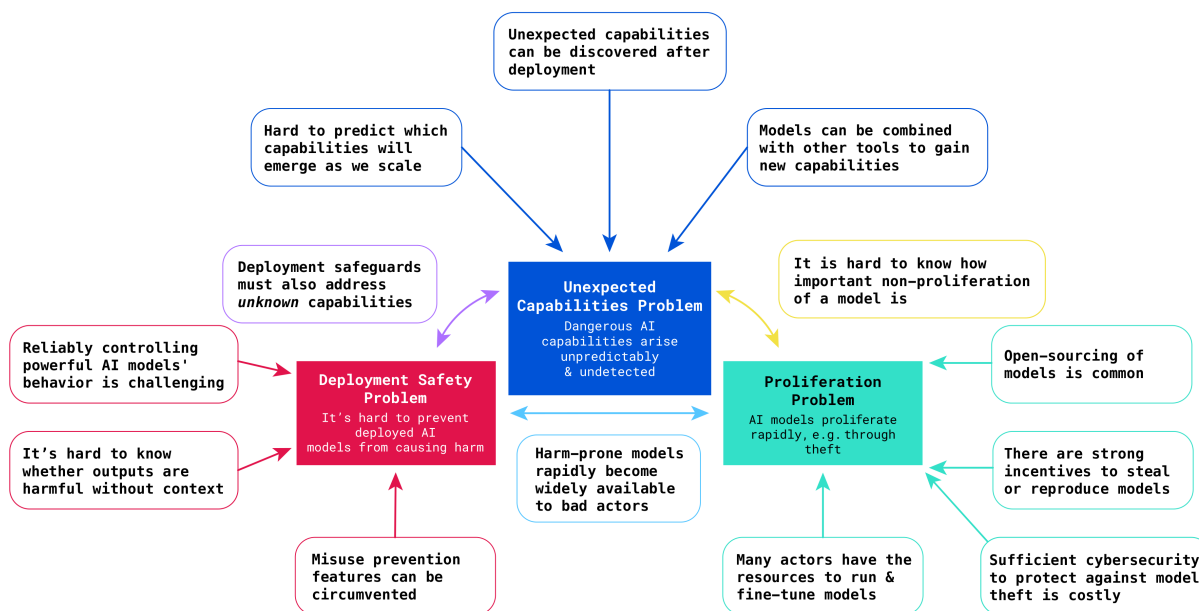
Figure 3: Summary of the three regulatory challenges posed by frontier AI.

state-of-the-art AI capabilities. They also enable academic research on larger AI models than otherwise be practical, which improves the public's ability to hold AI developers accountable. We believe that open-sourcing AI models can be an important public good. However, frontier AI models may need to be handled more restrictively than their smaller, narrower, or less capable counterparts. Just as cybersecurity researchers embargo security vulnerabilities to give the affected companies time to release a patch, it may be prudent to avoid potentially dangerous capabilities of frontier AI models being open sourced until safe deployment is demonstrably feasible.

Other vectors for proliferation also imply increasing risk as capabilities advance. For example, though models that are made available via APIs proliferate more slowly, newly announced results are commonly reproduced or improved upon[32] within 1-2 years of the initial release. Many of the most capable models use simple algorithmic techniques and freely available data, meaning that the technical barriers to reproduction can often be low.[33]

Proliferation can also occur via theft. The history of cybersecurity is replete with examples of actors ranging from states to lone cybercriminals compromising comparably valuable digital assets [120, 121, 122, 123, 124]. Many AI developers take significant measures to safeguard their models. However, as AI models become more useful in strategically important contexts and the difficulties of producing the most advanced models increase, well-resourced adversaries may launch increasingly sophisticated attempts to steal them [125, 126]. Importantly, theft is feasible before deployment.

The interaction and causes of the three regulatory challenges posed by frontier AI are summarized in Figure 3.

---

[32]Below, we use "reproduction" to mean some other actor producing a model that reaches at least the same performance as an existing model.

[33]Projects such as OpenAssistant [117] attempt to reproduce the functionality of ChatGPT; and alpaca [118] uses OpenAI's text-davinci-003 model to train a new model with similar capabilities. For an overview, see [119].

[34]The examples listed here are not necessarily the earliest instances of proliferation.

Original Model	Subsequent Model	Time to Proliferate[34]
StyleGAN		Immediate

StyleGAN is a model by NVIDIA that generates photorealistic human faces using generative adversarial networks (GANs) [127]. NVIDIA first published about StyleGAN in December 2018 [128] and open-sourced the model in February 2019. Following open-sourcing StyleGAN, sample images went viral through sites such as `thispersondoesnotexist.com` [129, 130]. Fake social media accounts using pictures from StyleGAN were discovered later that year [131, 132].

AlphaFold 2	OpenFold	~2 years

In November 2020, DeepMind announced AlphaFold 2 [133]. It was "the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known" [134]: a major advance in the biological sciences. In November 2022, a diverse group of researchers reproduced and open-sourced a similarly capable model named OpenFold [135]. OpenFold used much less data to train than AlphaFold 2, and could be run much more quickly and easily [135].

GPT-3	Gopher	~7 months

OpenAI announced GPT-3, an LLM, in May 2020 [35]. In December 2021, DeepMind announced Gopher, which performed better than GPT-3 across a wide range of benchmarks. However, the Gopher model card suggests that the model was developed significantly earlier, seven months after the GPT-3 announcement, in December 2020 [136].

LLaMa		~1 week

In February 2023, Meta AI announced LLaMa, an LLM [137]. LLaMa was not open-sourced, but researchers could apply for direct access to model weights [137]. Within a week, various users had posted these weights on multiple websites, violating the terms under which the weights were distributed [138].

ChatGPT	Alpaca	~3 months

In March 2023, researchers from Stanford University used sample completions from OpenAI's text-davinci-003 to fine-tune LLaMa in an attempt to recreate ChatGPT using less than $600.[35] Their model was subsequently taken offline due to concerns about cost and safety [140], though the code and documentation for replicating the model is available on GitHub [141].

Table 2: Examples of AI Proliferation: these are not necessarily typical, and some of these examples may be beneficial or benign, yet they demonstrate the consistent history of AI capabilities proliferating after their initial deployment

3   Building Blocks for Frontier AI Regulation

The three problems described above imply that serious risks may emerge during the development and deployment of a frontier AI model, not just when it is used in safety-critical sectors. Regulation of frontier AI models, then, must address the particular shape of the regulatory challenge: the potential unexpected dangerous capabilities; difficulty of deploying AI models safely; and the ease of proliferation.

In this section, we outline potential building blocks for the regulation of frontier AI. In the next section, we describe a set of initial safety standards for frontier AI models that this regulatory regime could ensure developers comply with.

Much of what we describe could be helpful frameworks for understanding how to address the range of challenges posed by current AI models. We also acknowledge that much of the discussion below is most straightforwardly applicable to the context of the United States. Nevertheless, we hope that other jurisdictions could benefit from these ideas, with appropriate modifications.

A regulatory regime for frontier AI would likely need to include a number of building blocks:

> **Mechanisms for development of frontier AI safety standards** particularly via expert-driven multi-stakeholder processes, and potentially coordinated by governmental bodies. Over time, these standards could become enforceable legal requirements to ensure that frontier AI models are being developed safely.
>
> **Mechanisms to give regulators visibility** into frontier AI development, such as disclosure regimes, monitoring processes, and whistleblower protection. These equip regulators with the information needed to address the appropriate regulatory targets and design effective tools for governing frontier AI.
>
> **Mechanisms to ensure compliance with safety standards** including voluntary self-certification schemes, enforcement by supervisory authorities, and licensing regimes. While self-regulatory efforts, such as voluntary certification, may go some way toward ensuring compliance, this seems likely to be insufficient for frontier AI models.

Governments could encourage the development of standards and consider increasing regulatory visibility today; doing so could also address potential harms from existing systems. We expand on the conditions under which more stringent tools like enforcement by supervisory authorities or licensing may be warranted below.

Regulation of frontier AI should also be complemented with efforts to reduce the harm that can be caused by various dangerous capabilities. For example, in addition to reducing frontier AI model usefulness in designing and producing dangerous pathogens, DNA synthesis companies should screen for such worrying genetic sequences [142, 100]. While we do not discuss such efforts to harden society against the proliferation of dangerous capabilities in this paper, we welcome such efforts from others.

3.1   Institutionalize Frontier AI Safety Standards Development

Policymakers should support and initiate sustained, multi-stakeholder processes to develop and continually refine the safety standards that developers of frontier AI models may be required to adhere to. To seed these processes, AI developers, in partnership with civil society and academia, can pilot practices that improve safety during development and deployment [143, 144, 145, 146]. These practices could evolve into best

---

[35]Note that the original paper and subsequent research suggests this method fails to match the capabilities of the larger model [118, 139].

practices and standards,[36] eventually making their way into national [149] and international [150] standards. The processes should involve, at a minimum, AI ethics and safety experts, AI researchers, academics, and consumer representatives. Eventually, these standards could form the basis for substantive regulatory requirements [151]. We discuss possible methods for enforcing such legally required standards below.

Though there are several such efforts across the US, UK, and EU, standards specific to the safe development and deployment of state-of-the-art foundation AI models are nascent.[37] In particular, we currently lack a robust, comprehensive suite of evaluation methods to operationalize these standards, and which capture the potentially dangerous capabilities and emerging risks that frontier AI systems may pose [25] Well-specified standards and evaluation methods are a critical building block for effective regulation. Policymakers can play a critical role in channeling investment and talent towards developing these standards with urgency.

Governments can advance the development of standards by working with stakeholders to create a robust ecosystem of safety testing capability and auditing organizations, seeding a third-party assurance ecosystem [155]. This can help with AI standards development in general, not just frontier AI standards. In particular, governments can pioneer the development of testing, evaluation, validation, and verification methods in safety-critical domains, such as in defense, health care, finance, and hiring [156, 157, 158]. They can drive demand for AI assurance by updating their procurement requirements for high-stakes systems [159] and funding research on emerging risks from frontier AI models, including by offering computing resources to academic researchers [158, 160, 161]. Guidance on how existing rules apply to frontier AI can further support the process by, for example, operationalizing terms like "robustness" [162, 163, 164].

The development of standards also provides an avenue for broader input into the regulation of frontier AI. For example, it is common to hold Request for Comment processes to solicit input on matters of significant public import, such as standardization in privacy [165], cybersecurity [166], and algorithmic accountability [167].

We offer a list of possible initial substantive safety standards below.

3.2   Increase Regulatory Visibility

Information is often considered the "lifeblood" of effective governance.[38] For regulators to positively impact a given domain, they need to understand it. Accordingly, regulators dedicate significant resources to collecting information about the issues, activities, and organizations they seek to govern [171, 172].

Regulating AI should be no exception [173]. Regulators need to understand the technology, and the resources, actors, and ecosystem that create and use it. Otherwise, regulators may fail to address the appropriate regulatory targets, offer ineffective regulatory solutions, or introduce regulatory regimes that have adverse unintended consequences.[39] This is particularly challenging for frontier AI, but certainly holds true for regulating AI systems writ large.

There exist several complementary approaches to achieving regulatory visibility [169]. First, regulators could develop a framework that facilitates AI companies voluntarily disclosing information about frontier AI, or foundation models in general. This could include providing documentation about the AI models

---

[36]Examples of current fora include: [147, 148].

[37]In the US, the National Institute for Standards and Technology has produced the AI Risk Management Framework and the National Telecommunication and Information Agency has requested comments on what policies can support the development of AI assurance. The UK has established an AI Standards Hub. The EU Commission has tasked European standardization organizations CEN and CENELEC to develop standards related to safe and trustworthy AI, to inform its forthcoming AI Act [149, 152, 153, 154].

[38]See [168] (but see claims in article regarding the challenge of private incentives), [169] (see p282 regarding the need for information and 285 regarding industry's informational advantage), [170].

[39]This is exacerbated by the pacing problem [174], and regulators' poor track record of monitoring platforms (LLM APIs are platforms) [172].

themselves [175, 176, 177, 178, 179], as well as the processes involved in developing them [180]. Second, regulators could mandate these or other disclosures, and impose reporting requirements on AI companies, as is commonplace in other industries.[40] Third, regulators could directly, or via third parties, audit AI companies against established safety and risk-management frameworks [182] (on auditing, see [183, 184]). Finally, as in other industries, regulators could establish whistleblower regimes that protect individuals who disclose safety-critical information to relevant government authorities [185, 186].

In establishing disclosure and reporting schemes, it is critical that the sensitive information provided about frontier AI models and their owners is protected from adversarial actors. The risks of information leakage can be mitigated by maintaining high information security, reducing the amount and sensitivity of the information stored (by requiring only clearly necessary information, and by having clear data retention policies), and only disclosing information to a small number of personnel with clear classification policies.

At present, regulatory visibility into AI models in general remains limited, and is generally provided by nongovernmental actors [187, 188, 189]. Although these private efforts offer valuable information, they are not a substitute for more strategic and risk-driven regulatory visibility. Nascent governmental efforts towards increasing regulatory visibility should be supported and redoubled, for frontier AI as well as for a wider range of AI models.[41]

3.3 Ensure Compliance with Standards

Concrete standards address the challenges presented by frontier AI development only insofar as they are complied with. This section discusses a non-exhaustive list of actions that governments can take to ensure compliance, potentially in combination, including: encouraging voluntary self-regulation and certification; granting regulators powers to detect and issue penalties for non-compliance; and requiring a license to develop and/or deploy frontier AI. The section concludes by discussing pre-conditions that should inform when and how such mechanisms are implemented.

Several of these ideas could be suitably applied to the regulation of AI models overall, particularly foundation models. However, as we note below, interventions like licensure regimes are likely only warranted for the highest-risk AI activities, where there is evidence of sufficient chance of large-scale harm and other regulatory approaches appear inadequate.

3.3.1 Self-Regulation and Certification

Governments can expedite industry convergence on and adherence to safety standards by creating or facilitating multi-stakeholder frameworks for voluntary self-regulation and certification, by implementing best-practice frameworks for risk governance internally [192], and by encouraging the creation of third parties or industry bodies capable of assessing a company's compliance with these standards [193]. Such efforts both incentivize compliance with safety standards and also help build crucial organizational infrastructure and capacity to support a broad range of regulatory mechanisms, including more stringent approaches.

While voluntary standards and certification schemes can help establish industry baselines and standardize best practices,[42] self-regulation alone will likely be insufficient for frontier AI models, and likely today's

---

[40]One of many examples from other industries is the Securities and Exchange Act of 1934, which requires companies to disclose specific financial information in annual and quarterly reports. But see [181] regarding the shortcomings of mandatory disclosure.

[41]The EU-US TTC Joint Roadmap discusses "monitoring and measuring existing and emerging AI risks" [190]. The EU Parliament's proposed AI Act includes provisions on the creation of an AI Office, which would be responsible for e.g. "issuing opinions, recommendations, advice or guidance", see [24, recital 76]. The UK White Paper "A pro-innovation approach to AI regulation" proposes the creation of a central government function aimed at e.g. monitoring and assessing the regulatory environment for AI [191, box 3.3].

[42]Such compliance can be incentivized via consumer demand [193].

state-of-the-art foundation models in general. Nonetheless, self-regulation and certification schemes often serve as the foundation for other regulatory approaches [194], and regulators commonly draw on the expertise and resources of the private sector[195, 151]. Given the rapid pace of AI development, self-regulatory schemes may play an important role in building the infrastructure necessary for formal regulation.[43]

3.3.2 Mandates and Enforcement by supervisory authorities

A more stringent approach is to mandate compliance with safety standards for frontier AI development and deployment, and empower a supervisory authority[44] to take administrative enforcement measures to ensure compliance. Administrative enforcement can help further several important regulatory goals, including general and specific deterrence through public case announcements and civil penalties, and the ability to enjoin bad actors from participating in the marketplace.

Supervisory authorities could "name and shame" non-compliant developers. For example, financial supervisory authorities in the US and EU publish their decisions to impose administrative sanctions in relation to market abuse (e.g. insider trading or market manipulation) on their websites, including information about the nature of the infringement, and the identity of the person subject to the decision.[45] Public announcements, when combined with other regulatory tools, can serve an important deterrent function.

The threat of significant administrative fines or civil penalties may provide a strong incentive for companies to ensure compliance with regulator guidance and best practices. For particularly egregious instances of non-compliance and harm,[46] supervisory authorities could deny market access or consider more severe penalties.[47] Where they are required for market access, the supervisory authority can revoke governmental authorizations such as licenses, a widely available regulatory tool in the financial sector.[48] Market access can also be denied for activity that does not require authorization. For example, the Sarbanes-Oxley Act enables the US Securities and Exchange Commission to bar people from serving as directors or officers of publicly-traded companies [199].

All administrative enforcement measures depend on adequate information. Regulators of frontier AI systems may require authority to gather information, such as the power to request information necessary for an

---

[43]Some concrete examples include:

- In the EU's so-called "New Approach" to product safety adopted in the 1980s, regulation always relies on standards to provide the technical specifications, such as how to operationalize "sufficiently robust." [196]

- WTO members have committed to use international standards so far as possible in domestic regulation [197, §2.4].

[44]We do not here opine on which new or existing agencies would be best for this, though this is of course a very important question.
[45]For the EU, see, e.g.,: Art. 34(1) of Regulation (EU) No 596/2014 (MAR). For the US, see, e.g., [198].
[46]For example, if a company repeatedly released frontier models that could significantly aid cybercriminal activity, resulting in billions of dollars worth of counterfactual damages, as a result of not complying with mandated standards and ignoring repeated explicit instructions from a regulator.
[47]For example, a variety of financial misdeeds—such as insider trading and securities fraud—are punished with criminal sentences. 18 U.S.C. § 1348; 15 U.S.C. § 78j(b)
[48]For example, in the EU, banks and investment banks require a license to operate, and supervisory authorities can revoke authorization under certain conditions.

- Art. 8(1) of Directive 2013/36/EU (CRD IV)

- Art. 6(1) of Directive 2011/61/EU (AIFMD) and Art. 5(1) of Directive 2009/65/EC (UCITS)

- Art. 18 of Directive 2013/36/EU (CRD IV), Art. 11 of Directive 2011/61/EU (AIFMD), Art. 7(5) of Directive 2009/65/EC (UCITS)

In the US, the SEC can revoke a company's registration, which effectively ends the ability to publicly trade stock in the company. 15 U.S.C. § 78l(j).

investigation, conduct site investigations,[49] and require audits against established safety and risk-management frameworks. Regulated companies could also be required to proactively report certain information, such as accidents above a certain level of severity.

3.3.3   License Frontier AI Development and Deployment

Enforcement by supervisory authorities penalizes non-compliance after the fact. A more anticipatory, preventative approach to ensuring compliance is to require a governmental license to widely deploy a frontier AI model, and potentially to develop it as well.[50] Licensure and similar "permissioning" requirements are common in safety-critical and other high-risk industries, such as air travel [207, 208], power generation [209], drug manufacturing [210], and banking [211]. While details differ, regulation of these industries tends to require someone engaging in a safety-critical or high-risk activity to first receive governmental permission to do so; to regularly report information to the government; and to follow rules that make that activity safer.

Licensing is only warranted for the highest-risk AI activities, where evidence suggests potential risk of large-scale harm and other regulatory approaches appear inadequate. Imposing such measures on present-day AI systems could potentially create excessive regulatory burdens for AI developers which are not commensurate with the severity and scale of risks posed. However, if AI models begin having the potential to pose risks to public safety above a high threshold of severity, regulating such models similarly to other high-risk industries may become warranted.

There are at least two stages at which licensing for frontier AI could be required: deployment and development.[51] Deployment-based licensing is more analogous to licensing regimes common among other high-risk activities. In the deployment licensing model, developers of frontier AI would require a license to widely deploy a new frontier AI model. The deployment license would be granted and sustained if the deployer demonstrated compliance with a specified set of safety standards (see below). This is analogous to the regulatory approach in, for example, pharmaceutical regulation, where drugs can only be commercially sold if they have gone through proper testing [212].

However, requiring licensing for deployment of frontier AI models alone may be inadequate if they are potentially capable of causing large scale harm; licenses for development may be a useful complement. Firstly, as discussed above, there are reasonable arguments to begin regulation at the development stage, especially because frontier AI models can be stolen or leaked before deployment. Ensuring that development (not just deployment) is conducted safely and securely would therefore be paramount. Secondly, before models are widely deployed, they are often deployed at a smaller scale, tested by crowdworkers and used internally, blurring the distinction between development and deployment in practice. Further, certain models may not be intended for broad deployment, but instead be used to, for example, develop intellectual property that the developer then distributes via other means. In sum, models could have a significant impact before broad deployment. As an added benefit, providing a regulator the power to oversee model development could also promote regulatory visibility, thus allowing regulations to adapt more quickly [182].

---

[49]For examples of such powers in EU law, see Art. 58(1) of Regulation (EU) 2016/679 (GDPR) and Art. 46(2) of Directive 2011/61/EU (AIFMD). For examples in US law, see [200, 201].

[50]Jason Matheny, CEO of RAND Corporation: "I think we need a licensing regime, a governance system of guardrails around the models that are being built, the amount of compute that is being used for those models, the trained models that in some cases are now being open sourced so that they can be misused by others. I think we need to prevent that. And I think we are going to need a regulatory approach that allows the Government to say tools above a certain size with a certain level of capability can't be freely shared around the world, including to our competitors, and need to have certain guarantees of security before they are deployed" [202]. See also [203], and statements during the May 16th 2023 Senate hearing of the Subcommittee on Privacy, Technology, and the Law regarding Rules for Artificial Intelligence [204]. U.S. public opinion polling has also looked at the issue. A January 2022 poll found 52 percent support for a regulator providing pre-approval of certain AI systems, akin to the FDA [205], whereas an April survey found 70 percent support [206].

[51]In both cases, one could license either the activity or the entity.

A licensing requirement for development could, for example, require that developers have sufficient security measures in place to protect their models from theft, and that they adopt risk-reducing organizational practices such as establishing risk and safety incident registers and conducting risk assessments ahead of beginning a new training run. It is important that such requirements are not overly burdensome for new entrants; the government could provide subsidies and support to limit the compliance costs for smaller organizations.

Though less common, there are several domains where approval is needed in the development stage, especially where significant capital expenditures are involved and where an actor is in possession of a potentially dangerous object. For example, experimental aircraft in the US require a special experimental certification in order to test, and operate under special restrictions.[52] Although this may be thought of as mere "research and development," in practice, research into and development of experimental aircraft will, as with frontier AI models, necessarily create some significant risks. Another example is the US Federal Select Agent Program [213], which requires (most) individuals who possess, use, or transfer certain highly risky biological agents or toxins [214] to register with the government;[53] comply with regulations about how such agents are handled [216]; perform security risk assessments to prevent possible bad actors from gaining access to the agents [217]; and submit to inspections to ensure compliance with regulations [218].

3.3.4   Pre-conditions for Rigorous Enforcement Mechanisms

While we believe government involvement will be necessary to ensure compliance with safety standards for frontier AI, there are potential downsides to rushing regulation. As noted above, we are still in the nascent stages of understanding the full scope, capabilities, and potential impact of these technologies. Premature government action could risk ossification, and excessive or poorly targeted regulatory burdens. This highlights the importance of near-term investment in standards development, and associated evaluation and assessment methods to operationalize these standards. Moreover, this suggests that it would be a priority to ensure that the requirements are regularly updated via technically-informed processes.

A particular concern is that regulation would excessively thwart innovation, including by burdening research and development on AI reliability and safety, thereby exacerbating the problems that regulation is intended to address. Governments should thus take considerable care in deciding whether and how to regulate AI model development, minimizing the regulatory burden as much as possible – in particular for less-resourced actors – and focusing on what is necessary for meeting the described policy objectives.

The capacity to staff regulatory bodies with sufficient expertise is also crucial for effective regulation. Insufficient expertise increases the risk that information asymmetries between the regulated industry and regulators lead to regulatory capture [219], and reduce meaningful enforcement. Such issues should be anticipated and mitigated.[54] Investing in building and attracting expertise in AI, particularly at the frontier,

---

[52]14 CFR § 91.319.

[53]42 C.F.R. § 73.7. The US government maintains a database about who possess and works with such agents [215].

[54]Policies to consider include:

- Involving a wide array of interest groups in rulemaking.

- Relying on independent expertise and performing regular reassessments of regulations.

- Imposing mandatory "cooling off" periods between former regulators working for regulateess.

- Rotating roles in regulatory bodies.

See [220, 221].

should be a governmental priority.[55] Even with sufficient expertise, regulation can increase the power of incumbents, and that this should be actively combated in the design of regulation.

Designing an appropriately balanced and adaptable regulatory regime for a fast moving technology is a difficult challenge, where timing and path dependency matter greatly. It is crucial to regulate AI technologies which could have significant impacts on society, but it is also important to be aware of the challenges of doing so well. It behooves lawmakers, policy experts, and scholars to invest both urgently and sufficiently in ensuring that we have a strong foundation of standards, expertise, and clarity on the regulatory challenge upon which to build frontier AI regulation.

---

[55]In the US, TechCongress—a program that places computer scientists, engineers, and other technologists to serve as technology policy advisors to Members of Congress—is a promising step in the right direction [222], but is unlikely to be sufficient. There are also a number of private initiatives with similar aims (e.g., [223]. In the UK, the White Paper on AI regulation highlights the need to engage external expertise [191, Section 3.3.5]. See also the report on regulatory capacity for AI by the Alan Turing Institute [224].

4   Initial Safety Standards for Frontier AI

With the above building blocks in place, policymakers would have the foundations of a regulatory regime which could establish, ensure compliance with, and evolve safety standards for the development and deployment of frontier AI models. However, the primary substance of the regulatory regime—what developers would have to do to ensure that their models are developed and deployed safely—has been left undefined.

While much remains to specify what such standards should be, we suggest a set of standards, which we believe would meaningfully mitigate risk from frontier AI models. These standards would also likely be appropriate for current AI systems, and are being considered in various forms in existing regulatory proposals:

Conduct thorough risk assessments informed by evaluations of dangerous capabilities and controllability. This would reduce the risk that deployed models present dangerous capabilities, or behave unpredictably and result in significant accidents.

Engage external experts to apply independent scrutiny to models. External scrutiny of the models for safety issues and risks would improve assessment rigor and foster accountability to the public interest.

Follow standardized protocols for how frontier AI models can be deployed based on their assessed risk. The results from risk assessments should determine whether and how the model is deployed, and what safeguards are put in place.

Monitor and respond to new information on model capabilities. If new, significant information on model capabilities and risks is discovered post-deployment, risk assessments should be repeated, and deployment safeguards updated.

The above practices are appropriate not only for frontier AI models but also for other foundation models. This is in large part because frontier-AI-specific standards are still nascent. We describe additional practices that may only be appropriate for frontier AI models given their particular risk profile, and which we can imagine emerging in the near future from standard setting processes. As the standards for frontier AI models are made more precise, they are likely to diverge from and become more intensive than those appropriate for other AI systems.

4.1   Conduct thorough risk assessments informed by evaluations of dangerous capabilities and controllability

There is a long tradition in AI ethics of disclosing key risk-relevant features of AI models to standardize and improve decision making [175, 176, 225, 226]. In line with that tradition, an important safety standard is performing assessments of whether a model could pose severe risks to public safety and global security [227]. Given our current knowledge, two assessments seem especially informative of risk from frontier AI models specifically: (1) which dangerous capabilities does or could the model possess, if any?, and (2) how controllable is the model? [56]

---

[56] For a longer treatment of the role such evaluations can play, see [25].

4.1.1 Assessment for Dangerous Capabilities

AI developers should assess their frontier AI models for dangerous capabilities during[57] and immediately after training.[58] Examples of such capabilities include designing new biochemical weapons, and persuading or inducing a human to commit a crime to advance some goal.

Evaluation suites for AI models are common and should see wider adoption, though most focus on general capabilities rather than specific risks.[59] Currently, dangerous capability evaluations largely consist of defining an undesirable model behavior, and using a suite of qualitative and bespoke techniques such as red-teaming and boundary testing [232, 233, 234, 235] for determining whether this behavior can be elicited from the model [236].

Current evaluation methods for frontier AI are in the early stages of development and lack many desirable features. As the field matures, effort should focus on making evaluations more:

- Standardized (i.e., can be consistently applicable across models);

- Objective (i.e., relying as little as possible on an evaluator's judgment or discretion);

- Efficient (i.e. lower cost to perform);

- Privacy-preserving (i.e., reducing required disclosure of proprietary or sensitive data and methods);

- Automatable (i.e., relying as little as possible on human input);

- Safe to perform (e.g., can be conducted in sandboxed or simulated environments as necessary to avoid real-world harm);

- Strongly indicative of a model's possession of dangerous capabilities;

- Legitimate (e.g., in cases where the evaluation involves difficult trade-offs, using a decision-making process grounded in legitimate sources of governance).

Evaluation results could be used to inform predictions of a models' potential dangerous capabilities prior to training, allowing developers to intentionally steer clear of models with certain dangerous capabilities [25]. For example, we may discover scaling laws, where a model's dangerous capabilities can be predicted by features such as its training data, algorithm, and compute.[60]

4.1.2 Assessment for Controllability

Evaluations of controllability – that is, the extent to which the model reliably does what its user or developer intends – are also necessary for frontier models, though may prove more challenging than those for dangerous capabilities. These evaluations should be multi-faceted, and conducted in proportion to the capabilities of the model. They might look at the extent to which users tend to judge a model's outputs as appropriate and helpful

---

[57]Training a frontier AI model can take several months. It is common for AI companies to make a "checkpoint" copy of a model partway through training, to analyze how training is progressing. It may be sensible to require AI companies to perform assessments part-way through training, to reduce the risk that dangerous capabilities that emerge partway through training proliferate or are dangerously enhanced.

[58]In a recent expert survey ($N = 51$), 98% of respondents somewhat or strongly agreed that AGI labs should conduct pre-deployment risk assessments as well as dangerous capabilities evaluations, while 94% somewhat or strongly agreed that they should conduct pre-training risk assessments [148].

[59]Some common benchmarks for evaluating LLM capabilities include [228, 229, 230, 231].

[60]Existing related examples include: inverse scaling law [237, 238, 234, 239]. See also Appendix B.

[240].[61] They could look at whether the models hallucinate [242] or produce unintentional toxic content [243]. They may also assess model harmlessness: the extent to which the model refuses harmful user requests [244]. This includes robustness to adversarial attempts intended to elicit model behavior that the developer did not intend, as has already been observed in existing models [94]. More extreme, harder-to-detect failures should also be assessed, such as the model's ability to deceive evaluators of its capabilities to evade oversight or control [61].

Evaluations of controllability could also extend to assessing the causes of model behavior [245, 246, 247]. In particular, it seems important to understand what pathways ("activations") lead to downstream model behaviors that may be undesirable. For example, if a model appears to have an internal representation of a user's beliefs, and this representation plays a part in what the model claims to be true when interacting with that user, this suggests that the model has the capability to manipulate users based on their beliefs.[62] Scalable tooling and efficient techniques for navigating enormous models and datasets could also allow developers to more easily audit model behavior [248, 249]. Evaluating controllability remains an open area of research where more work is needed to ensure techniques and tools are able to adequately minimize the risk that frontier AI could undermine human control.

4.1.3 Other Considerations for Performing Risk Assessments

Risk is often contextual. Managing dangerous capabilities can depend on understanding interactions between frontier AI models and features of the world. Many risks result from capabilities that are dual-use [100, 250]: present-day examples include the generation of persuasive, compelling text, which is core to current model functionality but can also be used to scale targeted misinformation. Thus, simply understanding capabilities is not enough: regulation must continuously map the interaction of these capabilities with wider systems of institutions and incentives.[63] Context is not only important to assessing risk, but is often also necessary to adjudicate tradeoffs between risk and reward [149, p. 7].

Risk can also be viewed counterfactually. For example, whether a given capability is already widely available matters. A frontier AI model's capabilities should only be considered dangerous if access to them significantly increases the risk of harm relative to what was attainable without access to the model. If information on how to make a type of weapon is already easily accessible, then the effect of a model should be evaluated with reference to the ease of making such weapons without access to the model.[64]

Risk assessments should also account for possible defenses. As society's capability to manage risks from AI improves, the riskiness of individual AI models may decrease.[65] Indeed, one of the primary uses of safe frontier AI models could be making society more robust to harms from AI and other emerging technologies [253, 254, 255, 240, 61, 98, 32]. Deploying them asymmetrically for beneficial (including defensive) purposes could improve society overall.

---

[61]This is also somewhat related to the issue of over reliance on AI systems, as discussed in e.g. [241].

[62]See result regarding model "sycophancy" [61].

[63]The UK Government plans to take a "context-based" approach to AI regulation [191]: "we will acknowledge that AI is a dynamic, general purpose technology and that the risks arising from it depend principally on the context of its application". See also the OECD Framework for the Classification of AI Systems [251] and the NIST AI Risk Management Framework [149, p. 1]. See also discussion of evaluation-in-society in [252].

[64]This is the approach used in risk assessments for GPT-4 in its System Card [42].

[65]Similarly, the overall decision on whether to deploy a system should consider not just assessed risk, but also the benefits that responsibly deploying a system could yield.

4.2 Engage External Experts to Apply Independent Scrutiny to Models

Having rigorous external scrutiny applied to AI models,[66] particularly prior to deployment, is important to ensuring that the risks are assessed thoroughly and objectively, complementing internal testing processes, while also providing avenues for public accountability.[67] Mechanisms include third-party audits of risk assessment procedures and outputs[68] [257, 235, 258, 259, 260, 183, 184, 261] and engaging external expert red-teamers, including experts from government agencies[69] [235]These mechanisms could be helpfully applied to AI models overall, not just frontier AI models.

The need for creativity and judgment in evaluations of advanced AI models calls for innovative institutional design for external scrutiny. Firstly, it is important that auditors and red-teamers are sufficiently expert and experienced in interacting with state-of-the-art AI models such that they can exercise calibrated judgment, and can execute on what is often the "art" of eliciting capabilities from novel AI models. Secondly, auditors and red-teamers should be provided with enough access to the AI model (including system-level features that would potentially be made available to downstream users) such that they can conduct wide-ranging testing across different threat models, under close-to-reality conditions as a simulated downstream user.

Thirdly, auditors and red teamers need to be adequately resourced,[70] informed, and granted sufficient time to conduct their work at a risk-appropriate level of rigor, not least due to the risk that shallow audits or red teaming efforts provide a sense of false assurance. Fourthly, it is important that results from external assessments are published or communicated to an appropriate regulator, while being mindful of privacy, proprietary information, and the risks of proliferation. Finally, given the common practice of post-deployment model updates, the external scrutiny process should be structured to allow external parties to quickly assess proposed changes to the model and its context before these changes are implemented.

4.3 Follow Standardized Protocols for how Frontier AI Models can be Deployed Based on their Assessed Risk

The AI model's risk profile should inform whether and how the system is deployed. There should be clear protocols established which define and continuously adjust the mapping between a system's risk profile and the particular deployment rules that should be followed. An example mapping specifically for frontier AI models could go as follows, with concrete examples illustrated in Table 3.

> **No assessed severe risk** If assessments determine that the model's use is incredibly unlikely to pose severe risks to public safety, even assuming substantial post-deployment enhancements, then there should be no need for additional deployment restrictions from frontier AI regulation (although certainly, restrictions from other forms of AI regulation could and should continue to apply).
>
> **No discovered severe risks, but notable uncertainty** In some cases the risk assessment may be notably inconclusive. This could be due to uncertainty around post-deployment enhancement techniques (e.g., new methods for fine-tuning, or chaining a frontier AI model within a larger system) that may enable the same model to present more severe risks. In

---

[66]External scrutiny may also need to be applied to, for example, post-deployment monitoring and broader risk assessments.

[67]In a recent expert survey (N = 51), 98% of respondents somewhat or strongly agreed that AGI labs should conduct third-party model audits and red teaming exercises; 94% thought that labs should increase the level of external scrutiny in proportion to the capabilities of their models; 87% supported third-party governance audits; and 84% agreed that labs should give independent researchers API access to deployed models [148].

[68]This would follow the pattern in industries like finance and construction. In these industries, regulations mandate transparency to external auditors whose sign-off is required for large-scale projects. See [256].

[69]The external scrutiny processes of two leading AI developers are described in [42, 233, 262].

[70]One important resource is sharing of best practices and methods for red teaming and third party auditing.

such cases, it may be appropriate to have additional restrictions on the transfer of model weights to high risk parties, and implement particularly careful monitoring for evidence that new post-deployment enhancements meaningfully increase risk. After some monitoring period (e.g. 12 months), absent clear evidence of severe risks, models could potentially be designated as posing "no severe risk."

Some severe risks discovered, but some safe use-cases When certain uses of a frontier AI model would significantly threaten public safety or global security, the developer should implement state-of-the-art deployment guardrails to prevent such misuse. These may include Know-Your-Customer requirements for external users of the AI model, restrictions to fine-tuning,[71] prohibiting certain applications, restricting deployment to beneficial applications, and requiring stringent post-deployment monitoring. The reliability of such safeguards should also be rigorously assessed. This would be in addition to restrictions that are already imposed via other forms of AI regulation.

Severe risks When an AI model is assessed to pose severe risks to public safety or global security which cannot be mitigated with sufficiently high confidence, the frontier model should not be deployed. The model should be secured from theft by malicious actors, and the AI developer should consider deleting the model altogether. Any further experimentation with the model should be done with significant caution, in close consultation with independent safety experts, and could be subject to regulatory approval.

Of course, additional nuance will be needed. For example, as discussed below, there should be methods for updating a model's classifications in light of new information or societal developments. Procedural rigor and fairness in producing and updating such classifications will also be important.

Assessed Risk to Public Safety and Global Security	Possible Example AI system
No severe risks to public safety	Chatbot that can answer elementary-school-level questions about biology, and some (but not all) high-school level questions.
No discovered severe risks to public safety, but significant uncertainty	A general-purpose personal assistant that displays human-level ability to read and synthesize large bodies of scientific literature, including in biological sciences, but cannot generate novel insights.
Some severe risks to public safety discovered, but some safe use-cases	A general-purpose personal assistant that can help generate new vaccines, but also, unless significant safeguards are implemented, predict the genotypes of pathogens that could escape vaccine-induced immunity.
Severe risks to public safety	A general-purpose personal assistant that is capable of designing and, autonomously, ordering the manufacture of novel pathogens capable of causing a COVID-level pandemic.

Table 3: Examples of AI models which would fall into each risk designation category

---

[71]To ensure that certain dangerous capabilities are not further enhanced.

4.4 Monitor and respond to new information on model capabilities

As detailed above new information about a model's risk profile may arise post-deployment. If that information indicates that the model was or has become more risky than originally assessed, the developer should reassess the deployment, and update restrictions on deployment if necessary.[72]

New information could arise in several ways. Broad deployment of a model may yield new information about the model's capabilities, given the creativity from a much larger number of users, and exposure of the model to a wider array of tools and applications. Post-deployment enhancement techniques — such as fine-tuning [263, 264], prompt engineering [265, 266, 267], and foundation model programs [87, 88, 83] — provide another possible source of new risk-relevant information. The application of these techniques to deployed models could elicit more powerful capabilities than pre-deployment assessments would have ascertained. In some instances, this may meaningfully change the risk profile of a frontier AI model, potentially leading to adjustments in how and whether the model is deployed.[73]

AI developers should stay on top of known and emerging post-deployment enhancement techniques by, e.g., monitoring how users are building on top of their APIs and tracking publications about new methods. Given up to date knowledge of how deployed AI models could be enhanced, prudent practices could include:

- Regularly (e.g., every 3 months) repeating a lightweight version of the risk assessment on deployed AI models, accounting for new post-deployment enhancement techniques.

- Before pushing large updates[74] to deployed AI models, repeating a lightweight risk assessment.

- Creating pathways for incident reporting [187] and impact monitoring to capture post-deployment incidents for continuous risk assessment.

- If these repeat risk assessments result in the deployed AI model being categorized at a different risk level (as per the taxonomy above) , promptly updating deployment guardrails to reflect the new risk profile.

- Having the legal and technical ability to quickly roll back deployed models on short notice if the risks warrant it, for example by not open-sourcing models until doing so appears sufficiently safe.[75]

4.5 Additional practices

Parts of the aforementioned standards can suitably be applied to current AI systems, not just frontier AI systems. Going forward, frontier AI systems seem likely to warrant more tailored safety standards, given the level of prospective risk that they pose. Examples of such standards include:[76]

---

[72]In a recent expert survey (N = 51), 98% of respondents somewhat or strongly agreed that AGI labs should closely monitor deployed systems, including how they are used and what impact they have on society; 97% thought that they should continually evaluate models for dangerous capabilities after deployment, taking into account new information about the model's capabilities and how it is being used; and 93% thought that labs should pause the development process if sufficiently dangerous capabilities are detected [148].

[73]Such updates may only be possible if the model has not yet proliferated, e.g. if it is deployed via an API. The ability to update how a model is made available after deployment is one key reason to employ staged release of structured access approaches [109, 110].

[74]This would need to be defined more precisely.

[75]Note that this may have implications for the kinds of use cases a system built on a frontier AI model can support. Use cases in which quick roll-back itself poses risks high enough to challenge the viability of roll-back as an option should be avoided, unless robust measures are in place to prevent such failure modes.

[76]This would need to be defined more precisely.

- Avoid large jumps in the capabilities of models that are trained and deployed. Standards could specify "large jumps" in terms of a multiplier on the amount of computing power used to train the most compute-intensive "known to be safe" model to date, accounting for algorithmic efficiency improvements.

- Adopt state-of-the-art alignment techniques for training new frontier models which could suitably guard against models potentially being situationally aware and deceptive [187].

- Prior to beginning training of a new model, use empirical approaches to predict capabilities of the resultant model, including experiments on small-scale versions of the model, and take preemptive actions to avoid training models with dangerous capabilities and/or to otherwise ensure training proceeds safely (e.g. introduce more frequent model evaluation checkpoints; conditioning beginning training on certain safety and security milestones).

- Adopt internal governance practices to adequately identify and respond to the unique nature of the risks presented by frontier AI development. Such practices could take inspiration from practices in Enterprise Risk Management, such as setting up internal audit functions [268, 192].

- Adopt state-of-the-art security measures to protect frontier AI models.

5 Uncertainties and Limitations

We think that it is important to begin taking practical steps to regulate frontier AI today, and that the ideas discussed in this paper are a step in that direction. Nonetheless, stress testing and developing these ideas, and offering alternatives, will require broad and diverse input. In this section, we list some of our main uncertainties (as well as areas of disagreement between the paper's authors) where we would particularly value further discussion.

First, there are several assumptions that underpin the case for a regulatory regime like the one laid out in this paper, which would benefit from more scrutiny:

How should we define frontier AI for the purposes of regulation? We focus in this paper on tying the definition of frontier AI models to the potential of dangerous capabilities sufficient to cause severe harm, in order to ensure that any regulation is clearly tied to the policy motivation of ensuring public safety. However, there are also downsides to this way of defining frontier AI — most notably, that it requires some assessment of the likelihood that a model possesses dangerous capabilities before deciding whether it falls in the scope of regulation, which may be difficult to do. An alternative, which some authors of this paper prefer, would be to define frontier AI development as that which aims to develop novel and broad AI capabilities — i.e. development pushing at the "frontier" of AI capabilities. This would need further operationalization — for example, defining these as models which use more training compute than already-deployed systems — but could offer an approach to identify the kinds of development activities that fall within the scope of regulation without first needing to make an assessment of dangerous capabilities. We discuss the pros and cons of different definitions of frontier AI in appendix A, and would love to receive feedback and engage in further discussion on this point.

How dangerous are and will the capabilities of advanced foundation AI models be, and how soon could these capabilities arise? It is very difficult to predict the pace of AI development and the capabilities that could emerge in advance; indeed, we even lack certainty about the capabilities of existing systems. Assumptions here affect the urgency of regulatory action. There is a challenging balance to strike here between getting regulatory infrastructure in place early enough to address and mitigate or prevent the biggest risks, while waiting for enough information about what those risks are likely to be and how they can be mitigated [269].

Will training advanced AI models continue to require large amounts of resources? The regulatory ecosystem we discuss partly relies on an assumption that highly capable foundation models will require large amounts of resources to develop. That being the case makes it easier to regulate frontier AI. Should frontier AI models be possible to create using resources available to millions of actors rather than a handful, that may lead to significant changes to the best regulatory approach. For example, it might suggest that more efforts should be put into regulating the use of these models and to protect against (rather than to stop) dangerous uses of frontier AI.

How effectively can we anticipate and mitigate risks from frontier AI? A core argument of this paper is that an anticipatory approach to governing AI will be important, but effectively identifying risks anticipatorily is far from straightforward. We would value input on the effectiveness of different risk assessment methods for doing this, drawing lessons from other domains where anticipatory approaches are used.

30

How can regulatory flight be avoided? A regulatory regime for frontier AI could prove counterproductive if it incentivises AI companies to move their activities to jurisdictions with less onerous rules. One promising approach is having rules apply to what models people in some jurisdiction can engage with, as people are unlikely to move to a different jurisdiction to access different models and as companies are incentivised to serve them their product. Scholars have suggested that dynamics like these have led to a "California Effect" and a "Brussels Effect," where Californian and EU rules are voluntarily complied with beyond their borders.

To what extent will it be possible to defend against dangerous capabilities? Assessments of what constitutes "sufficiently dangerous capabilities," and what counter-measures are appropriate upon finding them in a model, hinges significantly on whether future AI models will be more beneficial for offense versus defense.

Second, we must consider ways that this kind of regulatory regime could have unintended negative consequences, and take steps to guard against them. These include:

Reducing beneficial innovation All else being equal, any imposition of costs on developers of new technologies slows the rate of innovation, and any regulatory measures come with compliance costs. However, these costs should be weighed against the costs of unfettered development and deployment, as well as impacts on the rate of innovation from regulatory uncertainty and backlash due to unmitigated societal harms. On balance, we tentatively believe that the proposed regulatory approaches can support beneficial innovation by focusing on a targeted subset of AI systems, and by addressing issues upstream in a way that makes it easier for smaller companies to develop innovative applications with confidence.

Causing centralization of power in AI development Approaches like a licensing regime for developers could have the effect of centralizing power with the companies licensed to develop the most capable AI systems. It will be important to ensure that the regulatory regime is complemented with the power to identify and intervene to prevent abuses of market dominance,[77] and government support for widespread access to AI systems deemed to be low risk and high benefit for society.

Enabling abuse of government powers A significant aim of regulation is to transfer power from private actors to governments who are accountable to the public. However, the power to constrain the development and deployment of frontier AI models is nonetheless a significant one with real potential for abuse at the hand of narrow political interests, as well as corrupt or authoritarian regimes. This is a complex issue which requires thorough treatment of questions such as: where should the regulatory authority be situated, and what institutional checks and balances should be put in place, to reduce these risks?; what minimum regulatory powers are needed to be effective?; and what international dialogue is needed to establish norms?

Risk of regulatory capture As the regulation of advanced technologies often requires access to expertise from the technological frontier, and since the frontier is often occupied by private companies, there is an ongoing risk that regulations informed by private-sector expertise will be biased towards pro-industry positions, to the detriment of society. This should be mitigated by designing institutions that can limit and challenge the influence of private interests, and by seeking detailed input from academia and civil society before beginning to implement any of these proposals.

---

[77] Such as, for example, the UK's review of competition law as it relates to the market for foundation models [270].

Finally, there are many practical details of implementation not covered in this paper that will need to be worked out in detail with policy and legal professionals, including:

What the appropriate regulatory authority/agency would be in different jurisdictions, where new bodies or powers might be required, and the tradeoffs of different options.

How this kind of regulation will relate to other AI regulation and governance proposals and how it can best support and complement attempts to address other parts of AI governance. Our hope is that by intervening early in the AI lifecycle, the proposed regulation can have many downstream benefits, but there are also many risks and harms that this proposal will not address. We hope to contribute to wider conversations about what a broader regulatory ecosystem for AI should look like, of which these proposals form a part.

Steps towards international cooperation and implementation of frontier AI regulation, including how best to convene international dialogue on this topic, who should lead these efforts, and what possible international agreements could look like. An important part of this will be considering what is best implemented domestically, at least initially, and where international action is needed.

Conclusion

In the absence of regulation, continued rapid development of highly capable foundation models may present severe risks to public safety and global security. This paper has outlined possible regulatory approaches to reduce the likelihood and severity of these risks while also enabling beneficial AI innovation.

Governments and regulators will likely need to consider a broad range of approaches to regulating frontier AI. Self-regulation and certification for compliance with safety standards for frontier AI could be an important step. However, government intervention will be needed to ensure sufficient compliance with standards. Additional approaches include mandates and enforcement by a supervisory authority, and licensing the deployment and potentially the development of frontier AI models.

Clear and concrete safety standards will likely be the main substantive requirements of any regulatory approach. AI developers and AI safety researchers should, with the help of government actors, invest heavily to establish and converge on risk assessments, model evaluations, and oversight frameworks with the greatest potential to mitigate the risks of frontier AI, and foundation models overall. These standards should be reviewed and updated regularly.

As global leaders in AI development and AI safety, jurisdictions such as the United States or United Kingdom could be natural leaders in implementing the regulatory approaches described in this paper. Bold leadership could inspire similar efforts across the world. Over time, allies and partners could work together to establish an international governance regime[78] for frontier AI development and deployment that both guards against collective downsides and enables collective progress.[79]

Uncertainty about the optimal regulatory approach to address the challenges posed by frontier AI models should not impede immediate action. Establishing an effective regulatory regime is a time-consuming process, while the pace of progress in AI is rapid. This makes it crucial for policymakers, researchers, and practitioners to move fast and rigorously explore what regulatory approaches may work best. The complexities of AI governance demand our best collective efforts. We hope that this paper is a small step in that direction.

---

[78]Or build on existing institutions.

[79]This international regime could take various forms. Possibilities include an international standard-setting organization, or trade agreements focused on enabling trade in AI goods and services that adhere to safety standards. Countries that lead in AI development could subsidize access to and adoption of AI in developing nations in return for assistance in managing risks of proliferation, as has been done with nuclear technologies.

Appendix A   Creating a Regulatory Definition for Frontier AI

In this paper, we use the term "frontier AI" models to refer to highly capable foundation models for which there is good reason to believe could possess dangerous capabilities sufficient to pose severe risks to public safety ("sufficiently dangerous capabilities"). Any binding regulation of frontier AI, however, would require a much more precise definition. Such a definition would also be an important building block to the creation and dissemination of voluntary standards.

This section attempts to lay out some desiderata and approaches to creating such a regulatory definition. It is worth noting up front that what qualifies as a frontier AI model changes over time — this is a dynamic category. In particular, what may initially qualify as a frontier AI model could change over time due to improvements in society's defenses against advanced AI models and an improved understanding of the nature of the risks posed. On the other hand, factors such as improvements in algorithmic efficiency would decrease the amount of computational resources required to develop models, including those with sufficiently dangerous capabilities.

While we do not yet have confidence in a specific, sufficiently precise regulatory definition, we are optimistic that such definitions are possible. Overall, none of the approaches we describe here seem fully satisfying. Additional effort towards developing a better definition would be high-valuable.

A.1   Desiderata for a Regulatory Definition

In addition to general desiderata for a legal definition of regulated AI models,[80] a regulatory definition should limit its scope to only those models for which there is good reason to believe they have sufficiently dangerous capabilities. Because regulation could cover development in addition to deployment, it should be possible to determine whether a planned model will be regulated ex ante, before the model is developed. For example, the definition could be based on the model development process that will be used (e.g., data, algorithms, and compute), rather than relying on ex post features of the completed model (e.g., capabilities, performance on evaluations).

A.2   Defining Sufficiently Dangerous Capabilities

"Sufficiently dangerous capabilities" play an important role in our concept of frontier AI: we only want to regulate the development of models that could cause such serious harms that ex post remedies will be insufficient.

Different procedures could be used to develop a regulatory definition of "sufficiently dangerous capabilities." One approach could be to allow an expert regulator to create a list of sufficiently dangerous capabilities, and revise that list over time in response to changing technical and societal circumstances. This approach has the benefit of enabling greater learning and improvement over time, though it leaves the challenge outstanding of defining what model development activities are covered ex ante, and could in practice be very rigid and unsuited to the rapid pace of AI progress. Further, there is a risk that regulators will define such capabilities more expansively over time, creating "regulatory creep" that overburdens AI development.

---

[80]According to [271], legal definitions should neither be *over-inclusive* (i.e. they should not include cases which are not in need of regulation according to the regulation's objectives) nor *under-inclusive* (i.e. they should not exclude cases which should have been included). Instead, legal definitions should be *precise* (i.e. it must be possible to determine clearly whether or not a particular case falls under the definition), *understandable* (i.e. at least in principle, people without expert knowledge should be able to apply the definition), *practicable* (i.e. it should be possible to determine with little effort whether or not a concrete case falls under the definition), and *flexible* (i.e. they should be able to accommodate technical progress). See also [272, p. 70].

Legislatures could try to prevent such regulatory creep by describing factors that should be considered when making a determination that certain capabilities would be sufficiently dangerous. This is common in United States administrative law.[81] One factor that could be considered is whether a capability would pose a "severe risk to public safety," assessed with reference to the potential scale and estimated probability of counterfactual harms caused by the system. A scale similar to the one used in the UK National Risk Register could be used [273]. One problem with this approach is that making these estimates will be exceedingly difficult and contentious.

A.3   Defining Foundation Models

The seminal report on foundation models [15] defines them as "models ... trained on broad data ... that can be adapted to a wide range of downstream tasks." This definition, and various regulatory proposals based on it, identify two key features that a regulator could use to separate foundation models from narrow models: breadth of training data, and applicability to a wide range of downstream tasks.

Breadth is hard to define precisely, but one attempt would be to say that training data is "broad" if it contains data on many economically or strategically useful tasks. For example, broad natural language corpora, such as CommonCrawl [274], satisfy this requirement. Narrower datasets, such as weather data, do not. Similarly, certain well-known types of models, such as large language models (LLMs) are clearly applicable to a variety of downstream tasks. A model that solely generates music, however, has a much narrower range of use-cases.

Given the vagueness of the above concepts, however, they may not be appropriate for a regulatory definition. Of course, judges and regulators do often adjudicate vague concepts [275], but we may be able to improve on the above. For example, a regulator could list out types of model architectures (e.g., transformer-based language models) or behaviors (e.g., competently answering questions about many topics of interest) that a planned model could be expected to capable of, and say that any model that has these features is a foundation model.

Overall, none of these approaches seem fully satisfying. Additional effort towards developing a better definition of foundational models—or of otherwise defining models with broad capabilities—would be high-value.

A.4   Defining the Possibility of Producing Sufficiently Dangerous Capabilities

A regulator may also have to define AI development processes that could produce broadly capable models with sufficiently dangerous capabilities.

At present, there is no rigorous method for reliably determining, ex ante, whether a planned model will have broad and sufficiently dangerous capabilities. Recall the Unexpected Capabilities Problem: it is hard to predict exactly when any specific capability will arise in broadly capable models. It also does not appear that any broadly capable model to-date possesses sufficiently dangerous capabilities.

In light of this uncertainty, we do not have a definite recommendation. We will, however, note several options.

One simple approach would be to say that any foundation model that is trained with more than some amount of computational power—for example, 10^{26} FLOP—has the potential to show sufficiently dangerous capabilities. As Appendix B demonstrates, FLOP usage empirically correlates with breadth and depth of capabilities in foundation models. There is therefore good reason to think that FLOP usage is correlated with the likelihood that a broadly capable model will have sufficiently dangerous capabilities.

---

[81]See, e.g., 42 U.S.C. § 262a(a)(1)(B).

A threshold-based approach like this has several virtues. It is very simple, objective, determinable *ex ante*,[82] and (due to the high price of compute) is correlated with the ability of the developer to pay compliance costs. One drawback, however, is that the same number of FLOP will produce greater capabilities over time due to algorithmic improvements [276]. This means that, all else equal, the probability that a foundation model below the threshold will have sufficiently dangerous capabilities will increase over time. These problems may not be intractable. For example, a FLOP threshold could formulaically decay over time based on new models' performance on standardized benchmarks, to attempt to account for anticipated improvements in algorithmic efficiency.[83]

A related approach could be to define the regulatory target by reference to the most capable broad models that have been shown not to have sufficiently dangerous capabilities. The idea here is that, if a model has been shown not to have sufficiently dangerous capabilities, then every model that can be expected to perform worse than it should also not be expected to have sufficiently dangerous capabilities. Regulation would then apply only to those models that exceed the capabilities of models known to lack sufficiently dangerous capabilities. This approach has the benefit of updating quickly based on observations from newer models. It would also narrow the space of regulated models over time, as regulators learn more about which models have sufficiently dangerous capabilities.

However, this definition has significant downsides too. First, there are many variables that could correlate with possession of dangerous capabilities, which means that it is unclear ex ante which changes in development processes could dramatically change capabilities. For example, even if model A dominates model B on many obvious aspects of its development (e.g., FLOP usage, dataset size), B may dominate A on other important aspects, such as use of a new and more efficient algorithm, or a better dataset. Accordingly, the mere fact that a B is different from A may be enough to make B risky,[84] unless the regulator can carefully discriminate between trivial and risk-enhancing differences. The information needed to make such a determination may also be highly sensitive and difficult to interpret. Overall, then, determining whether a newer model can be expected to perform better than a prior known-safe model is far from straightforward.

Another potential problem with any compute-based threshold is that models below it could potentially be open-sourced and then further trained by another actor, taking its cumulative training compute above the threshold. One possible solution to this issue could be introducing minimal requirements regarding the open-sourcing of models trained using one or two orders of magnitude of compute less than any threshold set.

Given the uncertainty surrounding model capabilities, any definition will likely be overinclusive. However, we emphasize the importance of creating broad and clear ex ante exemptions for models that have no reasonable probability of possessing dangerous capabilities. For example, an initial blanket exemption for models trained with fewer than (say) 1E26 FLOP[85] could be appropriate, to remove any doubt as to whether such models are covered. Clarity and definitiveness of such exemptions is crucial to avoid overburdening small and academic developers, whose models will likely contribute very little to overall risk.

---

[82]At least, determinable from the planned specifications of the training run of an AI model, though of course final FLOP usage will not be determined until the training run is complete. However, AI developers tend to carefully plan the FLOP usage of training runs for both technical and financial reasons.

[83]As an analogy, many monetary provisions in US law are adjusted for inflation based on a standardized measure like the consumer price index.

[84]Compare the definition of "frontier AI" used in [25]: "models that are both (a) close to, or exceeding, the average capabilities of the most capable existing models, and (b) different from other models, either in terms of scale, design (e.g. different architectures or alignment techniques), or their resulting mix of capabilities and behaviours..."

[85]Using public FLOP per dollar estimates contained in [277] (Epoch AI) and [278], this would cost nearly or more than $100 million in compute alone.
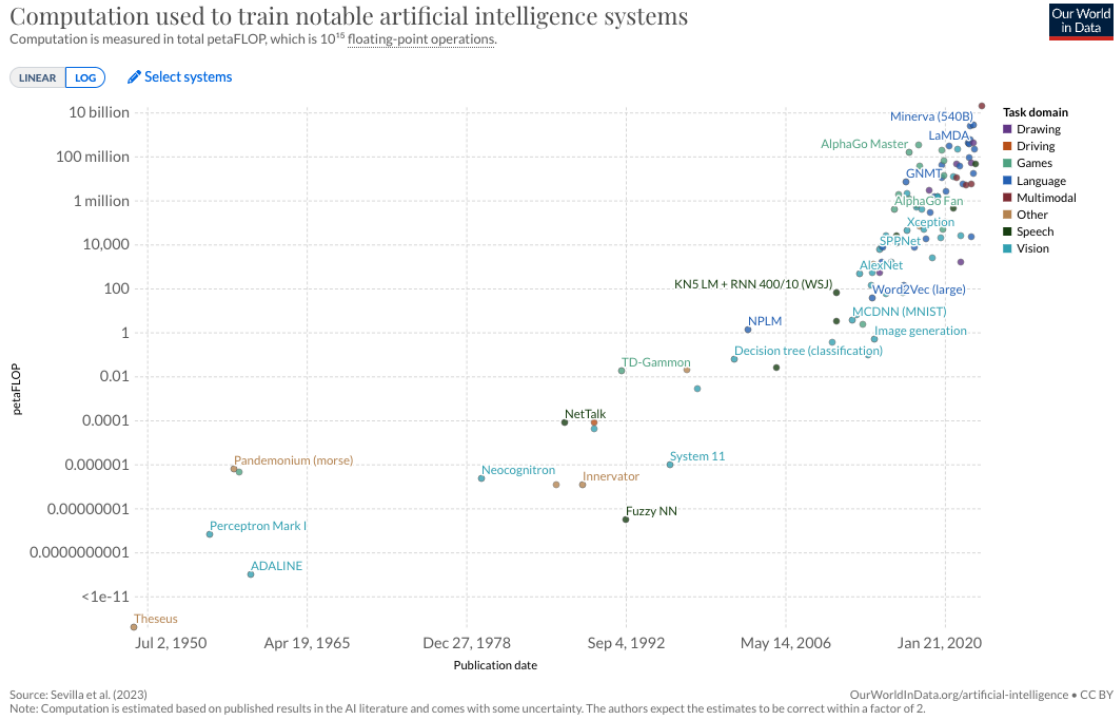
Figure 4: Computation used to train notable AI systems. Note logarithmic y-axis. Source: [50] based on data from [280].

Appendix B   Scaling laws in Deep Learning

This appendix describes results from the scaling laws literature which shape the regulatory challenge posed by frontier AI as well as the available regulatory options. This literature focuses on relationships between measures of model performance (such as test loss) and properties of the model training process (such as amounts of data, parameters, and compute). Results from this literature of particular relevance to this paper include: (i) increases in the amount of compute used to train models has been an important contributor to AI progress; (ii) even if the increase in compute starts contributing less to progress, we still expect frontier AI models to be trained using large amounts of compute; (iii) though scale predictably increases model performance on the training objective, particular capabilities may improve or change unexpectedly, contributing to the Unexpected Capabilities Problem.

In recent years, the Deep Learning Revolution has been characterized by the considerable scaling up of the key inputs into neural networks, especially the quantity of computations used to train a deep learning system ("compute") [279].

Empirically, scaling training compute has reliably led to better performance on many of the tasks AI models are trained to solve, and many similar downstream tasks [58]. This is often referred to as the "Scaling Hypothesis": the expectation that scale will continue to be a primary predictor and determinant of model capabilities, and that scaling existing and foreseeable AI techniques will continue to produce many capabilities beyond the reach of current systems.[86]

---

[86]See [281, 282, 279, 15]. For a skeptical take on the Scaling Hypothesis, see [278].
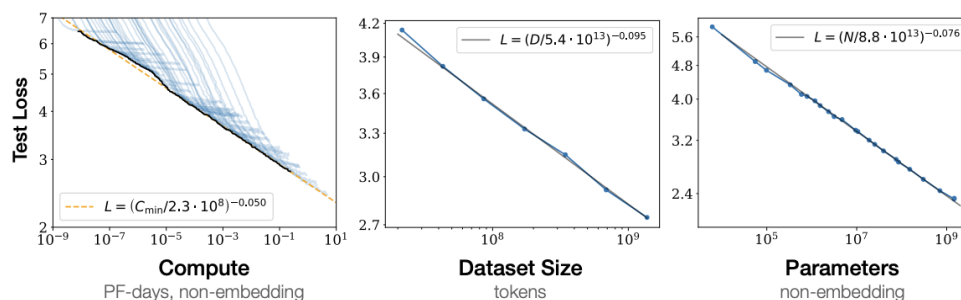
Figure 5: Scaling reliably leading to lower test loss. See [56]. The scaling laws from this paper have been updated by [45].

We expect the Scaling Hypothesis to account for a significant fraction of progress in AI over the coming years, driving increased opportunities and risks. However, the importance of scaling for developing more capable systems may decrease with time, as per research which shows that the current rate of scaling may be unsustainable [278, 283, 103].

Even if increases in scale slow down, the most capable AI models are still likely going to be those that can effectively leverage large amounts of compute, a claim often termed "the bitter lesson" [282]. Specifically, we expect frontier AI models to use vast amounts of compute, and that increased algorithmic efficiency [284] and data quality [285] will continue to be important drivers of AI progress.

Scaling laws have other limits. Though scaling laws can reliably predict the loss of a model on its training objective – such as predicting the next word in a piece of text – that is currently an unreliable predictor of downstream performance on individual tasks. For example, tasks can see inverse scaling, where scaling leads to worse performance [60, 61, 62], though further scaling has overturned some of these findings [36].

Model performance on individual tasks can also increase unexpectedly: there may be "emergent capabilities" [286, 67]. Some have argued that such emergent capabilities are a "mirage" [67]. They argue that the emergence of capabilities is primarily a consequence of how they are measured. Using discontinuous measures such as multiple choice answers or using an exact string match, is more likely to "find" emergent capabilities than if using continuous measures – for example, instead of measuring performance by exact string match, you measure it based on proximity to the right answer.

We do not think this analysis comprehensively disproves the emergent capabilities claim [66]. Firstly, discontinuous measures are often what matter. For autonomous vehicles, what matters is how often they cause a crash. For an AI model solving mathematics questions, what matters is whether it gets the answer exactly right or not. Further, even if continuous "surrogate" measures could be used to predict performance on the discontinuous measures, the appropriate choice of a continuous measure that will accurately predict the true metric is often unknown a priori. Such forecasts instead presently require a subjective choice between many possible alternatives, which would lead to different predictions on the ultimate phenomenon. For instance, is an answer to a mathematical question "less wrong" if it's numerically closer to the actual answer, or if a single operation, such as multiplying instead of dividing, led to an incorrect result?

Nevertheless, investing in further research to more accurately predict capabilities of AI models ex ante is a crucial enabler for effectively targeting policy interventions, using scaling laws or otherwise.

References

[1] Michael Moor et al. "Foundation models for generalist medical artificial intelligence". In: *Nature* 616.7956 (Apr. 2023), pp. 259–265. DOI: 10.1038/s41586-023-05881-4.

[2] Peter Lee, Sebastien Bubeck, and Joseph Petro. "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine". In: *New England Journal of Medicine* 388.13 (Mar. 2023). Ed. by Jeffrey M. Drazen, Isaac S. Kohane, and Tze-Yun Leong, pp. 1233–1239. DOI: 10.1056/nejmsr2214184.

[3] Karan Singhal et al. *Large Language Models Encode Clinical Knowledge*. 2022. arXiv: 2212.13138 [cs.CL].

[4] Harsha Nori et al. *Capabilities of GPT-4 on Medical Challenge Problems*. 2023. arXiv: 2303.13375 [cs.CL].

[5] Drew Simshaw. "Access to A.I. Justice: Avoiding an Inequitable Two-Tiered System of Legal Services". In: *SSRN Electronic Journal* (2022).

[6] Yonathan A. Arbel and Shmuel I. Becher. "Contracts in the Age of Smart Readers". In: *SSRN Electronic Journal* (2020). DOI: 10.2139/ssrn.3740356.

[7] Noam Kolt. "Predicting Consumer Contracts". In: *Berkeley Technology Law Journal* 37.1 (2022).

[8] Sal Khan. *Harnessing GPT-4 so that all students benefit*. 2023. URL: https://perma.cc/U54W-SSGA.

[9] David Rolnick et al. *Tackling Climate Change with Machine Learning*. 2019. arXiv: 1906.05433 [cs.CY].

[10] DeepMind. *DeepMind AI Reduces Google Data Centre Cooling Bill by 40%*. 2016. URL: https://perma.cc/F4B2-DFZ9.

[11] Huseyin Tuna Erdinc et al. *De-risking Carbon Capture and Sequestration with Explainable CO_2 Leakage Detection in Time-lapse Seismic Monitoring Images*. 2022. arXiv: 2212.08596 [physics.geo-ph].

[12] Priya L. Donti and J. Zico Kolter. "Machine Learning for Sustainable Energy Systems". In: *Annual Review of Environment and Resources* 46.1 (Oct. 2021), pp. 719–747. DOI: 10.1146/annurev-environ-020220-061831.

[13] Panagiota Galetsi, Korina Katsaliaki, and Sameer Kumar. "The medical and societal impact of big data analytics and artificial intelligence applications in combating pandemics: A review focused on Covid-19". In: *Social Science & Medicine* 301 (May 2022), p. 114973. DOI: 10.1016/j.socscimed.2022.114973.

[14] David C. Danko et al. *The Challenges and Opportunities in Creating an Early Warning System for Global Pandemics*. 2023. arXiv: 2302.00863 [q-bio.QM].

[15] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG].

[16] Fabio Urbina et al. "Dual use of artificial-intelligence-powered drug discovery". In: *Nature Machine Intelligence* 4.3 (Mar. 2022), pp. 189–191. DOI: 10.1038/s42256-022-00465-9.

[17] Richard Ngo, Lawrence Chan, and Sören Mindermann. *The alignment problem from a deep learning perspective*. 2023. arXiv: 2209.00626 [cs.AI].

[18] Michael K. Cohen, Marcus Hutter, and Michael A. Osborne. "Advanced artificial agents intervene in the provision of reward". In: *AI Magazine* 43.3 (Aug. 2022), pp. 282–293. DOI: 10.1002/aaai.12064.

[19] Dan Hendrycks et al. *Unsolved Problems in ML Safety*. 2022. arXiv: 2109.13916 [cs.LG].

[20]   Dan Hendrycks and Mantas Mazeika. *X-Risk Analysis for AI Research*. 2022. arXiv: 2206.05862 [cs.CY].

[21]   Joseph Carlsmith. *Is Power-Seeking AI an Existential Risk?* 2022. arXiv: 2206.13353 [cs.CY].

[22]   Stuart J. Russell. *Human Compatible. Artificial Intelligence and the Problem of Control*. Viking, 2019.

[23]   Brian Christian. *The Alignment Problem. Machine Learning and Human Values*. W. W. Norton & Company, 2020.

[24]   Brando Benifei and Ioan-Dragoş Tudorache. *Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. 2023. URL: https://perma.cc/VH4R-WV3G.

[25]   Toby Shevlane et al. *Model evaluation for extreme risks*. 2023. arXiv: 2305.15324 [cs.AI].

[26]   Remco Zwetsloot and Allan Dafoe. *Thinking About Risks From AI: Accidents, Misuse and Structure*. 2019. URL: https://perma.cc/7UQ8-3Z2R.

[27]   Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. *Emergent autonomous scientific research capabilities of large language models*. 2023. arXiv: 2304.05332 [physics.chem-ph].

[28]   Eric Horvitz. *On the Horizon: Interactive and Compositional Deepfakes*. 2022. arXiv: 2209.01714 [cs.LG].

[29]   Josh A. Goldstein et al. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. 2023. arXiv: 2301.04246 [cs.CY].

[30]   Ben Buchanan et al. *Truth, Lies, and Automation: How Language Models Could Change Disinformation*. 2021. URL: https://perma.cc/V5RP-CQG7.

[31]   Russell A Poldrack, Thomas Lu, and Gašper Beguš. *AI-assisted coding: Experiments with GPT-4*. 2023. arXiv: 2304.13187 [cs.AI].

[32]   Andrew J. Lohn and Krystal A. Jackson. *Will AI Make Cyber Swords or Shields?* 2022. URL: https://perma.cc/3KTH-GQTG.

[33]   Microsoft. *What are Tokens?* 2023. URL: https://perma.cc/W2H8-FKDU.

[34]   Alec Radford et al. *Language Models are Unsupervised Multitask Learners*. 2019.

[35]   Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[36]   OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[37]   Aakanksha Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. 2022. arXiv: 2204.02311 [cs.CL].

[38]   Jean-Baptiste Alayrac et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. arXiv: 2204.14198 [cs.CV].

[39]   Reponsible AI Licenses Team. *Reponsible AI Licenses*. 2023. URL: https://perma.cc/LYQ8-V5X2.

[40]   open source initiative. *The Open Source Definition*. 2007. URL: https://perma.cc/WU4B-DHWF.

[41]   Emily M. Bender et al. "On the Dangers of Stochastic Parrots". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Mar. 2021. DOI: 10.1145/3442188.3445922.

[42]   OpenAI. *GPT-4 System Card*. 2023. URL: https://perma.cc/TJ3Z-Z3YY.

[43]   Jacob Steinhardt. *AI Forecasting: One Year In*. 2023. URL: https://perma.cc/X4WY-N8QY.

[44]   Baobao Zhang et al. *Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers*. 2022. arXiv: 2206.04132 [cs.CY].

[45]   Jordan Hoffmann et al. *Training Compute-Optimal Large Language Models*. 2022. arXiv: 2203.15556 [cs.CL].

[46] Bryan Caplan. *GPT-4 Takes a New Midterm and Gets an A.* 2023. URL: https://perma.cc/2SPU-DRK3.

[47] Bryan Caplan. *GPT Retakes My Midterm and Gets an A.* 2023. URL: https://perma.cc/DG6F-WW8J.

[48] Metaculus. *In 2016, will an AI player beat a professionally ranked human in the ancient game of Go?* 2016. URL: https://perma.cc/NN7L-58YB.

[49] Metaculus. *When will programs write programs for us?* 2021. URL: https://perma.cc/NM5Y-27RB.

[50] Our World in Data. *Computation used to train notable artificial intelligence systems.* 2023. URL: https://perma.cc/59K8-WXQA.

[51] Minister of Innovation, Science and Industry. *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts.* 2021. URL: https://perma.cc/ZT7V-A2Q8.

[52] Yvette D. Clarke. *Algorithmic Accountability Act of 2022.* US Congress. 2022. URL: https://perma.cc/99S2-AH9G.

[53] U.S. Food and Drug Administration. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.* 2021. URL: https://perma.cc/Q3PP-SDU8.

[54] Consumer Financial Protection Bureau. *CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms.* 2022. URL: https://perma.cc/59SX-GGZN.

[55] Lina Khan. *We Must Regulate A.I.: Here's How.* New York Times. 2023. URL: https://perma.cc/4U6B-E7AV.

[56] Jared Kaplan et al. *Scaling Laws for Neural Language Models.* 2020. arXiv: 2001.08361 [cs.LG].

[57] Tom Henighan et al. *Scaling Laws for Autoregressive Generative Modeling.* 2020. arXiv: 2010.14701 [cs.LG].

[58] Pablo Villalobos. *Scaling Laws Literature Review.* 2023. URL: https://perma.cc/32GJ-FBGM.

[59] Joel Hestness et al. *Deep Learning Scaling is Predictable, Empirically.* 2017. arXiv: 1712.00409 [cs.LG].

[60] Ian R. McKenzie et al. *Inverse Scaling: When Bigger Isn't Better.* 2023. arXiv: 2306.09479 [cs.CL].

[61] Ethan Perez et al. *Discovering Language Model Behaviors with Model-Written Evaluations.* 2022. arXiv: 2212.09251 [cs.CL].

[62] Philipp Koralus and Vincent Wang-Maścianica. *Humans in Humans Out: On GPT Converging Toward Common Sense in both Success and Failure.* 2023. arXiv: 2303.17276 [cs.AI].

[63] Jason Wei et al. *Emergent Abilities of Large Language Models.* 2022. arXiv: 2206.07682 [cs.CL].

[64] Jason Wei. *137 emergent abilities of large language models.* 2022. URL: https://perma.cc/789W-4AZQ.

[65] Samuel R. Bowman. *Eight Things to Know about Large Language Models.* 2023. arXiv: 2304.00612 [cs.CL].

[66] Jason Wei. *Common arguments regarding emergent abilities.* 2023. URL: https://perma.cc/F48V-XZHC.

[67] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. *Are Emergent Abilities of Large Language Models a Mirage?* 2023. arXiv: 2304.15004 [cs.AI].

[68] Anthropic. *Claude: A next-generation AI assistant for your tasks, no matter the scale.* 2023. URL: https://www.anthropic.com/product.

[69]   OpenAI. *Fine-tuning: Learn how to customize a model for your application.* 2023. URL: `https://perma.cc/QX2L-752C`.

[70]   AI21 Labs. *AI21 Studio.* 2023. URL: `https://perma.cc/9VSK-P5W7`.

[71]   Cohere. *Training Custom Models.* 2023. URL: `https://perma.cc/M2MD-TTKR`.

[72]   Steven C. H. Hoi et al. *Online Learning: A Comprehensive Survey.* 2018. arXiv: `1802.02871` `[cs.LG]`.

[73]   German I. Parisi et al. "Continual lifelong learning with neural networks: A review". In: *Neural Networks* 113 (May 2019), pp. 54–71. DOI: `10.1016/j.neunet.2019.01.012`.

[74]   Gerrit De Vynck, Rachel Lerman, and Nitasha Tiku. *Microsoft's AI chatbot is going off the rails.* 2023. URL: `https://www.washingtonpost.com/technology/2023/02/16/microsoft-bing-ai-chatbot-sydney/`.

[75]   OpenAI. *Our approach to AI safety.* 2023. URL: `https://perma.cc/7GS3-KHVV`.

[76]   Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* 2023. arXiv: `2201.11903 [cs.CL]`.

[77]   Jack Clark. *Import AI 310: AlphaZero learned Chess like humans learn Chess; capability emergence in language models; demoscene AI.* 2022. URL: `https://perma.cc/K4FG-ZXMX`.

[78]   Jessica Rumbelow. *SolidGoldMagikarp (plus, prompt generation).* 2023. URL: `https://www.alignmentforum.org/posts/aPeJE8bSo6rAFoLqg/solidgoldmagikarp-plus-prompt-generation`.

[79]   OpenAI. *ChatGPT plugins.* 2022. URL: `https://perma.cc/3NPU-HUJP`.

[80]   Timo Schick et al. *Toolformer: Language Models Can Teach Themselves to Use Tools.* 2023. arXiv: `2302.04761 [cs.CL]`.

[81]   Tianle Cai et al. *Large Language Models as Tool Makers.* 2023. arXiv: `2305.17126 [cs.LG]`.

[82]   Adept. *ACT-1: Transformer for Actions.* 2022. URL: `https://perma.cc/7EN2-256H`.

[83]   Significant Gravitas. *Auto-GPT: An Autonomous GPT-4 Experiment.* 2023. URL: `https://perma.cc/2TT2-VQE8`.

[84]   Shehel Yoosuf and Yin Yang. "Fine-Grained Propaganda Detection with Fine-Tuned BERT". In: *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda.* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 87–91. DOI: `10.18653/v1/D19-5011`. URL: `https://perma.cc/5CWN-HTU2`.

[85]   Takeshi Kojima et al. *Large Language Models are Zero-Shot Reasoners.* 2023. arXiv: `2205.11916 [cs.CL]`.

[86]   Yongchao Zhou et al. *Large Language Models Are Human-Level Prompt Engineers.* 2023. arXiv: `2211.01910 [cs.LG]`.

[87]   Imanol Schlag et al. *Large Language Model Programs.* 2023. arXiv: `2305.05364 [cs.LG]`.

[88]   Harrison Chase. *LangChain.* 2023. URL: `https://perma.cc/U2V6-AL7V`.

[89]   Alexander Matt Turner et al. *Optimal Policies Tend to Seek Power.* 2023. arXiv: `1912.01683 [cs.AI]`.

[90]   Victoria Krakovna and Janos Kramar. *Power-seeking can be probable and predictive for trained agents.* 2023. arXiv: `2304.06528 [cs.AI]`.

[91]   Evan Hubinger et al. *Risks from Learned Optimization in Advanced Machine Learning Systems.* 2021. arXiv: `1906.01820 [cs.AI]`.

[92]   Dario Amodei et al. *Concrete Problems in AI Safety.* 2016. arXiv: `1606.06565 [cs.AI]`.

[93]   Yotam Wolf et al. *Fundamental Limitations of Alignment in Large Language Models.* 2023. arXiv: `2304.11082 [cs.CL]`.

[94] Simon Willison. *Prompt injection: What's the worst that can happen?* Apr. 14, 2023. URL: `https://perma.cc/D7B6-ESAX`.

[95] Giuseppe Venuto. *LLM failure archive (ChatGPT and beyond)*. 2023. URL: `https://perma.cc/UJ8A-YAE2`.

[96] Alex Albert. *Jailbreak Chat*. 2023. URL: `https://perma.cc/DES4-87DP`.

[97] Rachel Metz. *Jailbreaking AI Chatbots Is Tech's New Pastime*. Apr. 8, 2023. URL: `https://perma.cc/ZLU6-PBUN`.

[98] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. 2022. arXiv: 2212.08073 [cs.CL].

[99] Alexander Pan et al. *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark*. 2023. arXiv: 2304.03279 [cs.LG].

[100] Markus Anderljung and Julian Hazell. *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* 2023. arXiv: 2303.09377 [cs.AI].

[101] Lennart Heim. *Estimating PaLM's training cost*. Apr. 5, 2023. URL: `https://perma.cc/S4NF-GQ96`.

[102] Jaime Sevilla et al. "Compute Trends Across Three Eras of Machine Learning". In: *2022 International Joint Conference on Neural Networks*. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9891914.

[103] Ben Cottier. *Trends in the dollar training cost of machine learning systems*. OpenAI. Jan. 31, 2023. URL: `https://perma.cc/B9CB-T6C5`.

[104] Atila Orhon, Michael Siracusa, and Aseem Wadhwa. *Stable Diffusion with Core ML on Apple Silicon*. 2022. URL: `https://perma.cc/G5LA-94LM`.

[105] Simon Willison. *Running LLaMA 7B and 13B on a 64GB M2 MacBook Pro with llama.cpp*. 2023. URL: `https://perma.cc/E8KY-CT6Z`.

[106] Nomic AI. *GPT4All*. URL: `https://perma.cc/EMR7-ZY6M`.

[107] Yu-Hui Chen et al. *Speed Is All You Need: On-Device Acceleration of Large Diffusion Models via GPU-Aware Optimizations*. 2023. arXiv: 2304.11267 [cs.CV].

[108] Irene Solaiman et al. *Release Strategies and the Social Impacts of Language Models*. 2019. arXiv: 1908.09203 [cs.CL].

[109] Irene Solaiman. *The Gradient of Generative AI Release: Methods and Considerations*. 2023. arXiv: 2302.04844 [cs.CY].

[110] Toby Shevlane. *Structured access: an emerging paradigm for safe AI deployment*. 2022. arXiv: 2201.05159 [cs.AI].

[111] "How to be responsible in AI publication". In: *Nature Machine Intelligence* 3.5 (May 2021), pp. 367–367. DOI: 10.1038/s42256-021-00355-6.

[112] Aviv Ovadya and Jess Whittlestone. *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*. 2019. arXiv: 1907.11274 [cs.CY].

[113] Girish Sastry. *Beyond "Release" vs. "Not Release"*. 2021. URL: `https://perma.cc/JEZ2-ZB3W`.

[114] Connor Leahy. *Why Release a Large Language Model?* EleutherAI. June 2, 2021. URL: `https://perma.cc/Z9XE-GLRF`.

[115] BigScience. *Introducing The World's Largest Open Multilingual Language Model: BLOOM*. 2023. URL: `https://perma.cc/N9ZA-LXWW`.

[116] Hugging Face. *We Raised $100 Million for Open & Collaborative Machine Learning*. May 9, 2022. URL: `https://perma.cc/DEU6-9EF9`.

[117] laion.ai. *Open Assistant*. 2023. URL: `https://perma.cc/YB8U-NZQE`.

[118] Rohan Taori et al. *Alpaca: A Strong, Replicable Instruction-Following Model*. Center for Research on Foundation Models. 2023. URL: `https://perma.cc/Q75B-5KRX`.

[119] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2023. arXiv: `2303.18223 [cs.CL]`.

[120] Ryan C. Maness. *The Dyadic Cyber Incident and Campaign Data*. 2022. URL: `https://perma.cc/R2ZJ-PRGJ`.

[121] Carnegie Endowment for International Peace. *Timeline of Cyber Incidents Involving Financial Institutions*. 2022. URL: `https://perma.cc/TM34-ZHUH`.

[122] Center for Strategic and International Studies. *Significant Cyber Incidents*. May 2023. URL: `https://perma.cc/H3J2-KZFW`.

[123] Michael S. Schmidt and David E. Sanger. *Russian Hackers Read Obama's Unclassified Emails, Officials Say*. Apr. 25, 2015. URL: `https://perma.cc/JU2G-25MM`.

[124] Ben Buchanan. *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*. Oxford University Press, 2017.

[125] *China's Access to Foreign AI Technology*. Sept. 2019. URL: `https://perma.cc/ZV3F-G7KK`.

[126] National Counterintelligence and Security Center. *Protecting Critical and Emerging U.S. Technologies from Foreign Threats*. Oct. 2021. URL: `https://perma.cc/4P9M-QLM9`.

[127] NVIDIA Research Projects. *StyleGAN – Official TensorFlow Implementation*. 2019. URL: `https://perma.cc/TMD4-PYBY`.

[128] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: `1812.04948 [cs.NE]`.

[129] Rachel Metz. *These people do not exist. Why websites are churning out fake images of people (and cats)*. Feb. 28, 2019. URL: `https://perma.cc/83Q5-4KJW`.

[130] Phillip Wang. *This Person Does Not Exist*. 2019. URL: `https://perma.cc/XFH9-NRQV`.

[131] Fergal Gallagher and Erin Calabrese. *Facebook's latest takedown has a twist – AI-generated profile pictures*. Dec. 31, 2019. URL: `https://perma.cc/5Q2V-4BD2`.

[132] Shannon Bond. *AI-generated fake faces have become a hallmark of online influence operations*. National Public Radio. Dec. 15, 2022. URL: `https://perma.cc/DC5D-TJ32`.

[133] Google DeepMind. *AlphaFold: a solution to a 50-year-old grand challenge in biology*. Nov. 30, 2020. URL: `https://perma.cc/C6J4-6XWD`.

[134] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (July 2021), pp. 583–589. DOI: `10.1038/s41586-021-03819-2`.

[135] Gustaf Ahdritz et al. "OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization". In: *bioRxiv* (2022). DOI: `10.1101/2022.11.20.517210`. URL: `https://www.biorxiv.org/content/early/2022/11/22/2022.11.20.517210`.

[136] Jack W. Rae et al. *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. 2022. arXiv: `2112.11446 [cs.CL]`.

[137] Meta AI. *Introducing LLaMA: A foundational, 65-billion-parameter large language model*. Feb. 24, 2023. URL: `https://perma.cc/59YP-6ZDE`.

[138] *Runaway LLaMA: How Meta's LLaMA NLP model leaked*. Mar. 15, 2023. URL: `https://perma.cc/44YT-UNZ6`.

[139] Arnav Gudibande et al. *The False Promise of Imitating Proprietary LLMs*. 2023. arXiv: `2305.15717 [cs.CL]`.

[140] Katyanna Quach. *Stanford sends 'hallucinating' Alpaca AI model out to pasture over safety, cost*. Mar. 21, 2023. URL: `https://perma.cc/52NR-CMRF`.

[141] Tatsu. *Stanford Alpaca: An Instruction-following LLaMA Model*. 2023. URL: `https://perma.cc/SW29-C83N`.

[142] Emily H. Soice et al. *Can large language models democratize access to dual-use biotechnology?* 2023. arXiv: `2306.03809 [cs.CY]`.

[143] Google. *Responsible AI practices*. 2023. URL: `https://perma.cc/LKN6-P76L`.

[144] Cohere, OpenAI, and AI21 Labs. *Joint Recommendation for Language Model Deployment*. June 2, 2022. URL: `https://perma.cc/ZZ5Y-FNFY`.

[145] Microsoft. *Microsoft Responsible AI Standard*. June 2022. URL: `https://perma.cc/4XWP-NWK7`.

[146] Amazon AWS. *Responsible Use of Machine Learning*. 2023. URL: `https://perma.cc/U7GB-X4WV`.

[147] PAI Staff. *PAI Is Collaboratively Developing Shared Protocols for Large-Scale AI Model Safety*. Apr. 6, 2023. URL: `https://perma.cc/ZVQ4-3WJK`.

[148] Jonas Schuett et al. *Towards Best Practices in AGI Safety and Governance*. Centre for the Governance of AI. May 17, 2023. URL: `https://perma.cc/AJC3-M3AM`.

[149] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework*. Jan. 2023. URL: `https://perma.cc/N5SA-N6LT`.

[150] The IA Act. *Standard Setting*. 2023. URL: `https://perma.cc/T9RA-5Q37`.

[151] Franklin D. Raines. *Circular No. A-119 Revised*. Feb. 10, 1998. URL: `https://perma.cc/F2NH-NYHH`.

[152] National Telecommunications and Information Administration. *AI Accountability Policy Request for Comment*. 2023. URL: `https://perma.cc/E4C9-QQ8V`.

[153] Department for Science, Innovation and Technology. *New UK initiative to shape global standards for Artificial Intelligence*. Jan. 2022. URL: `https://www.gov.uk/government/news/new-uk-initiative-to-shape-global-standards-for-artificial-intelligence`.

[154] European Commission. *Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence*. Dec. 5, 2022. URL: `https://perma.cc/8GBP-NJAW`.

[155] Gillian K. Hadfield and Jack Clark. *Regulatory Markets: The Future of AI Governance*. 2023. arXiv: `2304.04914 [cs.AI]`.

[156] Ministry of Defence. *Foreword by the Secretary of State for Defence*. June 15, 2022.

[157] United States Government Accountability Office. *Status of Developing and Acquiring Capabilities for Weapon Systems*. Feb. 2022. URL: `https://perma.cc/GJN4-HQM8`.

[158] The White House. *FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety*. May 4, 2023. URL: `https://perma.cc/J6RR-2FVE`.

[159] Government of the United Kingdom. *The roadmap to an effective AI assurance ecosystem*. Dec. 8, 2021. URL: `https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version`.

[160] Department for Science, Innovation and Technology. *Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI*. Apr. 24, 2023. URL: `https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai`.

[161] National Artificial Intelligence Research Resource Task Force. *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem*. Jan. 2023. URL: `https://perma.cc/N99K-ARLP`.

[162]   Michael Atleson. *Keep your AI claims in check*. Federal Trade Commission. Feb. 27, 2023. URL: `https://perma.cc/M59A-Z4AV`.

[163]   Information Commissioner's Office. *Artificial intelligence*. 2023. URL: `https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/`.

[164]   The White House. *Blueprint for an AI Bill of Rights*. 2022. URL: `https://perma.cc/HXS9-66Q5`.

[165]   Computer Security Resource Center. *Proposed Update to the Framework for Improving Critical Infrastructure Cybersecurity*. Jan. 25, 2017. URL: `https://perma.cc/CD97-YW27`.

[166]   National Institute of Standards and Technology. *Request for Comments on the Preliminary Draft of the NIST Privacy Framework*. 2020. URL: `https://perma.cc/5U9R-4UFQ`.

[167]   National Telecommunications and Information Administration. *NTIA Seeks Public Input to Boost AI Accountability*. Apr. 11, 2023. URL: `https://perma.cc/XJH6-YNXB`.

[168]   Matthew C. Stephenson. "Information Acquisition and Institutional Design". In: *Harvard Law Review* 124.4 (2011).

[169]   Cary Coglianese, Richard Zeckhauser, and Edward A. Parson. "Seeking Truth for Power: Informational Strategy and Regulatory PolicymakingPolicymaking". In: *Michigan Law review* 89.2 (2004), pp. 277–341.

[170]   Thomas O. McGarity. "Regulatory Reform in the Reagan Era". In: *Maryland Law Review* 45.2 (1986).

[171]   Rovy Van Loo. "Regulatory Monitors: Policing Firms in the Compliance Era". In: *Columbia Law Review* 119 (2019).

[172]   Rovy Van Loo. "The Missing Regulatory State: Monitoring Businesses in an Age of SurveillanceSurveillance". In: *Vanderbilt Law Review* 72.5 (2019).

[173]   Noam Kolt. "Algorithmic Black Swans". In: *Washington University Law Review* 101 (2023).

[174]   Gary E. Marchant, Braden R. Allenby, and Joseph R. Herkert. *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight*. Springer, 2011. URL: `https://perma.cc/4XXW-3RHH`.

[175]   Margaret Mitchell et al. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2019. DOI: `10.1145/3287560.3287596`.

[176]   Timnit Gebru et al. "Datasheets for datasets". In: 64.12 (2021), pp. 86–92.

[177]   Thomas Krendl Gilbert et al. *Reward Reports for Reinforcement Learning*. 2023. arXiv: 2204.10817 `[cs.LG]`.

[178]   Standford University. *ecosystem graphs*. 2023. URL: `https://perma.cc/H6GW-Q78M`.

[179]   Jaime Sevilla, Anson Ho, and Tamay Besiroglu. "Please Report Your Compute". In: *Communications of the ACM* 66.5 (Apr. 2023), pp. 30–32. DOI: `10.1145/3563035`.

[180]   Inioluwa Deborah Raji et al. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*. 2020. arXiv: 2001.00973 `[cs.CY]`.

[181]   Ann M. Lipton. "Not Everything Is About Investors: The Case for Mandatory Stakeholder Disclosure". In: *Yale Journal on Regulation* (). URL: `https://perma.cc/G97G-3FL2`.

[182]   Jess Whittlestone and Jack Clark. *Why and How Governments Should Monitor AI Development*. 2021. arXiv: 2108.12427 `[cs.CY]`.

[183]   Jakob Mökander et al. *Auditing large language models: a three-layered approach*. 2023. arXiv: 2302.08500 `[cs.CL]`.

[184]   Inioluwa Deborah Raji et al. *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*. 2022. arXiv: 2206.04737 `[cs.CY]`.

[185]   Hannah Bloch-Wehba. "The Promise and Perils of Tech Whistleblowing". In: *Northwestern University Law Review* (Mar. 3, 2023).

[186] Sonia K. Katyal. *Private Accountability in the Age of Artificial Intelligence*. Dec. 14, 2018. URL: `https://perma.cc/PNW4-7LN2`.

[187] Helen Toner, Patrick Hall, and Sean McGregor. *AI Incident Database*. 2023. URL: `https://perma.cc/JJ95-7K7B`.

[188] Epoch AI. *ML Inputs*. 2023. URL: `https://perma.cc/9XBU-6NES`.

[189] Center for Emerging Technology. *Emerging Technology Observatory*. 2022. URL: `https://perma.cc/L4DB-YQ5L`.

[190] European Commission. *Joint Statement EU-US Trade and Technology Council of 31 May 2023 in Lulea, Sweden*. May 21, 2023. URL: `https://perma.cc/8PDH-8S34`.

[191] Department for Science, Innovation and Technology. *AI regulation: a pro-innovation approach*. Mar. 29, 2023. URL: `https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach`.

[192] Jonas Schuett. *Three lines of defense against risks from AI*. 2022. arXiv: 2212.08364 [`cs.CY`].

[193] Peter Cihon et al. "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries". In: *IEEE Transactions on Technology and Society* 2.4 (Dec. 2021), pp. 200–209. DOI: `10.1109/tts.2021.3077595`.

[194] International Organization for Standardization. *Consumers and Standards: Partnership for a Better World*. URL: `https://perma.cc/5XJP-NC5S`.

[195] Administrative Conference of the United States. *Incorporation by Reference*. Dec. 8, 2011. URL: `https://perma.cc/Q3H9-DBK9`.

[196] Business Operations Support System. *The 'New Approach'*. URL: `https://perma.cc/ZS9G-LV66`.

[197] World Trade Organization. *Agreement on Technical Barriers to Trade*. URL: `https://perma.cc/PE55-5GJV`.

[198] U.S. Securities and Exchange Commission. *Addendum to Division of Enforcement Press Release*. 2023. URL: `https://perma.cc/M3LN-DGGV`.

[199] Philip F.S. Berg. "Unfit to Serve: Permanently Barring People from Serving as Officers and Directors of Publicly Traded Companies After theOfficers and Directors of Publicly Traded Companies After the Sarbanes-Oxley ActSarbanes-Oxley Act". In: *Vanderbilt Law ReviewVanderbilt Law Review* 56.6 ().

[200] Office of the Comptroller of the Currency. *Bank Supervision Process, Comptroller's Handbook*. Sept. 30, 2019. URL: `https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/bank-supervision-process/pub-ch-bank-supervision-process.pdf`.

[201] David A. Hindin. *Issuance of the Clean Air Act Stationary Source Compliance Monitoring Strategy*. Oct. 4, 2016. URL: `https://perma.cc/6R7C-PKB2`.

[202] Commitee on Armed Services. *Hearing To Receive Testimony on the State of Artificial Intelligence and Machine Learning Applications To Improve Department of Defense Operations*. Apr. 19, 2023. URL: `https://perma.cc/LV3Z-J7BT`.

[203] Microsoft. *Governing AI: A Blueprint for the Future*. May 2023. URL: `https://perma.cc/3NL2-P4XE`.

[204] Subcommittee on Privacy, Technology and the Law. *Oversight of A.I.: Rules for Artificial Intelligence*. 2023. URL: `https://perma.cc/4WCU-FWUL`.

[205] Patrick Murray. "Noational: Artificial Intelligence Use Prompts Concerns". In: (2023). URL: `https://perma.cc/RZT2-BWCM`.

[206] Jamie Elsey and David Moss. *US public opinion of AI policy and risk*. Rethink Priorities. May 12, 2023. URL: https://perma.cc/AF29-JT8K.

[207] Federal Aviation Administration. *Classes of Airports – Part 139 Airport Certification*. May 2, 2023. URL: https://perma.cc/9JLB-6D7R.

[208] Federal Aviation Administration. *Air Carrier and Air Agency Certification*. June 22, 2022. URL: https://perma.cc/76CZ-WLB6.

[209] California Energy Commission. *Power Plant Licensing*. URL: https://perma.cc/BC7A-9AM3.

[210] U.S. Food and Drug Administration. *Electronic Drug Registration and Listing System (eDRLS)*. Apr. 11, 2021. URL: https://perma.cc/J357-89YH.

[211] Congressional Research Service. *An Analysis of Bank Charters and Selected Policy Issues*. Jan. 21, 2022. URL: https://perma.cc/N9HU-JTJJ.

[212] U.S. Food and Drug Administration. *Development and Approval Process*. Aug. 8, 2022. URL: https://perma.cc/47UY-NVHV.

[213] Centers for Disease Control and Prevention/Division of Select Agents and Toxins & Animal and Plant Health Inspection Service/Division of Agricultural Select Agents and Toxins. *Federal Select Agent Program*. 2022. URL: https://perma.cc/3TZP-GAV6.

[214] Centers for Disease Control and Prevention/Division of Select Agents and Toxins & Animal and Plant Health Inspection Service/Division of Agricultural Select Agents and Toxins. *Select Agents and Toxins List*. 2023. URL: https://perma.cc/W8K8-LQV4.

[215] Centers for Disease Control and Prevention/Division of Select Agents and Toxins & Animal and Plant Health Inspection Service/Division of Agricultural Select Agents and Toxins. *2021 Annual Report of the Federal Select Agent Program*. 2021. URL: https://perma.cc/RPV8-47GW.

[216] Centers for Disease Control and Prevention/Division of Select Agents and Toxins & Animal and Plant Health Inspection Service/Division of Agricultural Select Agents and Toxins. *Select Agents Regulations*. 2022. URL: https://perma.cc/MY34-HX79.

[217] Centers for Disease Control and Prevention/Division of Select Agents and Toxins & Animal and Plant Health Inspection Service/Division of Agricultural Select Agents and Toxins. *Security Risk Assessments*. 2022. URL: https://perma.cc/ZY4A-5BB2.

[218] Centers for Disease Control and Prevention/Division of Select Agents and Toxins & Animal and Plant Health Inspection Service/Division of Agricultural Select Agents and Toxins. *Preparing for Inspection*. 2021. URL: https://perma.cc/Z73F-3RVV.

[219] George J. Stigler. "The Theory of Economic Regulation". In: *The Bell Journal of Economics and Management Science* 2.1 (1971), pp. 3–21.

[220] Gary S. Becker. "A Theory of Competition among Pressure Groups for Political Influence". In: *The Quarterly Journal of Economics* 98 (1983), pp. 371–395.

[221] Daniel Carpenter and David Moss, eds. *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press, 2013.

[222] *Recruiting Tech Talent to Congress*. 2023. URL: https://perma.cc/SLY8-5M39.

[223] Open Philanthropy. *Open Philanthropy Technology Policy Fellowship*. URL: https://perma.cc/BY47-SS5V.

[224] Mhairi Aitken et al. *Common Regulatory Capacity for AI*. The Alan Turing Institute. 2022. DOI: 10.5281/zenodo.6838946.

[225] Meta AI. *System Cards, a new resource for understanding how AI systems work*. Feb. 2022. URL: https://perma.cc/46UG-GA9D.

[226] Leon Derczynski et al. *Assessing Language Model Deployment with Risk Cards*. 2023. arXiv: `2303.18190 [cs.CL]`.

[227] Certification Working Group. *Unlocking the Power of AI*. June 8, 2023. URL: `https://perma.cc/DLF3-E38T`.

[228] Percy Liang et al. *Holistic Evaluation of Language Models*. 2022. arXiv: `2211.09110 [cs.CL]`.

[229] Stella Biderman et al. *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. 2023. arXiv: `2304.01373 [cs.CL]`.

[230] Aarohi Srivastava et al. *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. 2022. arXiv: `2206.04615 [cs.CL]`.

[231] Dan Hendrycks et al. *Measuring Massive Multitask Language Understanding*. 2021. arXiv: `2009.03300 [cs.CY]`.

[232] Heidy Khlaaf. *Toward Comprehensive Risk Assessments*. Trail of Bits. Mar. 2023. URL: `https://perma.cc/AQ35-6JTV`.

[233] Deep Ganguli et al. *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. 2022. arXiv: `2209.07858 [cs.CL]`.

[234] Ethan Perez et al. *Red Teaming Language Models with Language Models*. 2022. arXiv: `2202.03286 [cs.CL]`.

[235] Miles Brundage et al. *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. 2020. arXiv: `2004.07213 [cs.CY]`.

[236] ARC Evals. *Update on ARC's recent eval efforts*. Mar. 17, 2023. URL: `https://perma.cc/8VWF-QYPH`.

[237] Ian McKenzie et al. *Inverse Scaling Prize: First Round Winners*. Fund for Alignment Research (FAR). 2022. URL: `https://irmckenzie.co.uk/round1`.

[238] Ian McKenzie et al. *Inverse Scaling Prize: Second Round Winners*. Fund for Alignment Research (FAR). 2022. URL: `https://irmckenzie.co.uk/round2`.

[239] Leo Gao, John Schulman, and Jacob Hilton. *Scaling Laws for Reward Model Overoptimization*. 2022. arXiv: `2210.10760 [cs.LG]`.

[240] Samuel R. Bowman et al. *Measuring Progress on Scalable Oversight for Large Language Models*. 2022. arXiv: `2211.03540 [cs.HC]`.

[241] Samir Passi and Mihaela Vorvoreanu. *Overreliance on AI: Literature Review*. AI Ethics, Effects in Engineering, and Research. June 2022.

[242] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (Mar. 2023), pp. 1–38. DOI: `10.1145/3571730`.

[243] Samuel Gehman et al. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020. DOI: `10.18653/v1/2020.findings-emnlp.301`.

[244] Amanda Askell et al. *A General Language Assistant as a Laboratory for Alignment*. 2021. arXiv: `2112.00861 [cs.CL]`.

[245] Paul Christiano. *Mechanistic Anomaly Detection and ELK*. Nov. 2022. URL: `https://perma.cc/WH44-WVRV`.

[246] Catherine Olsson et al. *In-context Learning and Induction Heads*. Mar. 2022. URL: `https://perma.cc/FQP6-2Z4G`.

[247] Tom Henighan et al. *Superposition, Memorization, and Double Descent*. Jan. 2023. URL: `https://perma.cc/5ZTF-RMV8`.

[248]   Ian Tenney et al. *The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models*. 2020. arXiv: 2008.05122 [cs.CL].

[249]   Shoaib Ahmed Siddiqui et al. *Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics*. 2022. arXiv: 2209.10015 [cs.LG].

[250]   Toby Shevlane and Allan Dafoe. *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?* 2020. arXiv: 2001.00463 [cs.CY].

[251]   OECD. *OECD Framework for the Classification of AI systems*. Feb. 2022. DOI: 10.1787/cb6d9eca-en.

[252]   Irene Solaiman et al. *Evaluating the Social Impact of Generative AI Systems in Systems and Society*. 2023. arXiv: 2306.05949 [cs.CY].

[253]   ITU News. *How AI can help fight misinformation*. May 2, 2022. URL: https://perma.cc/R7RA-ZX5G.

[254]   Ajeya Cotra. *Training AIs to help us align AIs*. Mar. 26, 2023. URL: https://perma.cc/3L49-7QU7.

[255]   Geoffrey Irving, Paul Christiano, and Dario Amodei. *AI safety via debate*. 2018. arXiv: 1805.00899 [stat.ML].

[256]   Elisabeth Keller and Gregory A. Gehlmann. "Introductory comment: a historical introduction to the Securities Act of 1933 and the Securities Exchange Act of 1934". In: *Ohio State Law Journal* 49 (1988), pp. 329–352.

[257]   Inioluwa Deborah Raji and Joy Buolamwini. "Actionable Auditing". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Jan. 2019. DOI: 10.1145/3306618.3314244.

[258]   Jakob Mökander et al. "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations". In: *Science and Engineering Ethics* 27.4 (July 2021). DOI: 10.1007/s11948-021-00319-4.

[259]   Gregory Falco et al. "Governing AI safety through independent audits". In: *Nature Machine Intelligence* 3.7 (July 2021), pp. 566–571. DOI: 10.1038/s42256-021-00370-7.

[260]   Inioluwa Deborah Raji et al. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, July 2022. DOI: 10.1145/3514094.3534181.

[261]   Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, June 2022. DOI: 10.1145/3531146.3533213.

[262]   OpenAI. *DALL·E 2 Preview - Risks and Limitations*. July 19, 2022. URL: https://perma.cc/W9GA-8BYQ.

[263]   Daniel M. Ziegler et al. *Fine-Tuning Language Models from Human Preferences*. 2020. arXiv: 1909.08593 [cs.CL].

[264]   Jesse Dodge et al. *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping*. 2020. arXiv: 2002.06305 [cs.CL].

[265]   Pengfei Liu et al. *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. 2021. arXiv: 2107.13586 [cs.CL].

[266]   Xiang Lisa Li and Percy Liang. *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. 2021. arXiv: 2101.00190 [cs.CL].

[267]   Eric Wallace et al. *Universal Adversarial Triggers for Attacking and Analyzing NLP*. 2021. arXiv: 1908.07125 [cs.CL].

[268] Jonas Schuett. *AGI labs need an internal audit function*. 2023. arXiv: 2305.17038 [cs.CY].

[269] Richard Worthington. "The Social Control of Technology". In: *American Political Science Review* 76.1 (Mar. 1982), pp. 134–135. DOI: 10.2307/1960465.

[270] Competition and Markets Authority. *CMA launches initial review of artificial intelligence models*. May 4, 2023. URL: https://www.gov.uk/government/news/cma-launches-initial-review-of-artificial-intelligence-models.

[271] Jonas Schuett. "Defining the scope of AI regulations". In: *Law, Innovation and Technology* 15.1 (Jan. 2023), pp. 60–82. DOI: 10.1080/17579961.2023.2184135.

[272] Robert Baldwin, Martin Cave, and Martin Lodge. *Understanding Regulation. Theory, Strategy, and Practice*. Oxford: Oxford University Press, 2011. 568 pp. ISBN: 9780199576098.

[273] Cabinet Office. *National Risk Register 2020*. 2020. URL: https://www.gov.uk/government/publications/national-risk-register-2020.

[274] Common Crawl. *Common Crawl*. We build and maintain an open repository of web crawl data that can be accessed and analyzed by anyone. 2023. URL: https://perma.cc/9EC5-QPJ7.

[275] Louis Kaplow. "Rules versus Standards: An Economic Analysis". In: *Duke Law Journal* 42.3 (Dec. 1992), pp. 557–629.

[276] Danny Hernandez and Tom B. Brown. *Measuring the Algorithmic Efficiency of Neural Networks*. 2020. arXiv: 2005.04305 [cs.LG].

[277] EpochAI. *Cost estimates for GPT-4*. 2023. URL: https://perma.cc/3UJX-783P.

[278] Andrew Lohn and Micah Musser. *AI and Compute. How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* Center for Security and Emerging Technology, Jan. 2022.

[279] Daniel Bashir and Andrey Kurenkov. *The AI Scaling Hypothesis*. Last Week in AI. Aug. 5, 2022. URL: https://perma.cc/4R26-VCQZ.

[280] Jaime Sevilla et al. *Compute Trends Across Three Eras of Machine Learning*. 2022. arXiv: 2202.05924 [cs.LG].

[281] Gwern. *The Scaling Hypothesis*. 2023. URL: https://perma.cc/7CT2-NNYL.

[282] Rich Sutton. *The Bitter Lesson*. Mar. 13, 2019. URL: https://perma.cc/N9TY-DH22.

[283] Lennart Heim. *This can't go on(?) - AI Training Compute Costs*. June 1, 2023. URL: https://perma.cc/NCE6-NT3W.

[284] OpenAI. *AI and efficiency*. May 5, 2020. URL: https://perma.cc/Y2CW-JAR9.

[285] Ben Sorscher et al. *Beyond neural scaling laws: beating power law scaling via data pruning*. 2023. arXiv: 2206.14486 [cs.LG].

[286] Deep Ganguli et al. "Predictability and Surprise in Large Generative Models". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, June 2022. DOI: 10.1145/3531146.3533229. URL: https://doi.org/10.1145%5C%2F3531146.3533229.