Increased Compute Efficiency and the Diffusion of AI Capabilities

Konstantin F. Pilz[*1†], **Lennart Heim**[*2], **Nicholas Brown**[3]

[1]Georgetown University, Washington, D.C., United States
[2]Centre for the Governance of AI, Oxford, United Kingdom
[3]Independent Researcher

Abstract

Training advanced AI models requires large investments in computational resources, or *compute*. Yet, as hardware innovation reduces the price of compute and algorithmic advances make its use more efficient, the cost of training an AI model to a given performance falls over time — a concept we describe as *increasing compute efficiency*. We find that while an *access effect* increases the number of actors who can train models to a given performance over time, a *performance effect* simultaneously increases the performance available to each actor. This potentially enables large compute investors to pioneer new capabilities, maintaining a performance advantage even as capabilities diffuse. Since large compute investors tend to develop new capabilities first, it will be particularly important that they share information about their AI models, evaluate them for emerging risks, and, more generally, make responsible development and release decisions. Further, as compute efficiency increases, governments will need to prepare for a world where dangerous AI capabilities are widely available — for instance, by developing defenses against harmful AI models or by actively intervening in the diffusion of particularly dangerous capabilities.

Extended version — https://arxiv.org/abs/2311.15377

1 Introduction

Over the past decade, computational resources (compute) emerged as a major driver of rapid advances in the field of artificial intelligence (AI). The amount of computational operations used in training the largest AI models has doubled approximately every six months since 2010 (Sevilla et al. 2022), enabling powerful new applications like high-quality image generators, coding assistants, and conversational chatbots (Yu et al. 2022; OpenAI 2023). Researchers have discovered empirical scaling laws that formalize the relationship between increased training compute and improved model performance across numerous domains.[1] These scaling laws and the current pace of training compute growth suggest that compute will likely continue to contribute to increased model performance for the foreseeable future (Sevilla et al. 2022; Villalobos 2023).[2]

1.1 Falling Training Cost

As compute budgets have surged, the cost of training an AI model to a given performance has fallen continuously. In 2017, training an image classifier to 93% classification accuracy on the ImageNet dataset cost over $1,000; by 2021, it cost only $5 — a reduction of over 99% (Zhang et al. 2022, p. 97). In 2020, OpenAI's GPT-3 cost at least $4.6 million in cloud compute to train (Li 2020); two years later, the company Mosaic claimed to achieve the same performance for a tenth of the price (Venigalla and Li 2022).

The cost of training falls due to two primary factors: (a) improvements in the price performance of AI hardware and (b) improvements in the efficiency of AI algorithms.

1.2 Hardware Price Performance

Hardware price performance determines the quantity of computational resources available for a given monetary investment. Innovations in hardware manufacturing and design increase price performance over time. For example, consumer prices for Intel's processors fell by 42% annually from 2004 to 2013, while their performance improved by 29% each year; both trends increased the computational performance available per dollar (Byrne, Oliner, and Sichel 2017). In the context of AI accelerators, Hobbhahn, Heim, and Aydos (2023) found that GPUs used in machine learning doubled in price performance approximately every two years. In addition, increased specialization, such as hardware-supported lower-precision number formats, further increased price performance by up to an order of magnitude (Hobbhahn, Heim, and Aydos 2023).

[*]These authors contributed equally.
[†]Corr.: lennart.heim@governance.ai, kfp15@georgetown.edu

[1]Scaling laws occur in a variety of domains, such as natural language processing (Hoffmann et al. 2022), protein structure prediction, acoustic modeling, and recommendation models (Chen et al. 2023; Droppo and Elibol 2021; Ardalani et al. 2022).

[2]Our model assumes that compute is a main determinant of an AI model's capabilities. However, training AI models with large amounts of compute is a sophisticated engineering challenge that requires talented researchers. Thus, simply having access to large amounts of compute does not guarantee a powerful AI model. We discuss additional factors and limitations in Section 4.

1.3 Algorithmic Efficiency

Falling AI training costs stem not just from more compute being available per dollar but also from ongoing improvements in *algorithmic efficiency*. Algorithmic efficiency determines the amount of computational resources required to execute a given algorithm. In the context of AI, algorithmic efficiency identifies the compute budget required to train or run inference on a model providing a given level of performance. In this paper, we will focus on the former, the efficiency of training.

Due to innovations across the AI stack, algorithmic efficiency has improved dramatically over the past decade. For example, Hernandez and Brown (2020) found that over the period of 2012 to 2019, improvements in image classification algorithms had led to a 97.7% reduction in the compute required to train a classifier to the performance of AlexNet. Erdil and Besiroglu (2023) extended the analysis to 2022 and found an even faster rate of algorithmic improvement in image classification, in which compute requirements halved every nine months. [3] Although limited analysis exists, algorithmic improvements have also impacted other fields. For instance, in language modeling, less compute-intensive models increasingly replicate the performance of previous, larger models (Mistral 2023; Hoffmann et al. 2022; Venigalla and Li 2022).

1.4 Compute Investment Efficiency

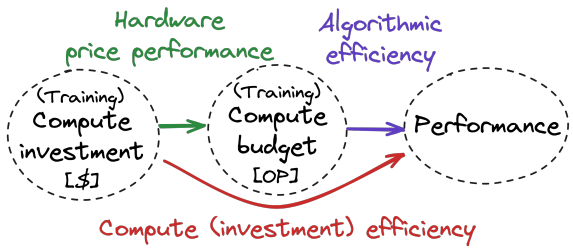

Figure 1: Hardware price performance is the conversion function between the training compute investment in dollars and the training compute budget in operations. Algorithmic efficiency is the subsequent conversion function between the training compute budget and the performance of the resulting AI model. *Compute (investment) efficiency* combines hardware price performance and algorithmic efficiency, relating training compute investment to the performance of the resulting model.

We introduce the concept of *compute investment efficiency* (abbreviated to *compute efficiency*), to refer to the relationship between the monetary spending on training compute and the performance of the resulting AI model (Figure 1).[4] Tracking compute efficiency over time captures advances in both the hardware and software used for AI training, including those advancements that depend on a mixture of hardware and software breakthroughs.[5]

In the following sections, we provide a conceptual model[6] of increasing compute efficiency, assess its impact on various actors, and evaluate its implications for competitive advantage and for the risks of dangerous capabilities. Finally, we discuss some of our assumptions and highlight relevant insights for policymakers and anyone else seeking to understand or respond to increased compute efficiency.

2 Effects of Compute Efficiency Increases

Similar to the effects of improved data efficiency in Tucker, Anderljung, and Dafoe (2020), we model the impact of increased compute efficiency through two effects: an *access effect* and a *performance effect*.

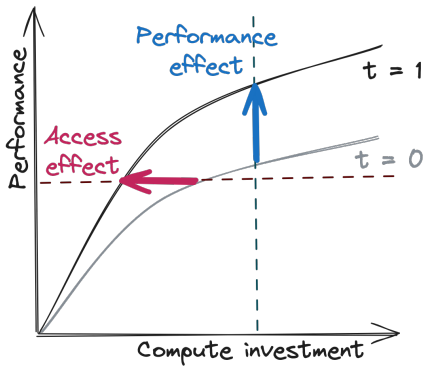

Figure 2: Compute efficiency improves between time $t = 0$ and $t = 1$, causing an access effect (red) and a performance effect (blue). *Figures are conceptual and do not make empirical claims about the slope of the curve.*

The Access Effect — Over time, training a model to a given level of performance[7] requires less compute investment. This *access effect* corresponds to a leftward shift on the performance vs. investment graph (Figure 2, red), allowing increasingly many actors to access models with a given level of performance.

The Performance Effect — Simultaneously, a fixed level of compute investment allows actors to train models to a higher performance over time. This *performance effect* corresponds to an upward shift in the performance vs. invest-

[3] Rather than smoothly doubling every nine months, algorithmic improvements usually occur in disjunct innovations. For instance, in image classifiers, those innovations included sparsity, batch normalization, and residual connections (Hernandez and Brown 2020).

[4] The model presented in this paper and many of its implications draw heavily from (Tucker, Anderljung, and Dafoe 2020) "Social and Governance Implications of Improved Data Efficiency". We translate the concepts and implications discussed in the context of increased data efficiency to the realm of compute efficiency.

[5] For instance, Hobbhahn, Heim, and Aydos (2023) identify the recent shift from FP32 precision to tensor-FP16/8 as one of the primary drivers of AI hardware improvements.

[6] For a formalization of our model and its assumptions, see extended version at https://arxiv.org/abs/2311.15377.

[7] We use *performance* to refer to hard metrics such as test loss or the accuracy on a test dataset. We later use *capabilities* to refer to qualitative metrics such as the types of problems an AI model can solve.

ment graph (Figure 2, blue). Compute efficiency improvements thus grant any actor access to increased performance.

3 Consequences

In this section, we illustrate the effects of our model through an example with three types of actors. We then assess what this implies for competitive advantage and the discovery and proliferation of dangerous capabilities.

3.1 Consequences of the Effects for Different Actors

To illustrate the impact of increased compute efficiency, we consider a scenario with three groups of actors:[8]

1. **Large compute investors**, who spend large amounts on training compute in order to advance the frontier of AI capabilities.
2. **Secondary actors,** who are able to make significant investments in training compute but who do *not* attempt to advance the frontier. Examples include large companies that integrate some AI-based tools into their products and smaller companies developing AI models on a limited budget.
3. **Compute-limited actors**, such as individuals, small collaborative projects, or academic researchers who can only afford limited compute investment.

We further assume that a model attains some relevant "Capability X" when it surpasses a given level of performance.[9] We now extrapolate repeated efficiency improvements and their effect on access to "Capability X".

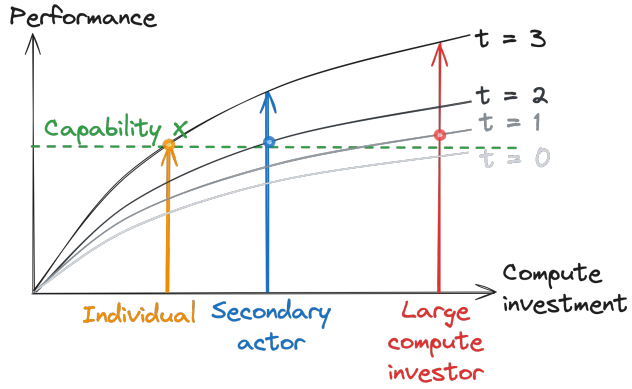

Figure 3: Diffusion of Capability X over time among three illustrative actors with varying levels of compute investment.

This extrapolation suggests the following insights:

(A) Large compute investors are the first to discover novel capabilities. As the performance effect increases the performance attainable to every actor, large compute investors are the first to discover novel capabilities (Figure 3, $t = 1$). For example, OpenAI's GPT-4 — likely the first training run to surpass 10^{25} FLOP (Epoch 2023) — showed previously unknown capabilities in mathematical reasoning, coding, and theory of mind (OpenAI 2023; Bubeck et al. 2023).

(B) Access to capabilities expands over time. As compute efficiency continues to improve, less and less investment is required to achieve a given level of performance. Over time, this lets more and more resource-limited actors recreate capabilities that were previously available only to large compute investors (Figure 3, $t = 2$). Eventually, even actors with very limited compute investments, such as research collectives or individuals, gain access to advanced capabilities (Figure 3, $t = 3$).

(C) Large compute investors can sustain their performance advantage as access expands. This is because the same efficiency increase that lets a smaller investor access the previous state-of-the-art (SOTA) often increases the current SOTA as well (Figure 3, $t = 3$). Assuming it is possible to derive additional value from increased performance, large investors will maintain an advantage. We explore this topic in more detail in the following section.

3.2 Implications for Competitive Advantage

Our model suggests several implications for the competitive advantage of different actors in the AI market.

Implications of Capability Diffusion With each efficiency increase, a broader range of actors can train AI models to a given performance, resulting in a wider variety of products. However, our model suggests that the impact on competition may often be limited because large compute investors benefit from the performance effect and retain an absolute advantage. Hence, smaller companies may face pressure to specialize, pursuing applications in areas with reduced competition from leaders.

For instance, various companies have recently developed smaller language models that perform significantly below OpenAI's GPT-4 on general benchmarks (Papers with Code 2023) but outperform it in specialized niches, such as finance (Wu et al. 2023), coding assistance (AWS 2023), or role-playing for entertainment (Character.AI 2023).

When AI capabilities finally diffuse even to low-resource actors like small research collectives and individuals, the range of use cases explored may increase significantly. This is because these groups have a broader range of interests than commercial entities do, so this diffusion could represent a qualitative shift in AI applications (Besiroglu et al. 2024; Ahmed and Wahed 2020).

Threshold Effects AI models sometimes demonstrate large improvements in capabilities even with only small increases in compute investment. These *performance thresholds* often occur in large language models, including for tasks like three-digit addition and general benchmarks like MMLU (Ganguli et al. 2022).

[8]For simplicity, we make the following assumptions about these groups: (1) Actors maintain a constant level of compute investment over time, (2) all actors have equal access to compute efficiency improvements, and (3) actors only have access to the capabilities of the models they train themselves. See sections 4 and 5 for a discussion of these assumptions.

[9]See Section 3.2 for a discussion of this assumption.

Our model suggests that when a compute efficiency advance first unlocks a qualitatively new capability, large compute investors are likely to be the only actors able to leverage it. Until subsequent advances broaden access (see Figure 3), these actors will enjoy reduced competition or even a full monopoly.

Performance Ceilings As developers scale their training compute investment, they may eventually encounter a region where further compute investment only marginally (if at all) increases a model's usefulness for real-world tasks. For example, the problem of identifying handwritten digits was effectively solved in the early 2000s, with classifiers reaching over 99% accuracy on the MNIST dataset (Liu et al. 2003). Similarly, face recognition technology has achieved impressively low error rates, down to 1:1000 on the VISA dataset (NIST 2023).

Performance ceilings can arise from either technical limitations or real-world constraints. Hestness et al. (2017) propose that, with sufficient training compute, an AI model may encounter an irreducible error region where its accuracy can no longer increase because the fixed training set and model architecture have extracted all possible information relevant to the task. Yet, the authors find this error region only in a simplified toy experiment and suggest that most AI applications are still far away from encountering it.

In practice, performance ceilings also occur when a model's performance continues to increase but no longer contributes to advances in real-world capabilities. For instance, in a voice-controlled application designed to recognize only a fixed set of commands, improving the underlying model's accuracy on less frequent words will not enhance its real-world functionality.

In our model, a performance ceiling creates an upper bound on achievable capability and thus dampens the performance effect, reducing the benefits leaders get from improved compute efficiency. Since the access effect remains unchanged, subsequent compute efficiency improvements let smaller actors approach the performance of the leaders who have already reached the ceiling. The performance ceiling, therefore, distributes capabilities evenly, and large compute investors slowly lose their performance lead.

However, under some circumstances, leaders may entrench their initial advantage and continue to dominate the field even as their performance advantage diminishes. Such *winner-take-all effects* occur, for instance, when large compute investors integrate AI models into their established ecosystems and thus create lock-in effects.[10] Alternatively, they may be able to use early profits to achieve economies of scale or apply exclusionary tactics to hamstring smaller competitors (Kuchinke and Vidal 2016).

Even without an apparent performance ceiling, AI models usually show reducing improvements from increasing compute investment or compute efficiency (Hoffmann et al. 2022). If returns diminish quickly, the field may encounter an effect similar to a performance ceiling, where the leaders' advantage reduces with increasing compute efficiency.

Zero-Sum Competition In certain domains, a relative performance advantage compared to competitors can lead to outsized benefits, even if the absolute performance difference is small. This is most notably the case in settings with zero-sum competition, [11] in which value is derived exclusively from performing better than an opponent. Zero-sum competition occurs in many domains, including algorithmic trading, in which a marginal speed advantage can let a trader exploit tiny market inefficiencies (Biais, Foucault, and Moinas 2015), law, in which the aim is to defeat an opponent's argument in court, and entertainment, in which competitors fight over a fixed pool of audience attention. Our model suggests that leaders may continue dominating in areas of zero-sum competition, even if their models only have a marginal performance advantage.

3.3 Implications for Dangerous Capabilities

Today's AI models already cause harm, for instance, by producing racist, sexist, or otherwise discriminating content or by aiding the suppression of minorities or political views (McGregor 2021; Creemers 2018). As models become more performant, they may develop capabilities that pose increasingly acute risks to safety and security, such as by automating cyberattacks (Amodei 2023), aiding in the design of bioweapons (Urbina et al. 2022; Nelson and Rose 2023), generating targeted disinformation (Sedova et al. 2021), or irreversibly evading human control (Carlsmith 2022; Ngo, Chan, and Mindermann 2023). Although some dangerous capabilities may first arise in highly specialized models that require only small amounts of compute, the majority likely first appear in compute-intensive general-purpose models (Anderljung et al. 2023).[12]

Discovery of Dangerous Capabilities As discussed in Section 3.3, we expect that large compute investors generally encounter novel capabilities first. This rule is likely to hold equally for benign and dangerous capabilities. However, without rigorous testing, developers may not immediately discover all dangerous capabilities of their models (Shevlane et al. 2023). Instead, some may become apparent only after being broadly available, including to potential malicious actors (Anderljung et al. 2023).

[10]For instance, Google currently integrates its chatbot Bard into its search engine and the Google workspace (Google 2023). Such integration may increase customers' switching costs, so they may continue using Google's AI model, even if competitors offer similar-performing models.

[11]Note that while leaders in these markets capture outsized benefits, they may not constitute zero-sum games in an economic context. The overall market can still expand due to rising model capabilities, effectively resulting in a positive-sum game where increased competition drives market growth.

[12]Not all dangerous capabilities require significant compute. For example, models predicting host-pathogen interactions could aid bioweapon development without large training costs. AlphaFold, an advanced protein predictor, required only 10^{20} FLOP, three orders of magnitude less than OpenAI's GPT-3 developed in the same year (Epoch 2022). Biological design tools similar to AlphaFold could play a crucial role in the development of bioweapons (Sandbrink 2023).

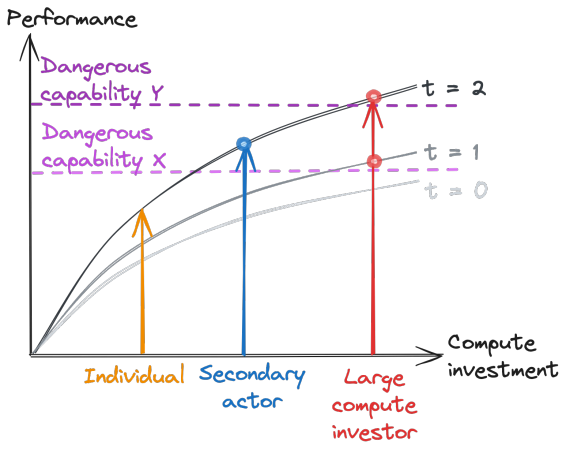

Figure 4: Example of three actors discovering novel dangerous capabilities X and Y at different points in time.

While only a small number of large compute investors will initially be able to train models that possess a novel dangerous capability X, compute efficiency improvements soon allow more and more actors to recreate such models (Figure 4). This complicates coordination and oversight, making it potentially difficult to entirely avoid harmful deployment.

Using Advanced Models for Defense One way to mitigate risks from the proliferation of dangerous capabilities is to invest in defensive measures. In particular, large compute investors may be able to use their ongoing performance advantage to detect and address threats posed by irresponsible or malicious actors (Anderljung et al. 2023, p. 25). For instance, large compute investors may offer cybersecurity tools for automatic threat detection and response that defend against attacks by less capable models developed by rogue actors (Lohn and Jackson 2022). Similarly, automated detection of disinformation could limit the impact of AI on epistemic security (Demartini, Mizzaro, and Spina 2020). Even as a model approaches a performance ceiling, large compute investors may be able to provide effective defensive measures by leveraging their superior quantity of inference compute (e.g., by deploying more and/or faster model instances.[13])

Importance of Offense-Defense Balance Nonetheless, some applications of AI capabilities may inherently favor offensive use, making it difficult to defend against them even with advantages in performance or scale of deployment. Garfinkel and Dafoe (2019) reviewed how the balance between offense and defense scales with increased investment across a number of military scenarios and found a high variance, with some technologies, such as missiles, likely favoring the offense.

Some AI capabilities may similarly favor the offense. For instance, even a very capable protein predictor may not easily find cures for toxic agents developed by a somewhat weaker model.

[13]A model instance refers to a single copy of a particular AI model. Once a model is trained, the developer can run many such instances using the available supply of compute.

4 Limitations

We highlight two key limitations of our conceptual model. For a longer discussion of limitations see extended version at https://arxiv.org/abs/2311.15377.

4.1 Compute Investment Scaling and Proprietary Efficiency Advances

Our model has so far assumed that compute investment stays constant over time. However, large compute investors have historically scaled their compute investment significantly faster than others, widening the investment gap to smaller actors (Besiroglu et al. 2024). Further, since they often employ a high number of talented researchers, large compute investors may develop proprietary hardware and algorithmic advances. These advances further widen the gap between large compute investors and other actors.

4.2 Other Means of Model Access

So far, we assumed actors can only access the capabilities of the models they create themselves. However, large compute investors can provide smaller actors access to their advanced models through different means: While product integrations and structured access protocols allow for controlled and reversible access, fully releasing the parameters of an advanced model causes the full and irreversible diffusion of its capabilities (Seger et al. 2023).

5 Discussion

We now discuss some of the broader implications of increasing compute efficiency. Given their importance, we focus on the implications for mitigating risks from the emergence and proliferation of dangerous capabilities. However, we also recognize the need for a better understanding of the implications of increasing compute efficiency on economics and competition and encourage further research on this topic.

Since large compute investors discover dangerous capabilities first, we argue their decisions deserve particular scrutiny. Yet, society also needs to prepare for the risks coming from the proliferation of these dangerous capabilities. We motivate research aiming at assessing and forecasting AI models' ability to cause harm and encourage defensive measures that could limit the negative consequences of proliferation.

5.1 Responsibility of Large Compute Investors

Many of the most dangerous capabilities, such as hacking or social manipulation, likely first arise in the highest-performing general-purpose AI models, developed by a small number of large compute investors. In choosing to train such advanced models, large compute investors have a responsibility to identify novel capabilities and the risks they pose. To reliably detect potentially dangerous capabilities of their models, developers can design and apply evaluations that test whether models are capable of causing harm (Shevlane et al. 2023). Furthermore, developers can employ

risk assessment procedures that review more complex risks that may not directly arise from a model's capabilities (Novelli et al. 2023; Koessler and Schuett 2023).

While some frontier AI developers actively try to develop practices for preventing harm caused by their AI models (Anthropic 2023a), incentives for industry-directed initiatives may not always align with addressing the most severe risks. Therefore, governments should scrutinize the decisions and methodologies AI developers use, particularly targeting well-resourced labs that concentrate on creating highly capable foundation models (Anderljung et al. 2023).

Overseeing Large-Scale Compute Infrastructure Assuming that the worst risks stem from the most capable AI models and that these models' performance directly results from the size of their training compute budget, then overseeing access to large-scale AI compute clusters may present an effective method of regulation. Moreover, the amount of compute available determines not only the trained model's capabilities but also the number of model instances the developer can deploy. Controlling access to large-scale compute resources could thus allow governments not only to monitor the most advanced capabilities but also to quickly address harms caused by the large-scale deployment of potentially dangerous models (O'Brien, Ee, and Williams 2023; Egan and Heim 2023).

5.2 Addressing Proliferation

As the access effect gives more and more actors the ability to reproduce models with dangerous capabilities using smaller and smaller amounts of compute it becomes increasingly challenging to regulate the creation of such models.

AI researchers can help prepare society for this proliferation of dangerous capabilities by continually assessing SOTA AI models for their misuse potential and developing benchmarks and tests that could aid governments in determining the current and future risks associated with various AI models.

Frontier AI developers may play a particularly crucial role in identifying and mitigating risks associated with the proliferation of advanced model capabilities. Given their knowledge advantage from developing the most capable models and their research capacity due to designated safety and ethics teams, these labs are in a strong position to investigate proliferation risks. To allow society to prepare for the impact of proliferation, governments may mandate frontier AI developers to share information about their advanced models publicly or with selected government organizations (Mulani and Whittlestone 2023; Department for Science, Innovation and Technology 2023). With insights into possible risks, governments could then evaluate various measures aimed at either preventing model proliferation or addressing its risks effectively.

Besides limiting the harmful use of proliferated models, governments may also increase oversight of crucial inputs that allow malicious actors to cause harm based on knowledge provided by these models (Anderljung and Hazell 2023). For instance, controls on laboratory equipment and DNA synthesis screening could prevent malicious actors from creating dangerous pathogens even if the knowledge required is widely available (The White House 2023; DiEuliis, Carter, and Gronvall 2017).

Using Advanced Models for Defense Due to the performance effect, large compute investors continually create the most powerful models. The resulting performance advantage potentially enables them to deploy AI models defensively to counteract risks caused by proliferation. However, the feasibility of defending depends on several crucial factors, such as the offense-defense balance, the gap between leaders' and proliferated models, and the regulatory environment.

Offense-Defense Balance As discussed in Section 3.3, the offense-defense balance may vary greatly between different areas. Should proliferating AI capabilities significantly favor the offense, it may become infeasible to use other models for defense, even if the latter are more advanced and deployed on a larger scale.

Proportion of Leaders' Performance Advantage Besides the offense-defense balance, the effectiveness of using advanced AI models for defense also depends on the performance gap between the defensive models and the proliferated capabilities.

The gap is largest when defending against compute-limited actors like individuals, but well-resourced actors like authoritarian states or irresponsible companies could train AI models much closer in performance to the frontier. As capabilities approach performance ceilings or diminishing returns set in, the gap shrinks, limiting defense feasibility. However, even near a ceiling, leaders can potentially leverage their large compute investment and access to compute in order to deploy a large number of model instances for defense and thus maintain a substantial advantage over less well-resourced actors using proliferated models.

Regulatory Environment Beyond just technical feasibility, the effectiveness of defensive AI strategies may also depend on regulatory frameworks. Large compute investors may not may not be incentivized to develop defensive measures by default, so governments may need to contract them or initiate collaborative projects. Moreover, certain defense approaches may require explicit regulatory permission in order to be active in sensitive domains. For example, cybersecurity defenses might need direct access to sensitive networks and permission to make real-time decisions autonomously without human oversight — which brings new risks. Similarly, protecting individuals from social manipulation through AI may require monitoring private communications, raising significant privacy concerns. Governments, therefore, may need to develop and enforce standards that preserve privacy in these defensive models while still giving them sufficient permissions to ensure their effectiveness.

5.3 Limiting Development and Proliferation of Dangerous Capabilities

If AI models strongly favor offensive applications or are an inherent threat due to uncontrollability, or if defensive solutions are too invasive, governments may have to limit the

proliferation of such capabilities in the first place. Given the difficulty in accurately predicting these factors beforehand, it would be prudent to establish precautionary mechanisms to manage potential risks.

Misaligned AI Models So far, we primarily discussed how to address the misuse of advanced AI models. However, an increasing number of AI researchers emphasize the possibility that future, highly capable general AI models could pursue autonomous goals, thus posing an inherent threat (Carlsmith 2022; Ngo, Chan, and Mindermann 2023; Russell 2019; Center for AI Safety 2023). They argue it will be difficult to ensure such models are safe and controllable. Some frontier AI labs have already established designated research teams focused on developing solutions for the safety of autonomous AI models with near or above human-level capabilities (Leike and Sutskever 2023; Anthropic 2023b).

First, if advanced AI models are difficult to control, it could be hard to leverage them for defense. Additionally, the gap between frontier AI developers and others could narrow if frontier AI developers prioritize safety over capability advances. Even if frontier AI developers accurately assess that a model is uncontrollable and avoid deploying it, increasing compute efficiency would still diffuse the ability to create such models over time, greatly increasing the risk that someone eventually creates an uncontrollable model.

Balancing Scrutiny and Proliferation Risks The critical period between the first discovery of a dangerous capability and its proliferation to malicious actors is crucial for developing societal resilience through regulation or defensive solutions. Hence, developers should be cautious about hastening the diffusion of advanced AI models. In particular, publishing the parameters of advanced AI models causes the immediate, irreversible diffusion of the model's capabilities. Rather than allowing anyone unrestricted access to advanced models, frontier AI developers should introduce structured access procedures to provide their model parameters only to responsible researchers (Seger et al. 2023; Solaiman 2023; Shevlane 2022; Bucknall and Trager 2023).

6 Conclusion

We found that increased compute efficiency results both in an access effect — making a given capability more widely available — and a performance effect — enabling a higher performance level for a given compute investment. The impacts of the two effects depend on how a model's performance converts to usefulness in real-world tasks. Threshold effects advantage large compute investors, who are the first to discover qualitatively new capabilities. Meanwhile, performance ceilings reduce the gap between leading and lagging actors, whereas, in zero-sum competition, even a marginal advantage allows outsized benefits for the leaders.

Our model suggests that large compute investors are at the frontier of capabilities and, therefore, likely the first to discover dangerous capabilities. To adequately address these capabilities, large compute investors should implement extensive capability evaluation and risk assessment procedures.

As compute efficiency increases, dangerous capabilities proliferate to an increasing number of actors. The effectiveness of societal measures to mitigate harm from this proliferation hinges on the duration between frontier AI labs disclosing these capabilities and their proliferation to malicious or irresponsible actors. To extend this critical period, governments should implement information-sharing frameworks with leading AI labs and thoroughly assess the risks of proliferation. Such assessments may necessitate access to advanced models by governments or independent bodies.

Once informed about dangerous capabilities, governments should begin increasing societal resilience against them. This could involve restricting certain inputs required to cause harm, such as laboratory equipment or compounds needed to develop dangerous pathogens. Furthermore, governments can contract or coordinate with leading AI developers to use their advanced models in defensive solutions that address risks caused by proliferation. However, the feasibility of AI models used for defense critically depends on the performance difference between the defending model and the proliferated one, as well as the fundamental offense-defense balance in the field.

Although increasing compute efficiency makes AI capabilities more widely available over time, regulating access to large-scale compute clusters can still increase oversight. Specifically, monitoring the largest compute clusters allows monitoring the most extensive training runs, which likely produce the models with the most advanced and potentially dangerous AI capabilities.

However, both compute oversight and defensive solutions may be inadequate should sufficiently dangerous capabilities arise, such as AI models enabling individuals to create significant harm or if AI models of a given performance are inherently uncontrollable. Governments may need to preemptively develop mechanisms that could restrict the development and proliferation of intolerably dangerous models, such as moratoria on specific research areas to avoid large-scale harm.

Acknowledgements

We thank Aaron Tucker, Markus Anderljung, and Allan Dafoe for their paper "Social and Governance Implications of Improved Data Efficiency" which inspired this paper and first introduced many of the ideas presented here. We are further grateful for feedback and discussion of the ideas in this paper, including from Markus Anderljung, Ben Garfinkel, Ben Harack, and many other members of the GovAI team. members. We thank Wes Cowley for copy editing and José Medina for formatting assistance.

References

Ahmed, N.; and Wahed, M. 2020. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. Issue: arXiv:2010.15581 arXiv:2010.15581 [cs].

Amodei, D. 2023. Written Testimony of Dario Amodei, Ph.D. Co-Founder and CEO, Anthropic. Subcommittee on Privacy, Technology, and the Law, United States Senate.

Anderljung, M.; Barnhart, J.; Korinek, A.; Leung, J.; O'Keefe, C.; Whittlestone, J.; Avin, S.; Brundage, M.; Bullock, J.; Cass-Beggs, D.; Chang, B.; Collins, T.; Fist, T.; Hadfield, G.; Hayes, A.; Ho, L.; Hooker, S.; Horvitz, E.; Kolt, N.; Schuett, J.; Shavit, Y.; Siddarth, D.; Trager, R.; and Wolf, K. 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. Issue: arXiv:2307.03718 arXiv:2307.03718 [cs].

Anderljung, M.; and Hazell, J. 2023. Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted? Issue: arXiv:2303.09377 arXiv:2303.09377 [cs].

Anthropic. 2023a. Anthropic's Responsible Scaling Policy.

Anthropic. 2023b. Core Views on AI Safety: When, Why, What, and How.

Ardalani, N.; Wu, C.-J.; Chen, Z.; Bhushanam, B.; and Aziz, A. 2022. Understanding Scaling Laws for Recommendation Models. Issue: arXiv:2208.08489 arXiv:2208.08489 [cs].

AWS. 2023. AI Code Generator – Amazon CodeWhisperer – AWS.

Besiroglu, T.; Bergerson, S. A.; Michael, A.; Heim, L.; Luo, X.; and Thompson, N. 2024. The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny? arxiv:2401.02452.

Biais, B.; Foucault, T.; and Moinas, S. 2015. Equilibrium fast trading. *Journal of Financial Economics*, 116(2): 292–313. Number: 2.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. Issue: arXiv:2303.12712 arXiv:2303.12712 [cs].

Bucknall, B. S.; and Trager, R. F. 2023. Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements. Whitepaper, Oxford Martin AI Governance Initiative, Centre for the Governance of AI.

Byrne, D. M.; Oliner, S. D.; and Sichel, D. E. 2017. How Fast are Semiconductor Prices Falling? *Review of Income and Wealth*, 64(3): 679–702. Number: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12308.

Carlsmith, J. 2022. Is Power-Seeking AI an Existential Risk? Issue: arXiv:2206.13353 arXiv:2206.13353 [cs].

Center for AI Safety. 2023. Statement on AI Risk.

Character.AI. 2023. character.ai.

Chen, B.; Cheng, X.; Geng, Y.-a.; Li, S.; Zeng, X.; Wang, B.; Gong, J.; Liu, C.; Zeng, A.; Dong, Y.; Tang, J.; and Song, L. 2023. xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. Pages: 2023.07.05.547496 Section: New Results.

Creemers, R. 2018. China's Social Credit System: An Evolving Practice of Control. Issue: 3175792.

Demartini, G.; Mizzaro, S.; and Spina, D. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. In *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, volume 43. IEEE Computer Society.

Department for Science, Innovation and Technology. 2023. Emerging processes for frontier AI safety. Policy Paper, UK Government.

DiEuliis, D.; Carter, S. R.; and Gronvall, G. K. 2017. Options for Synthetic DNA Order Screening, Revisited. *mSphere*, 2(4): 10.1128/msphere.00319–17. Number: 4 Publisher: American Society for Microbiology.

Droppo, J.; and Elibol, O. 2021. Scaling Laws for Acoustic Models. Issue: arXiv:2106.09488 arXiv:2106.09488 [cs, eess].

Egan, J.; and Heim, L. 2023. Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers. Issue: arXiv:2310.13625 arXiv:2310.13625 [cs].

Epoch. 2022. Parameter, compute and data trends in machine learning. Tex.copyright: CC-BY.

Epoch. 2023. AI Trends.

Erdil, E.; and Besiroglu, T. 2023. Algorithmic progress in computer vision. Issue: arXiv:2212.05153 arXiv:2212.05153 [cs].

Ganguli, D.; Hernandez, D.; Lovitt, L.; DasSarma, N.; Henighan, T.; Jones, A.; Joseph, N.; Kernion, J.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Drain, D.; Elhage, N.; Showk, S. E.; Fort, S.; Hatfield-Dodds, Z.; Johnston, S.; Kravec, S.; Nanda, N.; Ndousse, K.; Olsson, C.; Amodei, D.; Amodei, D.; Brown, T.; Kaplan, J.; McCandlish, S.; Olah, C.; and Clark, J. 2022. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764. ArXiv:2202.07785 [cs].

Garfinkel, B.; and Dafoe, A. 2019. How does the offense-defense balance scale? *Journal of Strategic Studies*, 42(6): 736–763. Number: 6 Publisher: Routledge _eprint: https://doi.org/10.1080/01402390.2019.1631810.

Google. 2023. Bard can now connect to your Google apps and services.

Hernandez, D.; and Brown, T. B. 2020. Measuring the Algorithmic Efficiency of Neural Networks. Issue: arXiv:2005.04305 arXiv:2005.04305 [cs, stat].

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; and Zhou, Y. 2017. Deep Learning Scaling is Predictable, Empirically. Issue: arXiv:1712.00409 arXiv:1712.00409 [cs, stat].

Hobbhahn, M.; Heim, L.; and Aydos, G. 2023. Trends in Machine Learning Hardware.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; Driessche, G. v. d.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Rae, J. W.; Vinyals, O.; and Sifre, L. 2022. Training Compute-Optimal Large Language Models. Issue: arXiv:2203.15556 arXiv:2203.15556 [cs].

Koessler, L.; and Schuett, J. 2023. Risk assessment at AGI companies: A review of popular risk assessment

techniques from other safety-critical industries. Issue: arXiv:2307.08823 arXiv:2307.08823 [cs].

Kuchinke, B. A.; and Vidal, M. 2016. Exclusionary strategies and the rise of winner-takes-it-all markets on the Internet. *Telecommunications Policy*, 40(6): 582–592. Number: 6.

Leike, J.; and Sutskever, I. 2023. Introducing Superalignment.

Li, C. 2020. OpenAI's GPT-3 Language Model: A Technical Overview.

Liu, C.-L.; Nakashima, K.; Sako, H.; and Fujisawa, H. 2003. Handwritten digit recognition: benchmarking of state-of-the-art techniques. *Pattern Recognition*, 36(10): 2271–2285. Number: 10.

Lohn, A.; and Jackson, K. 2022. Will AI Make Cyber Swords or Shields?

McGregor, S. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15458–15463. Number: 17.

Mistral. 2023. Mixtral of Experts. https://mistral.ai/news/mixtral-of-experts/.

Mulani, N.; and Whittlestone, J. 2023. Proposing a Foundation Model Information-Sharing Regime for the UK | GovAI Blog. Research Post, Centre for the Governance of AI.

Nelson, C.; and Rose, S. 2023. Report launch: examining risks at the intersection of AI and bio. Research Report, Centre for Long-Term Resilience.

Ngo, R.; Chan, L.; and Mindermann, S. 2023. The alignment problem from a deep learning perspective. Issue: arXiv:2209.00626 arXiv:2209.00626 [cs].

NIST. 2023. Face Recognition Technology Evaluation (FRTE) 1:1 Verification.

Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M.; and Floridi, L. 2023. Taking AI risks seriously: a new assessment model for the AI Act. *AI & SOCIETY*.

O'Brien, J.; Ee, S.; and Williams, Z. 2023. Deployment corrections: An incident response framework for frontier AI models. Research Report, Institute for AI Policy and Strategy.

OpenAI. 2023. GPT-4.

OpenAI. 2023. GPT-4 Technical Report. Issue: arXiv:2303.08774 arXiv:2303.08774 [cs].

Papers with Code. 2023. Papers with Code - MMLU Benchmark (Multi-task Language Understanding).

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin. ISBN 978-0-525-55862-0.

Sandbrink, J. B. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. Issue: arXiv:2306.13952 arXiv:2306.13952 [cs].

Sedova, K.; McNeill, C.; Johnson, A.; Joshi, A.; and Wulkan, I. 2021. AI and the Future of Disinformation Campaigns. CSET Policy Brief, Centre for Security and Emerging Technology.

Seger, E.; Dreksler, N.; Moulange, R.; Dardaman, E.; Schuett, J.; Wei, K.; Winter, C.; Arnold, M.; Ó hÉigeartaigh, S.; Korinek, A.; Anderljung, M.; Bucknall, B.; Chan, A.; Stafford, E.; Koessler, L.; Ovadya, A.; Garfinkel, B.; Bluemke, E.; Aird, M.; Levermore, P.; Hazell, J.; and Gupta, A. 2023. Open-Sourcing Highly Capable Foundation Models. Research paper, Centre for the Governance of AI.

Sevilla, J.; Heim, L.; Ho, A.; Besiroglu, T.; Hobbhahn, M.; and Villalobos, P. 2022. Compute Trends Across Three Eras of Machine Learning. Issue: arXiv:2202.05924 arXiv:2202.05924 [cs].

Shevlane, T. 2022. Structured access: an emerging paradigm for safe AI deployment. Issue: arXiv:2201.05159 arXiv:2201.05159 [cs].

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; Ho, L.; Siddarth, D.; Avin, S.; Hawkins, W.; Kim, B.; Gabriel, I.; Bolina, V.; Clark, J.; Bengio, Y.; Christiano, P.; and Dafoe, A. 2023. Model evaluation for extreme risks. Issue: arXiv:2305.15324 arXiv:2305.15324 [cs].

Solaiman, I. 2023. The Gradient of Generative AI Release: Methods and Considerations. Issue: arXiv:2302.04844 arXiv:2302.04844 [cs].

The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Tucker, A. D.; Anderljung, M.; and Dafoe, A. 2020. Social and Governance Implications of Improved Data Efficiency. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 378–384. ArXiv:2001.05068 [cs].

Urbina, F.; Lentzos, F.; Invernizzi, C.; and Ekins, S. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3): 189–191. Number: 3 Publisher: Nature Publishing Group.

Venigalla, A.; and Li, L. 2022. Mosaic LLMs (Part 2): GPT-3 quality for <$500k.

Villalobos, P. 2023. Scaling Laws Literature Review.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. BloombergGPT: A Large Language Model for Finance. Issue: arXiv:2303.17564 arXiv:2303.17564 [cs, q-fin] version: 2.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. Issue: arXiv:2206.10789 arXiv:2206.10789 [cs].

Zhang, D.; Maslej, N.; Brynjolfsson, E.; Etchemendy, J.; Lyons, T.; Manyika, J.; Ngo, H.; Niebles, J. C.; Sellitto, M.; Sakhaee, E.; Shoham, Y.; Clark, J.; and Perrault, R. 2022. The AI Index 2022 Annual Report. Technical report, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University.