THE ROLE OF COMPUTE THRESHOLDS FOR AI GOVERNANCE

MATTEO PISTILLO,[†] SUZANNE VAN ARSDALE,[††] LENNART HEIM[†††] & CHRISTOPH WINTER[††††]

ABSTRACT

Advances in artificial intelligence ("AI") could bring transformative changes in society. AI has the potential for immense opportunities and benefits across a wide range of sectors, from healthcare and drug discovery to public services, and it could broadly improve productivity and living standards. However, more capable AI models also have the potential to cause extreme harm. AI could be misused for more effective disinformation, surveillance, cyberattacks, and development of chemical and biological weapons. More capable models are also likely to possess unexpected dangerous capabilities not yet observed in existing models. Laws can mitigate these risks, but in doing so must identify which models pose the greatest dangers and thus warrant regulatory attention.

This Article discusses the role of training compute thresholds, which use training compute to determine which potentially dangerous models are subject to legal requirements, such as reporting and evaluations. Since the amount of compute used to train a model corresponds to performance, with occasional surprising leaps, a training compute threshold (1) can be used to target the desired level of performance and corresponding risk. Several further properties of compute make it an attractive regulatory target: it is (2) essential for training, (3) objective and quantifiable, (4) capable of being estimated before training, and (5) verifiable after training. Since the amount of compute necessary to train cutting-edge models costs millions of dollars and usually relies on specialized hardware, training compute thresholds also (6) enable regulators to narrowly target potentially dangerous AI systems without burdening small companies, academic institutions, and individual researchers.

However, training compute thresholds are not infallible. Training compute is not an exhaustive measurement of risk; It does not track all risks posed by AI and is not a precise indicator of how harmful a model may be. Technological changes, such as algorithmic innovation, could also significantly reduce how much compute is needed to train an advanced model. For these reasons, a training compute threshold should be treated as a filter and a trigger for further scrutiny, rather

† Joint first author. Institute for Law & AI, Cambridge, MA, USA. Email: matteo.pistillo@law-ai.org.

†† Joint first author. Institute for Law & AI, Cambridge, MA, USA. Email: suzanne.vanarsdale@law-ai.org.

††† RAND Corporation, Washington, DC, USA. Email: lheim@rand.org.

†††† Institute for Law & AI, Cambridge, MA, USA / Harvard University, Cambridge, MA, USA / University of Cambridge, Cambridge, England. Email: christoph_winter@fas.harvard.edu. We are grateful to Cullen O'Keefe, Mauricio Baker, Matthijs Maas, Charles Bullock, Daniel Bateyko, Mackenzie Arnold, and Leonie Koessler for comments, discussion, and critique. All comments were made in a personal capacity.

than an end in and of itself, and accompanied by a mechanism for updating the threshold.

Indeed, the United States and the European Union ("EU") have recognized the significance of compute in recent initiatives, which seek to ensure the safe and responsible development of AI in part by establishing training compute thresholds that trigger reporting requirements, capability evaluations, and incident monitoring. Beyond this, courts and regulators could rely on compute as an indicator of how much risk a given AI system poses when determining whether a legal condition or regulatory threshold has been met. Compute may play a role as an indicator of foreseeability of harm under tort law, as a proxy for threat to national or public security in risk assessments, or as a factor in regulatory impact analysis.

TABLE OF CONTENTS

I. INTRODUCTION

The idea of establishing a "compute threshold" and, more precisely, a "training compute threshold" has recently attracted significant attention from policymakers and commentators. In recent years, various scholars and AI labs have supported setting such a threshold,[1] as have governments around the world. On October 30, 2023, President Biden's Executive Order 14,110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence introduced the first living example of a compute threshold,[2] although it was one of many orders revoked by President Trump upon entering office.[3] The European Parliament and the European Council adopted the Artificial Intelligence Act, on June 13, 2024, providing for the establishment of a compute threshold.[4] On February 4, 2024, California State Senator Scott Wiener introduced Senate Bill 1047 that defined frontier AI models with a compute threshold.[5] The bill was approved by the California legislature, but it was ultimately vetoed by the State's Governor.[6] China may be

---

[1] For examples of scholars supporting the establishment of a training compute threshold, see Gillian Hadfield et al., *It's Time to Create a National Registry for Large AI Models*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 12, 2023), https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180 [https://perma.cc/DJJ2-HMEV]; Janet Egan & Lennart Heim, *Oversight for Frontier AI Through a Know-Your-Customer Scheme for Compute Providers*, ARXIV 3 (Oct. 20, 2023), https://doi.org/10.48550/arXiv.2310.13625 [https://perma.cc/Q2RM-927X]; Andrea Miotti & Akash Wasil, *Taking Control: Policies to Address Extinction Risks from Advanced AI*, ARXIV 9–11 (Oct. 31, 2023), https://doi.org/10.48550/arXiv.2310.20563 [https://perma.cc/FE27-RE63]; Sarah Bauerle Danzman et al., Comment Letter on Advance Notice of Proposed Rulemaking Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern (Sep. 29, 2023) [hereinafter Comment on ANPRM], https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/TREAS-DO-2023-0009-0049_attachment_1.pdf [https://perma.cc/J4Y3-PG7E], at 16–18; Sarah Bauerle Danzman et al., Comment Letter on Proposed Rule Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern (Aug. 4, 2024) [hereinafter Comment on Proposed Rule], https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/TREAS-DO-2024-0012-0041_attachment_1.pdf [https://perma.cc/2BFT-GWBF]; *see also* Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, ARXIV 9, 35–37 (Nov. 7, 2023), https://doi.org/10.48550/arXiv.2307.03718 [https://perma.cc/N62P-MKA4] (identifying compute thresholds as one of the options for defining a model's possibility of producing sufficiently dangerous capabilities); Kayla Matteucci et al., *AI Systems of Concern*, ARXIV 6 (Oct. 9, 2023), https://doi.org/10.48550/arXiv.2310.05876 [https://perma.cc/99PS-UDMV] (identifying compute as one of the potential indicators to identify and detect systems of concern). For examples of AI labs proposing such thresholds, see Sam Altman et al., *Governance of Superintelligence,* OPENAI (May 22, 2023), https://openai.com/blog/governance-of-superintelligence [https://perma.cc/VX72-JN2S] (proposing the introduction of a "capability (or resources like compute) threshold" as a "starting point" for the governance of superintelligence); Microsoft, *Governing AI: A Blueprint for the Future* (May 25, 2023), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw [https://perma.cc/BJ9Q-CXYR], at 21 (suggesting that a compute threshold may be "the best option on offer today" to define the material scope of regulated AI models).

[2] Exec. Order No. 14,110, § 4.2(b)–(c), 3 C.F.R. § 14110 (2024) (revoked by Exec. Order No. 14,148, § 2(ggg), 90 Fed. Reg. 8237 (Jan. 20, 2025)) [hereinafter Exec. Order on AI].

[3] Exec. Order No. 14,148, § 2(ggg), 90 Fed. Reg. 8237 (Jan. 20, 2025).

[4] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, art. 51(2). 2024 O.J. (L 144) 1, 83 [hereinafter EU AI Act].

[5] As introduced, the bill defined "covered models" to include models "trained using a quantity of computing power greater than 10^26 integer or floating-point operations." S.B. 1047, 2023–2024 Reg. Sess. (Cal. 2024) § 3 (as introduced in Senate, Feb. 7, 2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047 (choose "02/07/24 - Introduced" from dropdown"; then click "Go") [https://perma.cc/Q49X-M9JX]. The bill that ultimately passed in the Senate and Assembly additionally required the cost of compute to exceed $100 million and created a new category of "covered models," defined as those "created by fine-tuning a covered model using a quantity of computing power equal to or greater than three times 10^25 integer or floating-point operations, the cost of which, as reasonably assessed by the developer, exceeds ten million dollars ($10,000,000)." S.B. 1047, 2023–2024 Reg. Sess. (Cal. 2024) § 3 (as enrolled, Sept. 3, 2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047 (choose "09/03/24 - Enrolled" from dropdown"; then click "Go") [https://perma.cc/Y8GQ-8U95].

[6] *See* Office of Governor Gavin Newsom, *Governor Newsom Announces New Initiatives to Advance Safe and Responsible AI, Protect Californians* (Sept. 29, 2024), https://www.gov.ca.gov/2024/09/29/governor-newsom-announces-new-initiatives-to-advance-safe-and-responsible-ai-protect-californians/ [https://perma.cc/3VQJ-5PHW]; Office of Governor Gavin Newsom, *Veto Message* (Sept. 29, 2024), https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf [https://perma.cc/L6YC-J6VF].

considering similar measures, as indicated by recent discussions in policy circles.[7] While not perfect, compute thresholds are currently one of the best options available to identify potentially high-risk models and trigger further scrutiny. Yet, in spite of this, information about compute thresholds and their relevance from a policy and legal perspective remains dispersed.

This Article proceeds in two parts. Part I provides a technical overview of compute and how the amount of compute used in training corresponds to model performance and risk. It begins by explaining what compute is and the role compute plays in AI development and deployment. Compute refers to both computational infrastructure, the hardware necessary to develop and deploy an AI system, and the amount of computational power required to train a model, commonly measured in integer or floating-point operations. More compute is used to train notable models each year, and although the cost of compute has decreased, the amount of compute used for training has increased at a higher rate, causing training costs to increase dramatically.[8] This increase in training compute has contributed to improvements in model performance and capabilities, described in part by scaling laws. As models are trained on more data, with more parameters and training compute, they grow more powerful and capable. As advances in AI continue, capabilities may emerge that pose potentially catastrophic risks if not mitigated.[9]

Part II discusses why, in light of this risk, compute thresholds may be important to AI governance. Since training compute can serve as a proxy for the capabilities of AI models, a compute threshold can operate as a regulatory trigger, identifying what subset of models *might* possess more powerful and dangerous capabilities that warrant greater scrutiny, such as in the form of reporting and evaluations. Both the European Union AI Act and Executive Order 14,110 established compute thresholds for different purposes, and many more policy proposals rely on compute thresholds to ensure that the scope of covered models matches the nature or purpose of the policy. This Part provides an overview of policy proposals that expressly call for such a threshold, as well as proposals that could benefit from the addition of a compute threshold to clarify the scope of policies that refer broadly to "advanced systems" or "systems with dangerous capabilities." It then describes how, even absent a formal compute threshold, courts and regulators might rely on training compute as a proxy for how much risk a given AI system poses, even under existing law. This Part concludes with the advantages and limitations of using compute thresholds as a regulatory trigger.

---

[7] *See* Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars), CHINA L. SOC'Y. (Mar. 18, 2024), http://www.fxcxw.org.cn/dyna/content.php?id=26910 [https://perma.cc/5P7A-G7PE], art. 50(iii), art. 50–57, translated at Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars), CTR. FOR SEC. & EMERGING TECH. (May 2, 2024), https://cset.georgetown.edu/publication/china-ai-law-draft/ [https://perma.cc/SUX6-4DGA] ("Foundation models that have reached a certain level in aspects such as compute, parameters, or scale of use"); Matt Sheehan (@mattsheehan88), X (Mar. 21, 2024, 3:55 PM), https://x.com/mattsheehan88/status/1770902104795729936 [https://perma.cc/75UT-2B5J].

[8] This roughly follows Moore's Law. *See infra* Sec. I.B.

[9] *See infra* Sec. I.D.

II. COMPUTE AND THE SCALING HYPOTHESIS

A. What Is "Compute"?

The term "compute" serves as an umbrella term, encompassing several meanings that depend on context.

Commonly, the term "compute" is used to refer to *computational infrastructure*, i.e., the hardware stacks necessary to develop and deploy AI systems.[10] Many hardware elements are integrated circuits (also called chips or microchips), such as logic chips, which perform operations, and memory chips, which store the information on which logic devices perform calculations.[11] Logic chips cover a spectrum of specialization, ranging from general-purpose central processing units ("CPUs"), through graphics processing units ("GPUs") and field-programmable gate arrays ("FPGAs"), to application-specific integrated circuits ("ASICs") customized for specific algorithms.[12] Memory chips include dynamic random-access memory ("DRAM"), static random-access memory ("SRAM"), and NOT AND ("NAND") flash memory used in many solid state drives ("SSDs").[13]

Additionally, the term "compute" is often used to refer to how much computational power is required to train a specific AI system. Whereas the *computational performance* of a chip refers to how quickly it can execute operations and thus generate results, solve problems, or perform specific tasks, such as processing and manipulating data or training an AI system, "compute" refers to the *amount of computational power* used by one or more chips to perform a task, such as training a model. Compute is commonly measured in integer operations or floating-point operations ("OP" or "FLOP"),[14] expressing the *number of operations* that have been executed by one or more chips,

---

[10] Throughout this Article, the term "compute" refers specifically to "AI compute"—that is, the computational infrastructure that is specialized for AI development and deployment. *See* Organization for Economic Co-operation and Development (OECD), *A Blueprint for Building National Compute Capacity for Artificial Intelligence* (OECD Digital Economy Paper No. 350, 2023), https://doi.org/10.1787/876367e3-en [https://perma.cc/AAK2-SZ4D], at 20 ("AI computing resources ('AI compute') include one or more stacks of hardware and software used to support specialized AI workloads and applications in an efficient manner."); *see also* Saif M. Khan & Alexander Mann, *AI Chips: What They Are and Why They Matter*, CTR. FOR SEC. & EMERGING TECH. (Apr. 2020), https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/ [https://perma.cc/UMH3-X8ZF] (providing an overview of AI chips).

[11] Khan & Mann, *supra* note 10, at 33.

[12] *Id.* at 4–6, 20–21, 32–37 ("Different types of AI chips are useful for different tasks. GPUs are most often used for initially developing and refining AI algorithms; this process is known as 'training.' FPGAs are mostly used to apply trained AI algorithms to real world data inputs; this is often called 'inference.' ASICs can be designed for either training or inference."); *see also* Tim Hwang, *Computational Power and the Social Impact of Artificial Intelligence*, ARXIV 1 (Mar. 23, 2018), https://doi.org/10.48550/arXiv.1803.08971 [https://perma.cc/29GR-YSY9]; Konstantin Pilz & Lennart Heim, *Compute at Scale—A Broad Investigation into the Data Center Industry*, ARXIV 1 (Nov. 22, 2023), https://doi.org/10.48550/arXiv.2311.02651 [https://perma.cc/9VE6-4JHK]; *cf.* U.K., Dep't for Sci., Innovation & Tech., *Independent Review of The Future of Compute: Final Report and Recommendations* (Mar. 6, 2023), https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts [https://perma.cc/NL93-TPUZ] (defining compute as "computer systems where processing power, memory, data storage, and network are assembled at scale to tackle computational tasks beyond the capabilities of everyday computers").

[13] Khan & Mann, *supra* note 10, at 33.

[14] Integer and floating-point operations are specific kinds of arithmetic operations. Integer operations are basic arithmetic operations performed only with integers. Exec. Order on AI, *supra* note 2, § 3(r). Floating-point operations (FLOP) are basic arithmetic operations performed with numbers in floating-point notation. Floating-point numbers are a subset of the real numbers typically represented on computers by an integer of fixed precision scaled by an integer exponent of a fixed base (e.g., $12.345 = 12345 \times 10^{-3}$). Exec. Order on AI, *supra* note 2, § 3(m) ("The term 'floating-point operation' means any mathematical operation or assignment involving floating-point numbers, which are a subset of the real numbers typically represented on computers by an integer of fixed precision scaled by an integer exponent of a fixed base."). The compute threshold in Executive Order 14,110 refers to both integer operations and FLOP. Exec. Order on AI, *supra* note 2, § 4.2(ii) ("any model that was trained using a quantity of computing power greater than 10^{26} integer or floating-point

while the computational performance of those chips is measured in operations per second ("OP/s" or "FLOP/s"). In this sense, the amount of computational power used is roughly analogous to the distance traveled by a car.[15] Since large amounts of compute are used in modern computing, values are often reported in scientific notation such as 1e26 or 2e26, which refer to $1 \cdot 10^{26}$ and $2 \cdot 10^{26}$ respectively.

Compute is essential throughout the AI lifecycle. The AI lifecycle can be broken down into two phases: development and deployment.[16] In the first phase, *development*, developers design the model by choosing an architecture, the structure of the network, and initial values for hyperparameters (i.e., parameters that control the learning process, such as number of layers and training rate).[17] Enormous amounts of data, usually from publicly available sources, are processed and curated to produce high-quality datasets for training.[18] The model then undergoes "pre-training," in which the model is trained on a large and diverse dataset in order to build the general knowledge and features of the model, which are reflected in the weights and biases of the model.[19] Alternatively, developers may use an existing pre-trained model, such as OpenAI's GPT-4 ("Generative Pre-trained Transformer 4"). The term "foundation model" refers to models like these, which are trained on broad data and adaptable to many downstream tasks.[20] Performance and capabilities improvements are then possible using methods such as fine-tuning on task-specific datasets, reinforcement learning from human feedback ("RLHF"), teaching the model to use tools,

operations"). In contrast, the EU AI Act only refers to FLOP. EU AI Act, *supra* note 4, art. 51(2).

[15] The number of operations should not be confused with the speed of a chip, which is rather comparable to a car's travel speed (nor with the theoretical peak performance of a chip, which is comparable to a car's maximum travel speed). Speed does not explain the distance that a car has traveled, but only how fast a car can travel a given distance. Similarly, the speed of a chip does not explain the number of operations that a chip has performed.

[16] *See* U.K. Competition & Markets Authority, *AI Foundation Models: Initial Report* (Sept. 18, 2023), at 1, 10–12, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1185508/ Full_report_.pdf [https://perma.cc/M2TN-V7J6]; *see also* OECD, *supra* note 10, at 22 (defining the lifecycle as encompassing six phases: "(1) plan and design; (2) collect and process data; (3) build and use the model; (4) verify and validate the model; (5) deploy; and (6) operate and monitor the system"), citing *OECD Framework for the Classification of AI Systems* (OECD Digital Economy Papers, No. 323, 2022), https://doi.org/10.1787/cb6d9eca-en [https://perma.cc/F59S-TYMN], at 7, 22–23, and Figure 4 at 23 (noting that the phases "are not necessarily sequential").

[17] *See, e.g.*, Kizito Nyuytiymbiy, *Parameters and Hyperparameters in Machine Learning and Deep Learning*, TOWARDS DATA SCI. (Dec. 30, 2020).

[18] *See* Humza Naveed et al., *A Comprehensive Overview of Large Language Models*, ARXIV 5 (Oct. 17, 2024), https://doi.org/10.48550/arXiv.2307.06435 [https://perma.cc/4B5M-ETS4] (summarizing three data preprocessing techniques used for large language models: quality filtering, data deduplication, and privacy reduction); *cf.* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, ARXIV 8–9 & tbl.2.2 (July 22, 2020), https://doi.org/10.48550/arXiv.2005.14165 [https://perma.cc/7JK3-JQJ7] (noting that OpenAI filtered the Common Crawl dataset down from 45TB to 570GB, and that the curated dataset was used for 60% of the examples during training). Data can also be filtered in other ways, such as to remove personal information (such as names, addresses, and phone numbers), Naveed et al., at 6, or to reduce bias, L. Elisa Celis et al., *Data Preprocessing To Mitigate Bias: A Maximum Entropy Based Approach*, ARXIV (June 30, 2020), https://doi.org/10.48550/arXiv.1906.02164 [https://perma.cc/B9PF-5AMK] (discussing use of data preprocessing to mitigate bias from data containing human or social attributes that over- or under-represent certain groups).

[19] *See* Jishnu Mukhoti et al., *Fine-tuning Can Cripple Your Foundation Model; Preserving Features May Be the Solution*, ARXIV 2 (July 1, 2024), https://doi.org/10.48550/arXiv.2308.13320 [https://perma.cc/HM5D-8RL3] ("[T]he pre-training dataset of a foundation model, owing to its massive scale, contains information about several thousands of real-world concepts."); *see generally* Haifent Wang et al., *Pre-Trained Language Models and Their Applications*, 25 ENG'G 51 (2023); Dan Hendrycks et al., *Using Pre-Training Can Improve Model Robustness and Uncertainty*, ARXIV (Oct. 20, 2019), https://doi.org/10.48550/arXiv.1901.09960 [https://perma.cc/LXZ5-2PBQ] (describing the advantages of pre-training compared to training from scratch); Dumitru Erhan et al., *Why Does Unsupervised Pre-training Help Deep Learning?*, 11 J. OF MACH. LEARNING RSCH. 625 (Feb. 2010) (noting that it can be faster and more cost-effective to begin with one of the many pre-trained models available).

[20] *See* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ARXIV 3, 6–7 (July 12, 2022), https://doi.org/10.48550/arXiv.2108.07258 [https://perma.cc/DTY2-TYHQ].

and instruction tuning.[21] These enhancements are far less compute-intensive than pre-training, particularly for models trained on massive datasets.[22]

As of this writing, there is no agreed-upon standard for measuring "training compute." Estimates of "training compute" typically refer only to the amount of compute used during pre-training. More specifically, they refer to the amount of compute used during the final pre-training run, which contributes to the final machine learning model, and does not include any previous test runs or post-training enhancements, such as fine-tuning.[23] There are exceptions: for instance, the EU AI Act considers the *cumulative* amount of compute used for training by including all the compute "used across the activities and methods that are intended to enhance the capabilities of the model prior to deployment, such as pre-training, synthetic data generation and fine-tuning."[24] California Senate Bill 1047 addressed post-training modifications generally and fine-tuning in particular, providing that a covered model fine-tuned with more than 3e25 OP or FLOP would be considered a distinct "covered model," while one fine-tuned on less compute or subjected to unrelated post-training modifications would be considered a "covered model derivative."[25]

In the second phase, *deployment*, the model is made available to users and is used.[26] Users provide input to the model, such as in the form of a prompt, and the model makes predictions from

---

[21] Tom Davidson et al., *AI Capabilities Can Be Significantly Improved Without Expensive Retraining*, ARXIV (Dec. 12, 2023), https://doi.org/10.48550/arXiv.2312.07413 [https://perma.cc/N7TD-DSQY] (reviewing post-training enhancements and categorizing them as tool use, prompting methods, scaffolding, solution selection, and data generation); *see also* Paul Christiano et al., *Deep Reinforcement Learning from Human Preferences*, ARXIV (Feb. 17, 2023), https://doi.org/10.48550/arXiv.1706.03741 [https://perma.cc/RVY7-CJVV]; *see also*, OpenAI, *GPT-4 Technical Report*, ARXIV 12–13 (Mar. 4, 2024), https://doi.org/10.48550/arXiv.2303.08774 [https://perma.cc/ME4F-52XV] (noting that GPT-4 and prior models were fine-tuned using reinforcement learning from human feedback (RLHF) to "produce responses better aligned with the user's intent" and produce less harmful content); Shengyu Zhang et al., *Instruction Tuning for Large Language Models: A Survey*, ARXIV (Dec. 1, 2024), https://doi.org/10.48550/arXiv.2308.10792 [https://perma.cc/S6YA-QQ3Q].

[22] *See, e.g.*, Davidson et al., *supra* note 21, at 1 (noting that "fine-tuning costs are typically <1% of the original training cost."); Evani Radiya-Dixit & Xin Wang, *How Fine Can Fine-tuning Be? Learning Efficient Language Models*, ARXIV 1 (Apr. 24, 2020), https://doi.org/10.48550/arXiv.2004.14129 [https://perma.cc/CRT9-L5WC] ("Given a language model pre-trained on massive unlabeled text corpora, only very light supervised fine-tuning is needed to learn a task: the number of fine-tuning steps is typically five orders of magnitude lower than the total parameter count."); *see also Notable AI Models*, EPOCH (July 23, 2024), https://epochai.org/data/notable-ai-models [https://perma.cc/2GUD-UEWD] (reporting different estimates of pre-training compute for different models).

[23] *See* Jaime Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*, ARXIV 16 (Mar. 9, 2022) [hereinafter *Compute Trends*], https://doi.org/10.48550/arXiv.2202.05924 [https://perma.cc/GJ48-E64B] ("ML systems are often trained multiple times to choose better hyperparameters (e.g., number of layers or training rate). However, this information is often not reported in papers. Our dataset only annotates the compute used for the final training run."); Jaime Sevilla et al., *Compute Trends Across Three Eras of Machine Learning*, EPOCH (May 2, 2022) [hereinafter *Compute Trends Summary*], https://epochai.org/blog/compute-trends [https://perma.cc/642K-3YBN], at n.1 ("[W]e focus on the final training run of a ML system. This is primarily due to measurability—researchers generally do not mention the total compute or training time that does not directly contribute to the final machine learning model. We simply do not have sufficient information to determine the total compute through the entire experimentation process."); *see also* Neil C. Thompson et al., *The Computational Limits of Deep Learning*, ARXIV 6 (supplemental materials) (July 27, 2022), https://doi.org/10.48550/arXiv.2007.05558 [https://perma.cc/66GT-SWT6] (noting that "[t]o find all the data needed to estimate the computing power used to train a model can be quite challenging" due, for example, to only estimates being reported, certain data not being reported precisely, and errors or inconsistency in data sources); Jaime Sevilla et al., *Estimating Training Compute of Deep Learning Models*, EPOCH (Jan. 20, 2022) [hereinafter *Estimating Training Compute*], https://epochai.org/blog/estimating-training-compute [https://perma.cc/B3RT-9S4Q], app. C ("It is common to pre-train a large model on a large dataset and then fine-tune it on a smaller dataset. Similarly, it is common for researchers to manually train and tweak multiple versions of a system before they find the final architecture they use for training. We recommend counting the pre-training compute as part of the total training compute. However we do not recommend counting the tweak runs. While these are important, for reproducibility purposes it is the pre-training and fine-tuning of the final architecture that matters most. And pragmatically speaking information on the compute used to train previous versions while finding the right architecture is seldom reported.").

[24] EU AI Act, *supra* note 4, at Recital 111 & art. 51(2).

[25] S.B. 1047, 2023–2024 Reg. Sess. (Cal. 2024) § 3 (as enrolled, Sept. 3, 2024).

[26] U.K. Competition & Markets Authority, *supra* note 16, at 14–16 & fig.3.

this input in a process known as "inference."[27] The amount of compute needed for a single inference request is far lower than what is required for a training run.[28] However, for systems deployed at scale, the cumulative compute used for inference can surpass training compute by several orders of magnitude.[29] Consider, for instance, a large language model ("LLM"). During training, a large amount of compute is required over a smaller time frame within a closed system, usually a supercomputer. Once the model is deployed, each text generation leverages its own copy of the trained model, which can be run on a separate compute infrastructure. The model may serve hundreds of millions of users, each generating unique content and using compute with each inference request. Over time, the cumulative compute usage for inference can surpass the total compute required for training.

There are various reasons to consider compute usage at *different stages* of the AI lifecycle, which is discussed in Section I.E. For clarity, this Article uses "training compute" for compute used during the final pre-training run and "inference compute" for compute used by the model during a single inference, measured in the number of operations ("OP" or "FLOP"). Figure 1 illustrates a simplified version of the language model compute lifecycle.

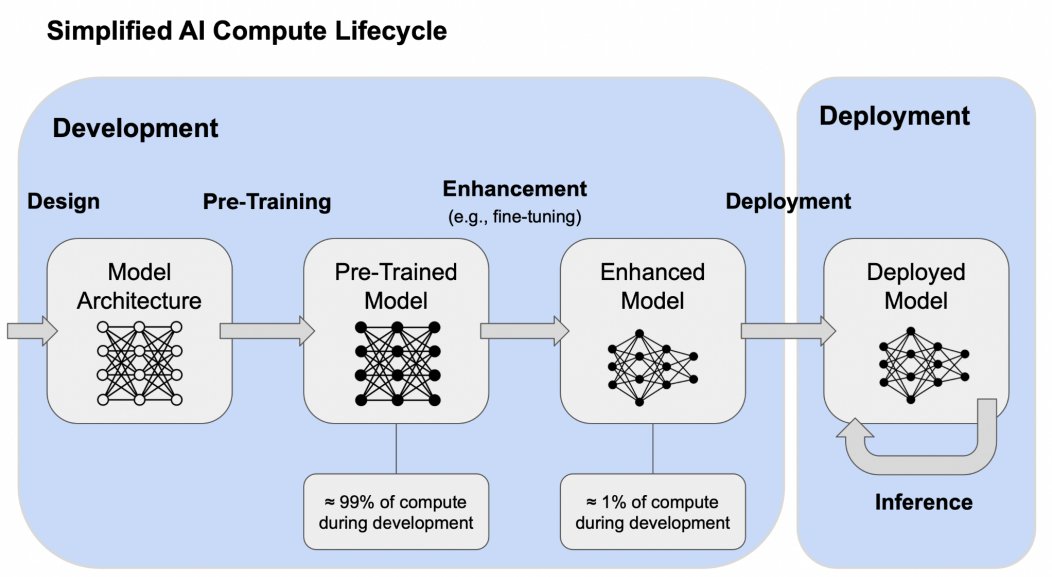

Figure 1: Simplified language model lifecycle

[27] *Id.* at n.22.

[28] Pablo Villalobos & David Atkinson, *Trading Off Compute in Training and Inference*, EPOCH (July 28, 2023), https://epochai.org/blog/trading-off-compute-in-training-and-inference [https://perma.cc/GE7N-QLYB] ("The cost of running a single inference is much smaller than the cost of the training process. A good rule of thumb is that the cost of an inference is close to the square root of the cost of training [], albeit with significant variability . . . . For example, for GPT-3, the cost of training was 3e23 FLOP, whereas the cost of a single inference is 3e11. So the cost of training is equivalent to performing 1e12 inferences."). Both training and inference compute correspond to the number of parameters in the model and size of the training dataset. *Id.*

[29] *Id.*; Dario Amodei & Danny Hernandez, *AI and Compute*, OPENAI (May 16, 2018), https://openai.com/index/ai-and-compute/ [https://perma.cc/Q4TA-SFCK] ("[T]he majority of neural net compute today is still spent on inference (deployment), not training."); OECD, *supra* note 10, at 22, citing IAN GOODFELLOW ET AL., DEEP LEARNING (2016) ("[W]hile a single training run is more computationally intensive than a single inference, the inferencing stage overall typically requires more compute in an AI system's lifecycle because ML systems are usually trained only a few times during their development phase, whereas inferencing is executed repeatedly every time a system is used during the lifetime of its deployment.").

B. What Is Moore's Law and Why Is It Relevant for AI?

In 1965, Gordon Moore forecasted that the number of transistors on an integrated circuit would double every year.[30] Ten years later, Moore revised his initial forecast to a two-year doubling period.[31] This pattern of exponential growth is now called "Moore's Law."[32] Similar rates of growth have been observed in related metrics, notably including the increase in computational performance of supercomputers;[33] as the number of transistors on a chip increases, so does computational performance (although other factors also play a role).[34]

A corollary of Moore's Law is that the cost of compute has fallen dramatically; a dollar can buy more FLOP every year.[35] Greater access to compute, along with greater spending from 2010 onwards (i.e., the so-called deep learning era),[36] has contributed to developers using ever more compute to train AI systems. Research has found that the compute used to train notable and frontier models has grown by 4–5x per year between 2010 and May 2024.[37]

---

[30] Gordon E. Moore, *Cramming More Components onto Integrated Circuits*, 38 ELECS. 114, 115 (1965). Transistors are one of the building blocks of modern electronic devices: small, electrical devices that contain a semiconductor material (such as silicon or germanium) and are used to amplify, control, and generate electrical signals.

[31] Gordon E. Moore, *Progress in Digital Integrated Electronics*, TECH. DIG. (1975), at 11–13. The frequently cited prediction of an 18-month doubling time was made by Intel executive David House, by considering not just the number of transistors, but also improvements in transistor speed. Michael Kanellos, *Moore's Law to Roll on for Another Decade*, CNET (Feb. 11, 2003), https://www.cnet.com/tech/tech-industry/moores-law-to-roll-on-for-another-decade/ [https://perma.cc/4F4P-XY3E].

[32] Ethan R. Mollick, *Establishing Moore's Law*, 28(3) IEEE ANNALS OF THE HISTORY OF COMPUTING 62–75 (July 2006).

[33] Max Roser, Hannah Ritchie & Edouard Mathieu, *What Is Moore's Law?*, OUR WORLD IN DATA (Mar. 28, 2023), https://ourworldindata.org/moores-law [https://perma.cc/C5J2-RC6Y].

[34] Henry Kressel, *The End of Moore's Law? Innovation in Computer Systems Continues at a High Pace*, ARTIFICIAL INTELLIGENCE IN SCIENCE: CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH (June 26, 2023), https://doi.org/10.1787/63e48242-en [https://perma.cc/V9J2-YHJF] ("The computing power of a system is a function of the available transistor capacity, the speed of transistor switching . . . , memory volume and interconnection speed."); *cf.* Marius Hobbhahn et al., *Trends in Machine Learning Hardware*, EPOCH (Nov. 9, 2023), https://epochai.org/blog/trends-in-machine-learning-hardware [https://perma.cc/Q6SQ-GED3] (suggesting that transistors count is a useful but imperfect metric of computational performance, as shown by the fact that the doubling time of the number of transistors, estimated at 2.89 years, is slightly slower than that of peak computational performance, estimated at 2.3 years).

[35] Gregory Arcuri & Sujai Shivakumar, *Moore's Law and Its Practical Implications*, CTR. FOR STRATEGIC & INT'L STUD. (Oct. 18, 2022), https://www.csis.org/analysis/moores-law-and-its-practical-implications [https://perma.cc/7V4G-3RAN]; Hobbhahn et al., *supra* note 34 (finding that the price-performance ratio, expressed in FLOP/$, has doubled every 2.1 years for machine learning GPUs and 2.5 years for general GPUs from 2004 to 2024).

[36] Although the cost of compute has decreased, the amount of compute used to train cutting-edge models has increased faster, causing training costs to increase dramatically. Neil Thompson et al., *The Importance of (Exponentially More) Computing Power*, ARXIV 16 (June 28, 2022), https://doi.org/10.48550/arXiv.2206.14007 [https://perma.cc/Z5J2-RZUP] ("Even after accounting for rapid hardware improvement rates, all [domains of AI studied] have shown enormous increases in the cost of the computing power being used;" however, "costs have not risen proportionally to these increases, principally because Moore's Law provided ever-cheaper computing power."); Thompson et al., *supra* note 23, at 4 (noting that in the 1960s and decades that followed, "the economic cost of running such models was largely stable over time" as the cost of compute decreased proportionally with the increase in compute requirements of the largest systems, but later "the amount of computing power used in the largest cutting-edge systems grew even faster, at approximately 10x per year from 2012 to 2019," at greater monetary cost); Ben Cottier, *Trends in the Dollar Training Cost of Machine Learning Systems*, EPOCH (Jan. 31, 2023), https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems [https://perma.cc/SL5T-7BDH] (finding that between 2009 and 2022 the cost of compute for the final training for notable models grew by approximately 0.5 orders of magnitude per year).

[37] Jaime Sevilla & Edu Roldán, *Training Compute of Frontier AI Models Grows by 4-5x Per Year*, EPOCH (May 28, 2024), https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year [https://perma.cc/X9RW-DPTU]; *see also Notable AI Models*, *supra* note 22 (dataset). This rate of growth is equivalent to training compute doubling every 5.2 to 6 months. For a discussion of earlier estimates, *see Compute Trends*, *supra* note 23, at 2 & tbl.3 (discussing earlier estimates and estimating that training compute for notable models doubled every 5.6 months between 2010 and 2022).

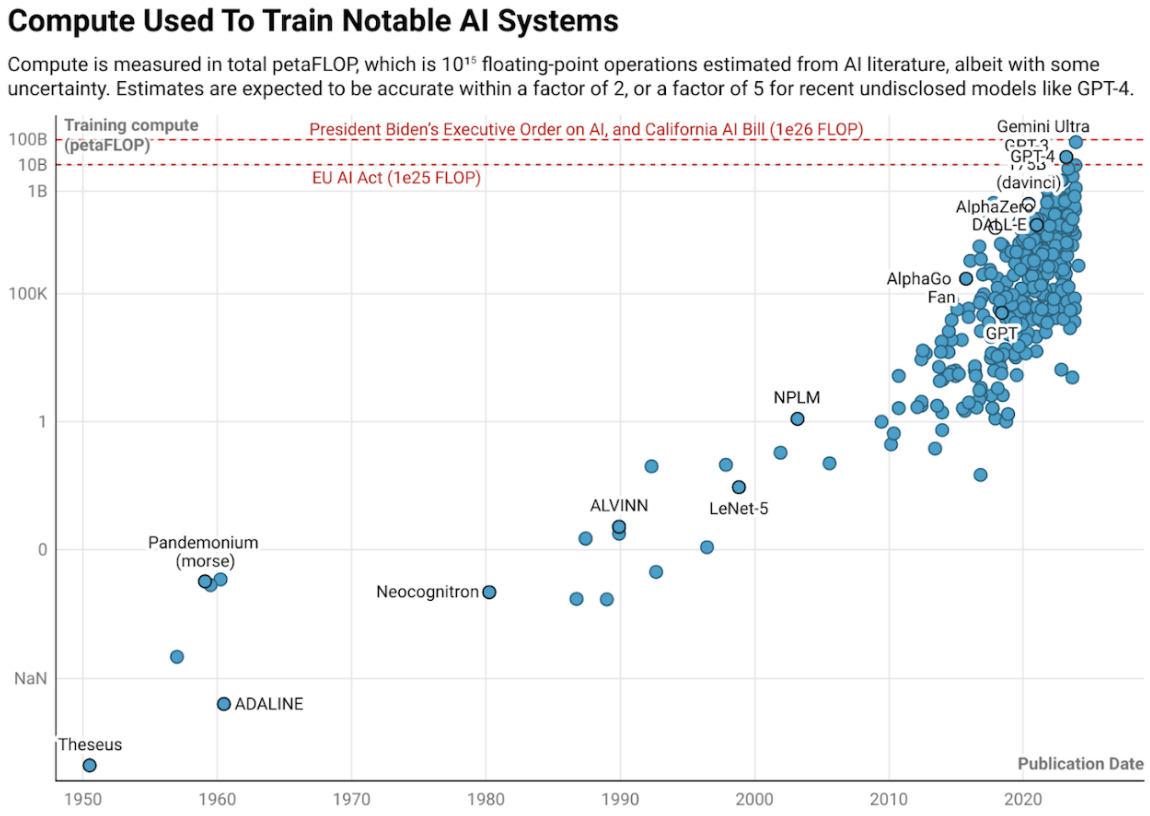

Figure 2: Compute used to train notable AI systems from 1950 to 2023[38]

However, the current rate of growth in training compute may not be sustainable. Scholars have cited the cost of training,[39] a limited supply of AI chips,[40] technical challenges with using that much hardware (such as managing the number of processors that must run in parallel to train larger models),[41] and environmental impact[42] as factors that could constrain the growth of training compute. Research in 2018 with data from OpenAI estimated that then-current trends of growth in

[38] Data for this chart was sourced from *Notable AI Models*, *supra* note 22.

[39] If current spending trajectories continued, the cost to train a frontier AI system would exceed the gross domestic product of the United States by 2036. Lennart Heim, *This Can't Go On(?)—AI Training Compute Costs*, BLOG.HEIM.XYZ (June 1, 2023), https://blog.heim.xyz/this-cant-go-on-compute-training-costs/ [https://perma.cc/7FDZ-VFLG]; *cf.* Andrew Lohn & Micah Musser, *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?*, CTR. FOR SEC. & EMERGING TECH., https://cset.georgetown.edu/publication/ai-and-compute/ [https://perma.cc/T2Z5-KRR6], at 10 & fig.2, 12 (Jan. 2022) (predicting, based on earlier numbers, that "the compute demand trendline should be expected to break within two to three years at the latest, and certainly well before 2026—if it hasn't done so already."); Ryan Carey, *Interpreting AI Compute Trends*, AI IMPACTS (July 10, 2018), https://aiimpacts.org/interpreting-ai-compute-trends/ [https://perma.cc/37R6-U8UF], at n.7 (extrapolating from their calculations, the cost of training would exceed the U.S. GDP, roughly 27 trillion dollars, by October 2025 to June 2027; to calculate, use the equation in note 7 and substitute the U.S. GDP, roughly 27 trillion, for the 200 billion used in the equation); Ben Cottier et al., *The Rising Costs of Training Frontier AI Models*, ARXIV (May 31, 2024), https://doi.org/10.48550/arXiv.2405.21015 [https://perma.cc/9GLB-BLZ4] (discussing the rising cost of training frontier AI models generally).

[40] *See* Lohn & Musser, *supra* note 39, at 1, 6, 14–15; Sevilla & Roldán, *supra* note 37.

[41] *See* Lohn & Musser, *supra* note 39, at 1, 6, 18–19 ("We estimate that the absolute upper limit of this trend's viability is at most a few years away, and that, in fact, the impending slowdown may have already begun.").

[42] *See* OECD, *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications* (OECD Digital Economy Paper No. 341, 2022), https://doi.org/10.1787/7babf571-en [https://perma.cc/F43Y-X94U]; Emma Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP*, *in* P19-1355 PROCS. 57TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 3645 (2019), http://dx.doi.org/10.18653/v1/P19-1355 [https://perma.cc/6HRF-AZFY]; Aimee van Wynsberghe, *Sustainable AI: AI for Sustainability and the Sustainability of AI*, 1 AI & ETHICS 213 (2021).

training compute could be sustained for at most 3.5 to 10 years (2022 to 2028), depending on spending levels and how the cost of compute evolves over time.[43] In 2022, that analysis was replicated with a more comprehensive dataset and suggested that this trend could be maintained for longer, for 8 to 18 years (2030 to 2040) depending on compute cost-performance improvements and specialized hardware improvements.[44]

C. What Are "Scaling Laws" and What Do They Say About AI Models?

Scaling laws describe the functional (mathematical) relationship between the amount of training compute and the performance of the AI model.[45] In this context, *performance* is a technical metric that quantifies "loss," which is the amount of error in the model's predictions. When loss is measured on a test or validation set that uses data not part of the training set, it reflects how well the model has generalized its learning from the training phase. The lower the loss, the more accurate and reliable the model is in making predictions on data it has not encountered during its training.[46] As training compute increases, alongside increases in parameters and training data, so does model performance, meaning that greater training compute reduces the errors made.[47] Increased training compute also corresponds to an increase in *capabilities*.[48] Whereas performance

---

[43] Carey, *supra* note 39; *see also* Ben Garfinkel, *Reinterpreting "AI and Compute,"* AI IMPACTS (Feb. 9, 2019), https://aiimpacts.org/reinterpreting-ai-and-compute/ [https://perma.cc/4359-QGNX] (suggesting a more pessimistic interpretation of the same data: "if we were previously underestimating the rate at which computing power was increasing, this means that we were *overestimating* how sustainable its growth is").

[44] Heim, *supra* note 39.

[45] More precisely, while this analysis focuses on compute, scaling laws describe the power-law relationship between performance and *three* technical variables: the amount of compute used to train the model, the number of parameters, and the size of the training dataset. *See infra* note 47. Training compute, parameter count, and dataset size are interconnected variables—in particular, more compute is required to train a model with more parameters or a larger dataset. *Cf. Estimating Training Compute*, *supra* note 23 (describing how the number of FLOP used to train an AI model can be calculated through information about the model's architecture and amount of training data); Amodei & Hernandez, *supra* note 29 ("we directly counted the number of FLOPs (adds and multiplies) in the described architecture per training example and multiplied by the total number of forward and backward passes during training."); Villalobos & Atkinson, *supra* note 28.

[46] For instance, OpenAI tested GPT-4's final loss, among other test sets, on an internal database that was different from training data. OpenAI, *supra* note 21, at 2–3 & fig.1 (explaining that loss "tends to be less noisy than other measures across different amounts of training compute" and reporting that a power law fit to smaller models highly accurately predicted GPT-4's final loss).

[47] Deep Ganguli et al., *Predictability and Surprise in Large Generative Models*, ARXIV 2, 4 (Oct. 3, 2022), https://doi.org/10.48550/arXiv.2202.07785 [https://perma.cc/TUB3-FAKR] ("[T]he relationship between scale and model performance is often so predictable that it can be described in a lawful relationship—a scaling law. . . . [T]he general performance of large generative models tends to exhibit smooth and predictable growth as a function of scale—larger systems tend to do increasingly better on a broad range of tasks."); Jared Kaplan et al., *Scaling Laws for Neural Language Models*, ARXIV 2–3 (Jan. 23, 2020), https://doi.org/10.48550/arXiv.2001.08361 [https://perma.cc/TFF4-F4EC] ("Performance depends strongly on scale, weakly on model shape: Model performance depends most strongly on scale, which consists of three factors: the number of model parameters N (excluding embeddings), the size of the dataset D, and the resulting amount of compute C used for training. Within reasonable limits, performance depends very weakly on other architectural hyperparameters such as depth vs. width."); Joel Hestness et al., *Deep Learning Scaling Is Predictable, Empirically*, ARXIV 1 (Dec. 1, 2017), https://doi.org/10.48550/arXiv.1712.00409 [https://perma.cc/ZA67-3SXG] ("present[ing] a large scale empirical characterization of generalization error and model size growth as training sets grow."); Lohn & Musser, *supra* note 39, at 21 ("Both compute and parameter size are critical ingredients for increasing the performance of a model under the current deep learning paradigm, and there are diminishing returns associated with scaling up one without the other.").

[48] *See* Ganguli et al., *supra* note 47, at 2–6 ("In most cases, these scaling laws predict a continued increase in certain capabilities as models get larger. . . . More precisely, by general capability scaling we mean two things. First, the training (and test) loss improves predictably with scale on a broad data distribution. Second, this improvement in loss tends to correlate on average with increased performance on a number of downstream tasks."); Konstantin Pilz, Lennart Heim & Nicholas Brown, *Increased Compute Efficiency and the Diffusion of AI Capabilities*, ARXIV 7–8 (Feb. 13, 2024), https://doi.org/10.48550/arXiv.2311.15377 [https://perma.cc/D2XX-6JYR] ("For illustration, a language model that achieves a certain *performance* on next-word prediction may gain the *capability* to solve coding problems. . . . Depending on their nature, benchmarks can capture either the performance of a model or its capabilities."); Pablo Villalobos, *Scaling Laws Literature Review*, EPOCH (Jan. 26, 2023), https://epochai.org/blog/scaling-laws-literature-review [https://perma.cc/WB5N-TXRH]; *see also* EU AI Act, *supra* note 4, at Recital 111 ("According to the state of the art at the time of entry into force of this Regulation,

refers to a technical metric, such as test loss, capabilities refer to the ability to complete concrete tasks and solve problems in the real world, including in commercial applications.[49] Capabilities can also be assessed using practical and real-world tests, such as standardized academic or professional licensing exams, or with benchmarks developed for AI models. Common benchmarks include "Beyond the Imitation Game" ("BIG-Bench"), which comprises 204 diverse tasks that cover a variety of topics and languages,[50] and the "Massive Multitask Language Understanding" benchmark ("MMLU"), a suite of multiple-choice questions covering 57 subjects.[51] To evaluate the capabilities of Google's PaLM 2 and OpenAI's GPT-4, developers relied on BIG-Bench and MMLU as well as exams designed for humans, such as the SAT and AP exams.[52]

Training compute has a relatively smooth and consistent relationship with technical metrics like training loss. Training compute also corresponds to real-world capabilities, but not in a smooth and predictable way. This is due in part to occasional surprising leaps, discussed in Section I.D, and subsequent enhancements such as fine-tuning, which can further increase capabilities using far less compute.[53] Despite being unable to provide a full and accurate picture of a model's final capabilities, training compute still provides a reasonable basis for estimating the base capabilities (and corresponding risk) of a foundation model. Figure 3 shows the relationship between an increase in training compute and dataset size, and performance on the MMLU benchmark.

---

the cumulative amount of computation used for the training of the general-purpose AI model measured in floating point operations is one of the relevant approximations for model capabilities."). *But see* Rylan Schaeffer et al., *Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive?*, ARXIV (June 6, 2024), https://doi.org/10.48550/arXiv.2406.04391 [https://perma.cc/J9DZ-MGHP].

[49] Pilz, Heim & Brown, *supra* note 48, at 7 ("Capabilities refer to a more qualitative metric, such as the problems an AI model can solve in the real world.").

[50] *See* Srivastava et al., *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*, ARXIV (June 12, 2023), https://doi.org/10.48550/arXiv.2206.04615 [https://perma.cc/F7LS-J2EE].

[51] *See* Dan Hendrycks et al., *Measuring Massive Multitask Language Understanding*, ARXIV (Jan. 12, 2021), https://doi.org/10.48550/arXiv.2009.03300 [https://perma.cc/C4UV-Q4LE].

[52] *See* Google, *PaLM 2 Technical Report*, ARXIV 9–23 (Sept. 13, 2023), https://doi.org/10.48550/arXiv.2305.10403 [https://perma.cc/99QS-3PQX]; OpenAI, *supra* note 21, at tbl.1 & n.5 (noting also that BIG-bench was excluded from the benchmark results because portions of it were inadvertently mixed into the training set).

[53] *See* Sara Hooker, *On the Limitations of Compute Thresholds as a Governance Strategy*, ARXIV 13 (July 31, 2024), https://doi.org/10.48550/arXiv.2407.05694 [https://perma.cc/7QAX-AZHL].

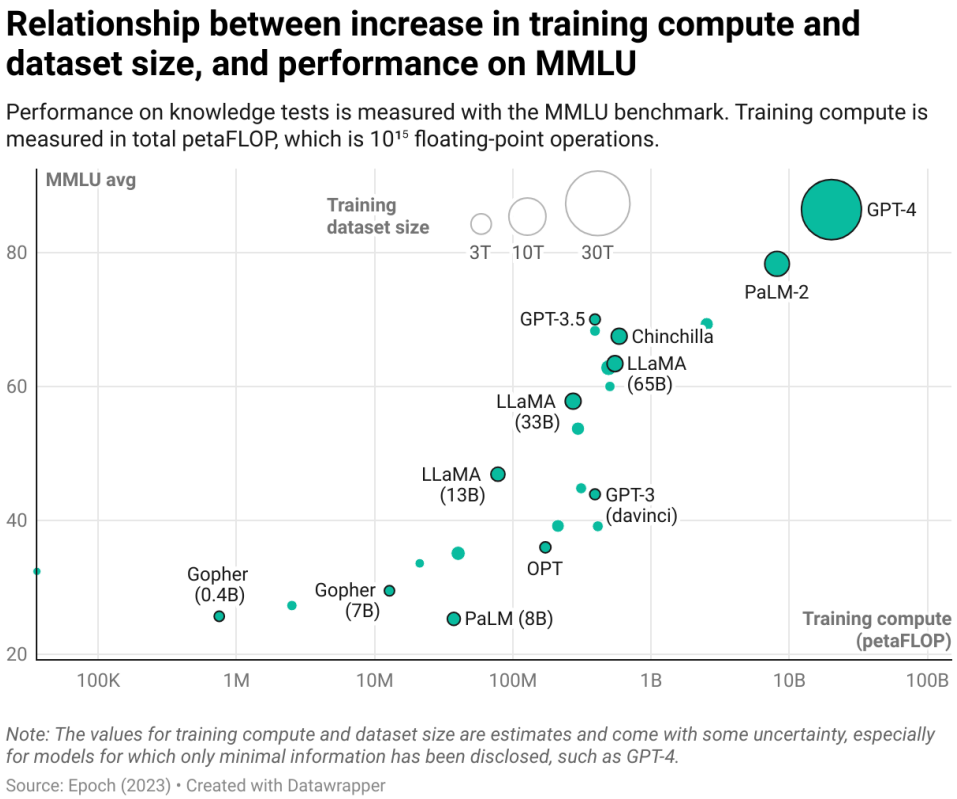

Figure 3: Relationship between increase in training compute and dataset size, and performance on MMLU[54]

In light of the correlation between training compute and performance, the "scaling hypothesis" states that scaling training compute will predictably continue to produce even more capable systems, and thus more compute is important for AI development.[55] Some have taken this hypothesis further, proposing a "Bitter Lesson:" that "the only thing that matters in the long run is the leveraging of comput[e]."[56] Since the emergence of the deep learning era, this hypothesis has been sustained by the increasing use of AI models in commercial applications, whose development and commercial success have been significantly driven by increases in training compute.[57]

---

[54] *See Artificial Intelligence: Performance on Knowledge Tests vs. Training Computation*, OUR WORLD IN DATA, https://ourworldindata.org/grapher/ai-performance-knowledge-tests-vs-training-computation [https://perma.cc/44QL-XP9Z]; David Owen, *How Predictable Is Language Model Benchmark Performance?*, EPOCH (June 9, 2023), https://epochai.org/blog/how-predictable-is-language-model-benchmark-performance [https://perma.cc/X8GE-7K6K].

[55] Gwern Branwen, *The Scaling Hypothesis* (2020), https://gwern.net/scaling-hypothesis [https://perma.cc/7CJR-EPD2] (proposing the scaling hypothesis); *see also* Anderljung et al., *supra* note 1, at 37 ("[S]caling training compute has reliably led to better performance on many of the tasks AI models are trained to solve, and many similar downstream tasks. 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."). *See generally* Thompson et al., *supra* note 36, at 19 (finding that "computing power (and implicitly the algorithm changes needed to harness it) account for half or more of all improvement" and arguing for the "importance of exponentially more computing power.").

[56] Rich Sutton, *The Bitter Lesson* (Mar. 13, 2019), http://www.incompleteideas.net/IncIdeas/BitterLesson.html [https://perma.cc/CB2B-7Q3Y]. More recently, see Matthew Barnett, *A Compute-Based Framework for Thinking About the Future of AI*, EPOCH (Aug. 10, 2023), https://epochai.org/blog/a-compute-based-framework-for-thinking-about-the-future-of-ai [https://perma.cc/SK4Q-ME8V] (arguing that compute will ultimately be most important for explaining progress in the foreseeable future).

[57] *See Compute Trends*, *supra* note 23 (describing the compute trends in the deep learning and large-scale era).

Two factors weigh against the scaling hypothesis. First, scaling laws describe more than just the performance improvements based on training compute; they describe the optimal ratio of the size of the dataset, the number of parameters, and the training compute budget.[58] Thus, a lack of abundant or high-quality *data* could be a limiting factor. Researchers estimate that, if training datasets continue to grow at current rates, language models will fully utilize human-generated public text data between 2026 and 2032,[59] while image data could be exhausted between 2030 and 2060.[60] Specific tasks may be bottlenecked earlier by the scarcity of high-quality data sources.[61] There are, however, several ways that data limitations might be delayed or avoided, such as synthetic data generation and using additional datasets that are not public or in different modalities.[62]

Second, *algorithmic innovation* permits performance gains that would otherwise require prohibitively expensive amounts of compute.[63] Research estimates that every 9 months, improved algorithms for image classification[64] and LLMs[65] contribute the equivalent of a doubling of training compute budgets. Algorithmic improvements include more efficient utilization of data[66]

---

[58] *See supra* note 47 and accompanying text on scaling laws.

[59] Pablo Villalobos et al., *Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data*, ARXIV 6–7, 9 (June 4, 2024), https://doi.org/10.48550/arXiv.2211.04325 [https://perma.cc/6NKQ-JG2U] (noting that full utilization may occur even earlier if models are "overtrained" with more data to be more compute-efficient during inference).

[60] Pablo Villalobos et al., *Will We Run Out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning*, ARXIV 1, 5–6 (Oct. 26, 2022), https://doi.org/10.48550/arXiv.2211.04325 [https://perma.cc/LNC9-3GFX].

[61] *Cf.* Barnett, *supra* note 56 (noting that, even if data does not constrain general AI progress, particular tasks may be bottlenecked).

[62] Villalobos et al., *supra* note 59, at 7–9 (discussing AI-generated data, multimodal and transfer learning using data from other domains or modalities, and non-public data); *see also* Villalobos et al., *supra* note 60, at 7–9 (discussing AI-generated data, multimodal and transfer learning, using non-public data, and other techniques); Jiaxin Huang et al., *Large Language Models Can Self-Improve*, ARXIV (Oct. 25, 2022), https://doi.org/10.48550/arXiv.2210.11610 [https://perma.cc/ZY69-BLS7] (discussing synthetic data); Ronen Eldan & Yuanzhi Li, *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?*, ARXIV (May 24, 2023), https://doi.org/10.48550/arXiv.2305.07759 [https://perma.cc/23A5-WSQB] (introducing TinyStories, a synthetic dataset of short stories usable to train and evaluate smaller language models); Armen Aghajanyan et al., *Scaling Laws for Generative Mixed-Modal Language Models*, ARXIV (Jan. 10, 2023), https://doi.org/10.48550/arXiv.2301.03728 [https://perma.cc/T6PU-4J5E] (discussing multi-modal training, which uses multiple data types). Data production could also result from certain shifts, such as large-scale adoption of self-driving cars that provide road video recordings, or from significant spending in domains where high-quality data is needed. Villalobos et al., *supra* note 60, at 2–3.

[63] *See* Danny Hernandez & Tom Brown, *Measuring the Algorithmic Efficiency of Neural Networks*, ARXIV 5–7 (May 8, 2020), https://doi.org/10.48550/arXiv.2005.04305 [https://perma.cc/5F6H-8QTB]; Lohn & Musser, *supra* note 39, at 23 (recommending a "shift towards efficiency in both algorithms and hardware rather than massive increases in compute usage"); Anderljung et al., *supra* note 1, at 34 ("[F]actors such as improvements in algorithmic efficiency would decrease the amount of computational resources required to develop models, including those with sufficiently dangerous capabilities."); Microsoft, *supra* note 1, at 21 ("The amount of compute used to train a model . . . is imperfect in several ways and unlikely to be durable into the future, especially as algorithmic improvements lead to compute efficiencies or new architectures altogether.").

[64] Ege Erdil & Tamay Besiroglu, *Revisiting Algorithmic Progress*, EPOCH (Dec. 12, 2022), https://epochai.org/blog/revisiting-algorithmic-progress [https://perma.cc/H4MH-BD96] (revisiting earlier research by Hernandez & Brown to include later data and to avoid sensitivity to the exact benchmark and threshold pair chosen, and noting uncertainty in the estimate: "our 95% CI spans 4 to 25 months"); Hernandez & Brown, *supra* note 63 (finding a 44-fold improvement in image classification algorithmic efficiency over the period of 2012 to 2019, corresponding to doubling every 16 months). *See generally* Thorsten Koch et al., *Progress in Mathematical Programming Solvers from 2001 to 2020*, 10 EURO J. ON COMPUTATIONAL OPTIMIZATION (2022) (finding that for solving Linear Programs (LP) and Mixed Integer Linear Programs (MILP), computer hardware got about 20 times faster, and the algorithms improved by a factor of about nine for LP and around 50 for MILP); Katja Grace, *Algorithmic Progress in Six Domains*, MACH. INTEL. RSCH. INST. (2013), https://intelligence.org/files/AlgorithmicProgress.pdf [https://perma.cc/9JXF-MH2T], at 49 (finding that gains from algorithmic progress have been roughly fifty to one hundred percent as large as those from hardware progress).

[65] Anson Ho et al., *Algorithmic Progress in Language Models*, ARXIV 6 (Mar. 9, 2024), https://doi.org/10.48550/arXiv.2403.05812 [https://perma.cc/L7EM-NRMF] ("[W]e find that the median doubling time for effective compute is 8.4 months, with a 95% confidence interval of 4.5 to 14.3 months.").

[66] Villalobos et al., *supra* note 59, at 9; *see also* Niklas Muennighoff et al., *Scaling Data-Constrained Language Models*, ARXIV 1–2 (Oct. 26, 2023), https://doi.org/10.48550/arXiv.2305.16264 [https://perma.cc/PB8Z-24CG] (finding that repeating data improves

and parameters, the development of improved training algorithms, or new architectures.[67] Over time, the amount of training compute needed to achieve a given capability is reduced, and it may become more difficult to predict performance and capabilities on that basis (although scaling trends of new algorithms could be studied and perhaps predicted). The governance implications of this are multifold, including that increases in training compute may become less important for AI development and that many more actors will be able to access the capabilities previously restricted to a limited number of developers.[68] Still, responsible frontier AI development may enable stakeholders to develop understanding, safety practices, and (if needed) defensive measures for the most advanced AI capabilities before these capabilities proliferate.

D. Are High-Compute Systems Dangerous?

Advances in AI could deliver immense opportunities and benefits across a wide range of sectors, from healthcare and drug discovery[69] to public services.[70] However, more capable models may come with greater risk, as improved capabilities could be used for harmful and dangerous ends. While the degree of risk posed by current AI models is a subject of debate,[71] future models

---

performance, but the value of repetition "eventually decays to zero").

[67] *See, e.g.*, Julie Keisler et al., *An Algorithmic Framework for the Optimization of Deep Neural Networks Architectures and Hyperparameters*, ARXIV (May 14, 2024), https://doi.org/10.48550/arXiv.2303.12797 [https://perma.cc/4NF8-ZUGV] (proposing an algorithmic framework to automatically generate efficient deep neural networks and optimize their associated hyperparameters); Benjamin Doerr & Carola Doerr, *Theory of Parameter Control for Discrete Black-Box Optimization: Provable Performance Gains Through Dynamic Parameter Choices*, ARXIV (Nov. 7, 2020), https://doi.org/10.48550/arXiv.1804.05650 [https://perma.cc/3TRN-GABQ] (surveying existing works of parameter control in the context of evolutionary algorithms); Xin-She Yang et al., *A Framework for Self-Tuning Optimization Algorithm*, ARXIV (Dec. 19, 2013) https://doi.org/10.48550/arXiv.1312.5667 [https://perma.cc/6VM2-EWRW] (presenting a framework for self-tuning algorithms so that, instead of tuning the parameters, an algorithm to be tuned can be used to tune the algorithm itself); Andrey Petrushov & Boris Krasnopolsky, *Automated Tuning for the Parameters of Linear Solvers*, ARXIV (Sept. 27, 2023), https://doi.org/10.48550/arXiv.2303.15451 [https://perma.cc/V5YV-VF62] (proposing an optimization algorithm for tuning the numerical method parameters); Hanxiao Liu et al., *Hierarchical Representations for Efficient Architecture Search*, ARXIV (Feb. 22, 2018), https://doi.org/10.48550/arXiv.1711.00436 [https://perma.cc/EU3A-HFQ9] (reporting a surge of interest in using algorithms to automate the manual process of architecture design).

[68] *See* Pilz, Heim & Brown, *supra* note 49, at 9–15 (describing the effects of increased compute efficiency).

[69] For instance, Google DeepMind recently announced that the AI tool Graph Networks for Materials Exploration (GNoME) enabled the discovery of 2.2 million new crystals. Amil Merchant et al., *Scaling Deep Learning for Materials Discovery*, 624 NATURE 80 (Nov. 29, 2023). For further examples, see Debleena Paul et al., *Artificial Intelligence in Drug Discovery and Development*, 26 DRUG DISCOVERY TODAY 80 (2021); Jonathan M. Stokes et al., *A Deep Learning Approach to Antibiotic Discovery*, 180(4) CELL 688 (2020); Asmaa Ibrahim et al., *Artificial Intelligence in Digital Breast Pathology: Techniques and Applications*, 49 BREAST 267 (2020).

[70] Jamie Berryhill et al., *Hello, World: Artificial Intelligence and Its Use in the Public Sector* (OECD Working Paper on Public Governance No. 36, 2019), https://doi.org/10.1787/726fd39d-en [https://perma.cc/AG2R-4W6X]; *Federal AI Use Case Inventories*, AI.GOV (Sept. 1, 2023), https://ai.gov/ai-use-cases/ [https://perma.cc/5LVA-FEFV]; Rachel Wright, *Artificial Intelligence in the States*, COUNCIL STATE GOV'TS (Dec. 5, 2023), https://www.csg.org/2023/12/05/artificial-intelligence-in-the-public-sector-how-are-states-harnessing-the-power-of-ai/ [https://perma.cc/7QL2-VEMK].

[71] Nikhil Mulani & Jess Whittlestone, *Proposing a Foundation Model Information-Sharing Regime for the UK*, CTR. FOR THE GOVERNANCE OF AI (June 16, 2023), https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk [https://perma.cc/7EP6-R4HC] ("The degree of risk posed by current foundation models is contentious."). Some argue that current AI systems already pose catastrophic risks in various domains. *See* Benjamin S. Bucknall, & Shiri Dori-Hacohen, *Current and Near-Term AI as a Potential Existential Risk Factor*, *in* AAI/ACM CONF. ON AI, ETHICS, & SOC'Y 119–129 (2022), https://doi.org/10.1145/3514094.3534146 [https://perma.cc/Y2C4-TVA9] (proposing the hypothesis that certain already-documented effects of AI can act as existential risk factors). Others contend that they do not pose existential risks but might in the future. *See* U.K., Department for Science, Innovation and Technology, *Future Risks of Frontier AI* (Oct. 2023), https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf [https://perma.cc/C2GB-W2AQ], at 2, 25 (concluding that "[g]iven the significant uncertainty, there is insufficient evidence to rule out that future Frontier AI, if misaligned, misused or inadequately controlled, could pose an existential threat," discussing the debate on AI and existential risk, and outlining several pathways of risk); Altman et al., *supra* note 1 (arguing that the level of risks posed by today's models "feel commensurate with other Internet technologies and society's likely approaches seem appropriate," while future systems may "have power beyond any technology yet created").

may pose catastrophic and existential risks as capabilities improve.[72] Some of these risks are expected to be closely connected to the unexpected emergence of dangerous capabilities and the dual-use nature of AI models.

As discussed in Section I.C, increases in compute, data, and the number of parameters lead to predictable improvements in model performance (test loss) and general but somewhat less predictable improvements in capabilities (real-world benchmarks and tasks). However, scaling up these inputs to a model can also result in *qualitative* changes in capabilities in a phenomenon known as "emergence."[73] That is, a larger model might unexpectedly display *emergent capabilities* not present in smaller models, suddenly able to perform a task that smaller models could not.[74] During the development of GPT-3, early models had close-to-zero performance on a benchmark for addition, subtraction, and multiplication. Arithmetic capabilities appeared to emerge suddenly in later models, with performance jumping substantially above random at $2 \cdot 10^{22}$ FLOP and continuing to improve with scale.[75] Similar jumps were observed at different thresholds, and for different models, on a variety of tasks.[76]

Some have contested the concept of emergent capabilities, arguing that what appear to be emergent capabilities in large language models are explained by the use of discontinuous measures, rather than by sharp and unpredictable improvements or developments in model

---

[72] Yoshua Bengio et al., *Managing Extreme AI Risks Amid Rapid Progress*, 384 SCI. 842, 843 (May 20, 2024) ("[A]longside advanced AI capabilities come large-scale risks."); Samuel Bowman, *Eight Things to Know About Large Language Models*, ARXIV 8 (Apr. 2, 2023), https://doi.org/10.48550/arXiv.2304.00612 [https://perma.cc/7UF3-BQ63] ("[I]t is reasonable to expect a substantial increase and a substantial qualitative change in the range of misuse risks and model misbehaviors that emerge from the development and deployment of LLMs."); Ganguli et al., *supra* note 47, at 2 ("[R]isks . . . may become more severe as the models increase in capability."); Dario Amodei et al., *Concrete Problems in AI Safety*, ARXIV 2 (July 25, 2016), https://doi.org/10.48550/arXiv.1606.06565 [https://perma.cc/GS82-MT4Z], at 2 ("As AI capabilities advance and as AI systems take on increasingly important societal functions, we expect the fundamental challenges discussed in this paper to become increasingly important."); Matteucci et al., *supra* note 1, at 6 ("[T]oday's most advanced AI systems are characterized by the need for very large training compute . . . and high load (parameter count), which are directly linked (via scaling laws) to higher capabilities, and therefore to a higher potential for harm.").

[73] Philip W. Anderson, *More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science*, 177(4047) SCI. 393, 393–96 (1972) (popularizing the concept); *see also* Rylan Schaeffer et al., *Are Emergent Abilities of Large Language Models a Mirage?*, ARXIV 1 (May 22, 2023), https://doi.org/10.48550/arXiv.2304.15004 [https://perma.cc/L583-GHW8] ("The idea of emergence was popularized by Nobel Prize-winning physicist P.W. Anderson's "More Is Different," which argues that as the complexity of a system increases, new properties may materialize that cannot be predicted even from a precise quantitative understanding of the system's microscopic details.").

[74] Jason Wei et al., *Emergent Abilities of Large Language Models,* ARXIV 2 (Oct. 26, 2022), https://doi.org/10.48550/arXiv.2206.07682 [https://perma.cc/2CZF-JK2P]; Ganguli et al., *supra* note 47, at 4 ("Though performance is predictable at a general level, performance on a specific task can sometimes emerge quite unpredictably and abruptly at scale."); Bowman, *supra* note 72, at 2–4 ("Often, a model can fail at some task consistently, but a new model trained in the same way at five or ten times the scale will do well at that task."); Anderljung et al., *supra* note 1, at 10–11 ("[S]pecific capabilities can significantly improve quite suddenly."); Yonadav Shavit, *What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*, ARXIV 3, 4, 18 (May 30, 2023), https://doi.org/10.48550/arXiv.2303.11341 [https://perma.cc/KQ6N-HTDP] (citing Wei et al. and Ganguli et al.); David Owen, *How Predictable Is Language Model Benchmark Performance?*, ARXIV 7 (Jan. 9, 2024), https://doi.org/10.48550/arXiv.2401.04757 [https://perma.cc/NGK9-PE8B] (citing Aarohi Srivastava et al., *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*, ARXIV (June 12, 2023), https://doi.org/10.48550/arXiv.2206.04615 [https://perma.cc/A8ZG-HT4Y] (noting that, while "overall model capabilities are predictable with scale," "[i]ndividual tasks are highly variable in their scaling, and the sharp emergence of capabilities can make it difficult to predict performance."). As summarized during the U.K.'s AI Safety Summit in November 2023, "it is very likely we will continue to be surprised by what future AI systems can do, in ways that are not necessarily predicted or intended by their creators." U.K., Department for Science, Innovation and Technology, *AI Safety Summit 2023: Roundtable Chairs' Summaries, 1 November* (Nov. 1, 2023), https://www.gov.uk/government/publications/ai-safety-summit-1-november-roundtable-chairs-summaries/ai-safety-summit-2023-roundtable-chairs-summaries-1-november--2 [https://perma.cc/6MC2-ULQV].

[75] Wei, *supra* note 74, at 3–4 & fig.2A.

[76] *Id.*

capabilities with scale.[77] However, discontinuous measures are often meaningful, as when the correct answer or action matters more than how close the model gets to it. As Anderljung and others explain: "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."[78] Given the difficulties inherent in choosing an appropriate continuous measure and determining how it corresponds to the relevant discontinuous measure,[79] it is likely that capabilities will continue to seemingly emerge.

Together with emerging capabilities come emerging *risks*. Like many other innovations, AI systems are dual-use by nature, with the potential to be used for both beneficial and harmful ends.[80] Executive Order 14,110 recognized that some models may "pose a serious risk to security, national economic security, national public health or safety" by "substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear weapons; enabling powerful offensive cyber operations . . . ; [or] permitting the evasion of human control or oversight through means of deception or obfuscation."[81]

Predictions and evaluations will likely adequately identify many capabilities before deployment, allowing developers to take appropriate precautions. However, systems trained at a greater scale may possess novel capabilities, or improved capabilities that surpass a critical threshold for risk, yet go undetected by evaluations.[82] Some of these capabilities may appear to emerge only after post-training enhancements, such as fine-tuning or more effective prompting methods. A system may be capable of conducting offensive cyber operations, manipulating people in conversation, or providing actionable instructions on conducting acts of terrorism,[83] and still be deployed without the developers fully comprehending unexpected and potentially harmful behaviors. Research has already detected unexpected behavior in current models. For instance, during the recent U.K. AI Safety Summit on November 1, 2023, Apollo Research showed that

---

[77] *See generally* Schaeffer et al., *supra* note 73; Thomas Woodside, *Emergent Abilities in Large Language Models: An Explainer*, CTR FOR SEC. & EMERGING TECH. (Apr. 16, 2024), https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/ [https://perma.cc/YW7D-E7CN] (noting that Schaeffer et al. show that capabilities that appear to emerge suddenly are often more predictable if they can be decomposed into metrics that improve continuously, and that its results were not unforeseen by Wei et al.).

[78] Anderljung et al., *supra* note 1, at 38, app. B; *see also* Boaz Barak, *Emergent Abilities and Grokking: Fundamental, Mirage, or Both?*, WINDOWS ON THEORY (Dec. 23, 2023), https://windowsontheory.org/2023/12/22/emergent-abilities-and-grokking-fundamental-mirage-or-both/ [https://perma.cc/QJ2Y-J2N7].

[79] Anderljung et al., *supra* note 1, at 38, app. B.

[80] *See, e.g.*, Fabio Urbina et al., *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 NATURE MACH. INTEL. 189, 189–91 (2022); Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, ARXIV (Dec. 1, 2024), https://doi.org/10.48550/arXiv.1802.07228 [https://perma.cc/HLY6-4WKK]; *cf.* Lucie-Aimée Kaffee et al., *Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing*, ARXIV 1–4 (Oct. 30, 2023), https://doi.org/10.48550/arXiv.2304.08315 [https://perma.cc/8G36-BL88].

[81] Exec. Order on AI, *supra* note 2, § 3(k) (defining "dual-use foundation model").

[82] Ganguli et al., *supra* note 47, at 4, 6–8 ("Large generative models are open-ended and can take in a varying range of inputs concerning arbitrary domains. As a result, certain capabilities (or even entire areas of competency) may be unknown until an input happens to be provided that solicits such knowledge. Even after a model is trained, creators and users may not be aware of most of its (possibly harmful) capabilities.").

[83] Toby Shevlane et al., *Model Evaluation for Extreme Risks*, ARXIV 1 & tbl.1 (Sept. 22, 2023), https://doi.org/10.48550/arXiv.2305.15324 [https://perma.cc/9BDQ-MAER]; Dan Hendrycks et al., *Unsolved Problems in ML Safety*, ARXIV 7 (June 16, 2022), https://doi.org/10.48550/arXiv.2109.13916 [https://perma.cc/L5CS-KK2A] (observing that future models may make the synthesis of harmful or illegal content seamless, such as videos of child exploitation, suggestions for evading the law, or instructions for building bombs).

GPT-4 can take illegal actions like insider trading and then lie about its actions without being instructed to do so.[84] Since the capabilities of future foundation models may be challenging to predict and evaluate, "emergence" has been described as "both the source of scientific excitement and anxiety about unanticipated consequences."[85]

Not all risks come from large models. Smaller models trained on data from certain domains, such as biology or chemistry, may pose significant risks if repurposed or misused.[86] When MegaSyn, a generative molecule design tool used for drug discovery, was repurposed to find the most toxic molecules instead of the least toxic, it found tens of thousands of candidates in under six hours, including known biochemical agents and novel compounds predicted to be as or more deadly.[87] The amount of compute used to train DeepMind's AlphaFold, which predicts three-dimensional protein structures from the protein sequence, is minimal compared to frontier language models.[88] While scaling laws can be observed in a variety of domains, the amount of compute required to train models in some domains may be so low that a compute threshold is not a practical restriction on capabilities.

Broad consensus is forming around the need to test, monitor, and restrict systems of concern.[89] The role of compute thresholds, and whether they are used at all, depends on the nature of the risk and the purpose of the policy: does it target risks from emergent capabilities of frontier

---

[84] *Our Research on Strategic Deception Presented at the UK's AI Safety Summit*, APOLLO RSCH. (Nov. 6, 2023), https://www.apolloresearch.ai/research/summit-demo [https://perma.cc/FK67-ZAJF]; *see also* Jérémy Scheurer et al., *Technical Report: Large Language Models Can Strategically Deceive Their Users When Put Under Pressure*, ARXIV 1 (July 15, 2024), https://doi.org/10.48550/arXiv.2311.07590 [https://perma.cc/X6JD-2V94].

[85] Bommasani et al., *supra* note 20, at 3.

[86] *See generally* TODD KUIKEN, CONG. RSCH. SERV. R47849, ARTIFICIAL INTELLIGENCE IN THE BIOLOGICAL SCIENCES: USES, SAFETY, SECURITY, AND OVERSIGHT 2 (2023), at 2; Cassidy Nelson & Sophie Rose, *Understanding AI-Facilitated Biological Weapon Development*, CTR. FOR LONG-TERM RESILIENCE (Oct. 2023), https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio [https://perma.cc/4QQB-9KTQ]; Sarah R. Carter et al., *The Convergence of Artificial Intelligence and the Life Sciences*, NUCLEAR THREAT INITIATIVE (Oct. 2023), https://www.nti.org/wp-content/uploads/2023/10/NTIBIO_AI_Executive-Summary_FINAL.pdf [https://perma.cc/64JT-YFYZ], at 23–30.

[87] Urbina et al., *supra* note 77, at 189–191; *cf.* Jonas B. Sandbrink, *Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools*, ARXIV (Dec. 23, 2023), https://doi.org/10.48550/arXiv.2306.13952 [https://perma.cc/A4WU-PHJ6] (discussing similar biological risks from AI).

[88] Lohn & Musser, *supra* note 39, at 21 (noting that "not all progress requires record-breaking levels of compute" and, for instance, "AlphaFold is revolutionizing aspects of computational biochemistry and only required a few weeks of training on 16 TPUs" and "current top performing image classifier only needed two days to train on 512 TPUs"); *see also* Sterlin Sawaya et al., *The Potential For Dual-Use of Protein-Folding Prediction*, F3 MAG. 152 (2021), https://unicri.it/sites/default/files/2021-12/21_dual_use.pdf [https://perma.cc/9LVZ-QTNH] (raising concerns about potentially malicious uses of protein-folding algorithms).

[89] A research team at the Centre for the Governance of AI surveyed leading experts from labs, academia, and civil society. The vast majority (98%) agreed that, among others, pre-deployment risk assessments, dangerous capabilities evaluations and safety restrictions on model usage are necessary. Jonas Schuett et al., *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion*, ARXIV 2, 8 (May 11, 2023), https://doi.org/10.48550/arXiv.2305.07153 [https://perma.cc/9CHH-NDAA]. Microsoft has suggested focusing on "highly capable systems, increasingly autonomous systems, and systems that cross the digital physical divide," such as those that: (a) take decisions or actions affecting large-scale networked systems; (b) process or direct physical inputs and outputs; (c) operate autonomously or semi-autonomously; (d) pose a significant potential risk of large-scale harm, including physical, economic, or environmental harm. Microsoft, *supra* note 1, at 14. For further examples of dangerous capabilities, see Jonas Schuett, *Defining the Scope of AI Regulations*, 15(1) L., INNOVATION & TECH. 60, 60–82, 75 (Mar. 3, 2023) (identifying as potential sources of risk the capability to: (a) physically interact with their environment; (b) make automated decisions; (c) make decisions which have a legal or similarly significant effect) and Matthijs Maas, *Concepts in Advanced AI Governance: A Literature Review of Key Terms and Definitions, AI Foundations Report 3*, INST. FOR L. & AI (Oct. 2023), https://law-ai.org/wp-content/uploads/2023/10/website-PDF-version-Concepts-in-advanced-AI-governance_-A-literature-review-of-key-terms-and-definitions.pdf [https://perma.cc/X78G-UEVA], at 44–49 (presenting a taxonomy of critical capabilities that may result in significant risk).

models,[90] risks from models with more narrow but dangerous capabilities,[91] or other risks from AI?

E. Does Compute Usage Outside of Training Influence Performance and Risk?

In light of the relationship between training compute and performance expressed by scaling laws, training compute is a common proxy for how capable and powerful AI models are and the risks that they pose.[92] However, compute used outside of training can also influence performance, capabilities, and corresponding risk.

As discussed in Section I.A, training compute typically does not refer to *all* compute used during development, but is instead limited to compute used during the final pre-training run.[93] This definition excludes subsequent (post-training) enhancements, such as fine-tuning and prompting methods, which can significantly improve capabilities (see *supra* Figure 1) using far less compute; many current methods can improve capabilities the equivalent of a 5x increase in training compute, while some can improve them by more than 20x.[94]

The focus on training compute also misses the significance of compute used for inference, in which the trained model generates output in response to a prompt or new input data.[95] Inference is the biggest compute cost for models deployed at scale, due to the frequency and volume of requests they handle.[96] While developing an AI model is far more computationally intensive than a single inference request, it is a one-time task. In contrast, once a model is deployed, it may receive numerous inference requests that, in aggregate, exceed the compute expenditures of training. Some have even argued that inference compute could be a bottleneck in scaling AI, if inference compute costs scaling with training compute grow too large.[97]

Greater availability of inference compute could enhance malicious uses of AI by allowing the model to process data more rapidly and enabling the operation of multiple instances in parallel. For example, AI could more effectively be used to carry out cyber attacks, such as a distributed

---

[90] For examples of laws that address large-scale AI risk, see Exec. Order on AI, *supra* note 2 (U.S.); EU AI Act, *supra* note 4 (European Union); Measures for the Management of Generative Artificial Intelligence Services (China); National Information Security Standardization Technical Committee (TC260), Safety Requirement Guidelines (China); *see also* Bill No. 2,338/2023, Dispõe sobre o uso da Inteligência Artificial (introduced May 3, 2023) (Brazil).

[91] *See, e.g.*, Exec. Order on AI, *supra* note 2, § 4.2(b) (establishing a lower compute threshold for models "using primarily biological sequence data"); *cf.* Artificial Intelligence and Biosecurity Risk Assessment Act of 2023, S. 2399, 118th Cong. (2023) (charging the Department of Health and Services with evaluating whether advanced AI could be used to develop various biosecurity threats); Strategy for Public Health Preparedness and Response to Artificial Intelligence Threats Act of 2023, S. 2346, 118th Cong. (2023) (proposing broader responsibilities for HHS including development of a plan focused on risks that AI might pose to national health security).

[92] *See, e.g.*, Exec. Order on AI, *supra* note 2, § 4.2(b); EU AI Act, *supra* note 4, art. 51 (establishing a presumption that AI models above 1e25 FLOP have "high impact capabilities"); *infra* notes 114–124, 158–180 and accompanying text (discussing the use of compute thresholds in existing and proposed law).

[93] *See supra* notes 23–25 and accompanying text.

[94] Davidson et al., *supra* note 21, at 1, tbl.1, 4–5 (summarizing post-training enhancements and their corresponding compute-equivalent gain).

[95] *See* U.K. Competition & Markets Authority, *supra* note 16, at 14, n.22 ("Inference refers to each time the model is called upon to make a make a prediction based on new data.").

[96] *See supra* notes 27–29 and accompanying text.

[97] Dylan Patel & Gerald Wong, *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*, SEMIANALYSIS (July 10, 2023), https://www.semianalysis.com/p/gpt-4-architecture-infrastructure [https://perma.cc/DM9A-NVCA].

denial-of-service ("DDoS") attack,[98] to manipulate financial markets,[99] or to increase the speed, scale, and personalization of disinformation campaigns.[100]

Compute used outside of development may also impact model performance. Specifically, some techniques can increase the performance of a model at the cost of more compute used during inference.[101] Developers could therefore choose to improve a model beyond its current capabilities or to shift some compute expenditures from training to inference, in order to obtain equally-capable systems with less training compute. Users could also prompt a model to use similar techniques during inference, for example by (1) using "few-shot" prompting, in which initial prompts provide the model with examples of the desired output for a type of input,[102] (2) using chain-of-thought prompting, which uses few-shot prompting to provide examples of reasoning,[103] or (3) simply providing the same prompt multiple times and selecting the best result. Some user-side techniques to improve performance might increase the compute used during a single inference, while others would leave it unchanged (while still increasing the total compute used, due to multiple inferences being performed).[104] Meanwhile, other techniques—such as pruning,[105] weight sharing,[106] quantization,[107] and distillation[108]—can reduce compute used during inference while maintaining or even improving performance, and they can further reduce inference compute at the cost of lower

---

[98] *Cf.* Jugal Shroff et al., *Enhanced Security Against Volumetric DDoS Attacks Using Adversarial Machine Learning*, 2022 WIRELESS COMMC'NS & MOBILE COMPUTING 5757164 (Mar. 11, 2022), https://doi.org/10.1155/2022/5757164 [https://perma.cc/2NQ3-NLBD].

[99] *See* Alessio Azzutti et al., *Machine Learning, Market Manipulation, and Collusion on Capital Markets: Why the "Black Box" Matters*, 43 U. PA. J. INT'L L. 79, 94–103 (2021).

[100] *See generally* Katerina Sedova et al., *AI and the Future of Disinformation Campaigns Part 2: A Threat Model*, CTR. FOR SEC. & EMERGING TECHNOLOGY (Dec. 2021), https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/ [https://perma.cc/E8UX-6PXK].

[101] Villalobos & Atkinson, *supra* note 28 (reviewing four techniques: varying the scaling policy, pruning, Monte Carlo Tree Search, and repeated sampling of the model and filtering for the best result); Davidson et al., *supra* note 21.

[102] *See* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, ARXIV 2, 4, 22 (July 22, 2020), https://doi.org/10.48550/arXiv.2005.14165 [https://perma.cc/E4SJ-7ZTU] (describing meta-learning, "which in the context of language models means the model develops a broad set of skills and pattern recognition abilities at training time, and then uses those abilities at inference time to rapidly adapt to or recognize the desired task," and further distinguishing between zero-, one-, and few-shot "depending on how many demonstrations are provided at inference time" and further noting that "one- and few-shot performance is often much higher than true zero-shot performance").

[103] *See generally* Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, ARXIV (Jan. 10, 2023), https://doi.org/10.48550/arXiv.2201.11903 [https://perma.cc/H4LM-YDUN].

[104] *Cf. id.* at 6 (finding that, for chain-of-thought prompting to improve performance, the model must actually use additional compute to express intermediate steps via natural language and cannot provide an abbreviated output).

[105] Pruning is the practice of removing parameters (such as weights) that are redundant or not sufficiently informative. *See id.*; Song Han et al., *Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding*, *in* 4TH INT'L CONF. ON LEARNING REPRESENTATIONS (2016) (reporting no loss of accuracy with models pruned by "removing the redundant connections, keeping only the most informative connections"); Yihui He et al., *Channel Pruning for Accelerating Very Deep Neural Networks*, *in* IEEE INT'L CONF. ON COMPUT. VISION 1398 (2017), https://doi.org/10.1109/ICCV.2017.155 [https://perma.cc/FX25-VKJ9].

[106] Weight sharing in neural networks, and particularly in convolutional neural networks (CNNs), is the practice of using the same weights across different connections. Jordan Ott, *Learning in the Machine: To Share or Not To Share?*, 126 NEURAL NETWORKS 235, 235–249 (2020); Xin Chen et al., *Fitting the Search Space of Weight-sharing NAS with Graph Convolutional Networks*, *in* THIRTY-FIFTH AAAI CONF. ON A.I. 7065 (2021).

[107] Quantization is the practice of reducing the precision of numbers used to represent model parameters. *See, e.g.*, Benoit Jacob et al., *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*, *in* PROC. IEEE/CVF CONF. ON COMPUT. VISION & PATTERN RECOGNITION 2704 (2018) (describing the approach of " quantiz[ing] the weights and / or activations of a CNN from 32 bit floating point into lower bit-depth representations"); Darryl Lin et al., *Fixed Point Quantization of Deep Convolutional Networks*, *in* PROC. 33RD INT'L CONF. MACH. LEARNING 2849 (2016); Zhongnan Qu et al., *Adaptive Loss-Aware Quantization for Multi-bit Networks*, ARXIV (July 4, 2020), https://doi.org/10.48550/arXiv.1912.08883 [https://perma.cc/F4SU-3PNR].

[108] Distillation is the practice of training a smaller, simpler model to replicate the behavior of a larger, more complex model. *See* Geoffrey Hinton et al., *Distilling the Knowledge in a Neural Network*, NIPS DEEP LEARNING & REPRESENTATION LEARNING WORKSHOP (2015), https://doi.org/10.48550/arXiv.1503.02531 [https://perma.cc/3SFV-KZAH].

performance.

Beyond model characteristics such as parameter count, other factors can also affect the amount of compute used during inference in ways that may or may not improve performance, such as input size (compare a short prompt to a long document or high-resolution image) and batch size (compare one input provided at a time to many inputs in a single prompt).[109] Thus, for a more accurate indication of model capabilities, compute used to run a single inference[110] for a given set of prompts could be considered alongside other factors, such as training compute. However, doing so may be impractical, as data about inference compute (or architecture useful for estimating it) is rarely published by developers,[111] different techniques could make inference more compute-efficient, and less information is available regarding the relationship between inference compute and capabilities.

While companies might be hesitant to increase inference compute at scale due to cost, doing so may still be worthwhile in certain circumstances, such as for more narrowly deployed models or those willing to pay more for improved capabilities. For example, OpenAI offers dedicated instances for users who want more control over system performance, with a reserved allocation of compute infrastructure and the ability to enable features such as longer context limits.[112]

Over time, compute usage during the AI development and deployment process may change. It was previously common practice to train models with supervised learning, which uses annotated datasets. In recent years, there has been a rise in self-supervised, semi-supervised, and unsupervised learning, which use data with limited or no annotation but require more compute.[113]

---

[109] *Cf.* Andrew G. Howard et al., *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, ARXIV (Apr. 17, 2017), https://doi.org/10.48550/arXiv.1704.04861 [https://perma.cc/RRH6-FQUU] (in the context of mobile and embedded vision applications, finding that computational power depends on the number of input channels, *M*, which represent the data points, such as, for an image, the number of pixels multiplied by one if in greyscale or three if in color with separate red, green, and blue values); Tim Yarally et al., *Batching for Green AI – An Exploratory Study on Inference*, ARXIV (July 21, 2023), https://doi.org/10.48550/arXiv.2307.11434 [https://perma.cc/JG4L-UC73] (examining the effect of input batching on energy consumption and response times of neural networks for computer vision); Yuriy Kochura et al., *Batch Size Influence on Performance of Graphic and Tensor Processing Units During Training and Inference Phases*, ARXIV (Dec. 31, 2018), https://doi.org/10.48550/arXiv.1812.11731 [https://perma.cc/L8R4-GES3] (investigating scaling of training and inference performance with an increase of batch size and dataset size); Zhoujun Cheng et al., *Batch Prompting: Efficient Inference with Large Language Model APIs*, ARXIV (Oct. 24, 2023), https://doi.org/10.48550/arXiv.2301.08721 [https://perma.cc/K66E-7278] (proposing a prompting approach that enables LLMs to run inference in batches, instead of one sample at a time, as a solution to reduce inference costs).

[110] *See* Villalobos & Atkinson, *supra* note 28 ("[W]e must distinguish between the cost of running a single inference, which is a technical characteristic of the model, and the aggregate cost of all the inferences over the lifetime of a model, which additionally depends on the number of inferences run.").

[111] *See* Pilz, Heim & Brown, *supra* note 48, at n.15 ("Over the last year, we observe that publication norms have entered a new phase. Frontier AI developers are reluctant to share even basic details of their models, such as architecture and compute used. . . .").

[112] Greg Brockman et al., *Introducing ChatGPT and Whisper APIs*, OPENAI (Mar. 1, 2023), https://openai.com/blog/introducing-chatgpt-and-whisper-apis [https://perma.cc/HJ98-36HB].

[113] *See* Bommasani et al., *supra* note 20, at 4–5 (describing self-supervised learning); Alec Radford et al., *Improving Language Understanding by Generative Pre-Training*, OPENAI (2018), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [https://perma.cc/7USE-8UTB], at 2–3 (describing semi-supervised and unsupervised learning).

III. THE ROLE OF COMPUTE THRESHOLDS FOR AI GOVERNANCE

A. How Can Compute Thresholds Be Used in AI Policy?

Compute can be used as a proxy for the capabilities of AI systems, and compute thresholds can be used to define the limited subset of high-compute models subject to oversight or other requirements.[114] Their use depends on the context and purpose of the policy. Compute thresholds serve as intuitive starting points to identify potential models of concern,[115] perhaps alongside other factors.[116] They operate as a trigger for greater scrutiny or specific requirements. Once a certain level of training compute is reached, a model is presumed to have a higher risk of displaying dangerous capabilities (and especially unknown dangerous capabilities) and, hence, is subject to stricter oversight and other requirements.

Compute thresholds have already entered AI policy. The EU AI Act requires model providers to assess and mitigate systemic risks, report serious incidents, conduct state-of-the-art tests and model evaluations, ensure cybersecurity, and report serious incidents if a compute threshold is crossed.[117] Under the EU AI Act, a general-purpose model that meets the initial threshold is presumed to have high-impact capabilities and associated systemic risk.[118]

In the United States, Executive Order 14,110 directed agencies to propose rules based on compute thresholds. Although it was revoked by President Trump's Executive Order 14,148,[119] many actions have already been taken and rules have been proposed for implementing Executive Order 14,110. For instance, the Department of Commerce's Bureau of Industry and Security issued a proposed rule on September 11, 2024[120] to implement the requirement that AI developers and cloud service providers report on models above certain thresholds, including information about (1) "any ongoing or planned activities related to training, developing, or producing dual-use foundation models," (2) the results of red-teaming, and (3) the measures the company has taken to meet safety objectives.[121] The executive order also imposed know-your-customer ("KYC") monitoring and reporting obligations on U.S. cloud infrastructure providers and their foreign resellers, again with a preliminary compute threshold.[122] On January 29, 2024, the Bureau of Industry and Security issued a proposed rule implementing those requirements.[123] The proposed

---

[114] *See generally* Anderljung et al., *supra* note 1, at 9, 35–37 (discussing the advantages and limitations of compute as one of several options); Matteucci et al., *supra* note 1, at 5–6 (expecting intrinsic danger to come only from systems that have very high capabilities and therefore suggest to "only subject a small subset of all AI systems to such evaluations").

[115] *See, e.g.*, Matteucci et al., *supra* note 1, at 6; Leonie Koessler et al., *Risk Thresholds for Frontier AI*, CTR. FOR THE GOVERNANCE OF AI (June 16, 2023), https://www.governance.ai/research-paper/risk-thresholds-for-frontier-ai [https://perma.cc/448C-QNRE], at 3.

[116] Further definitional elements are discussed in Charlie Bullock et al., *Legal Considerations for Defining "Frontier Model"* (Inst. for L. & AI, Working Paper No. 2-2024), (Inst. for L. & AI, Working Paper No. 3-2024), https://law-ai.org/wp-content/uploads/2024/09/Legal-Considerations-for-Defining-Frontier-Model.pdf [https://perma.cc/MUR4-CDME].

[117] EU AI Act, *supra* note 4, at art. 55.

[118] EU AI Act, *supra* note 4, at Recital 111 and art. 51.

[119] Exec. Order No. 14,148, § 2(ggg), 90 Fed. Reg. 8237 (Jan. 20, 2025).

[120] Bureau Indus. & Sec., Establishment of Reporting Requirements for the Development of Advanced Artificial Intelligence Models and Computing Clusters, 89 Fed. Reg. 73612 (proposed Sept. 11, 2024) (to be codified at 15 C.F.R. pt. 702).

[121] Exec. Order on AI, *supra* note 2, § 4.2(a)–(b).

[122] *Id.* § 4.2(c).

[123] Bureau Indus. & Sec., Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities, 89 Fed. Reg. 5698 (proposed Jan. 29, 2024) (to be codified at 15 C.F.R. pt. 7); Press Release, U.S. Dep't of Com.,

rule noted that training compute thresholds may determine the scope of the rule; the program is limited to foreign transactions to "train a large AI model with potential capabilities that could be used in malicious cyber-enabled activity," and technical criteria "may include the compute used to pre-train the model exceeding a specified quantity." [124] The fate of these rules is uncertain, as all rules and actions taken pursuant to Executive Order 14,110 will be reviewed to ensure that they are consistent with the AI policy set forth in Executive Order 14,179, Removing Barriers to American Leadership in Artificial Intelligence.[125] Any rules of actions identified as inconsistent are directed to be suspended, revised, or rescinded.[126]

Numerous policy proposals have likewise called for compute thresholds. Scholars and developers alike have expressed support for a licensing or registration regime,[127] and a compute threshold could be one of several ways to trigger the requirement.[128] Compute thresholds have also been proposed for determining the level of KYC requirements for compute providers (including cloud providers).[129] The Framework to Mitigate AI-Enabled Extreme Risks, proposed by U.S. Senators Romney, Reed, Moran, and King, would include a compute threshold for requiring notice of development, model evaluation, and pre-deployment licensing.[130]

Other AI regulations and policy proposals do not explicitly call for the introduction of compute thresholds but could still benefit from them. A compute threshold could clarify when specific obligations are triggered in laws and guidance that refer more broadly to "advanced systems" or "systems with dangerous capabilities," as in the voluntary guidance for "organizations developing the most advanced AI systems" in the Hiroshima Process International Code of

---

Commerce Proposes Rule to Advance U.S. National Security Interests and Implement Biden-Harris Administration's AI Executive Order and National Cybersecurity Strategy (Jan. 29, 2024), https://www.bis.doc.gov/index.php/documents/about-bis/newsroom/press-releases/3443-2024-01-29-bis-press-release-infrastructure-as-as-service-know-your-customer-nprm-final/file [https://perma.cc/789D-6KLE].

[124] Bureau Indus. & Sec., *supra* note 123.

[125] Exec. Order No. 14,179, § 5(a), 90 Fed. Reg. 8741 (Jan. 23, 2025).

[126] *Id.* § 5(a).

[127] *See, e.g.*, Sam Altman, Written Testimony of Sam Altman Before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law (2023), https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf [https://perma.cc/HLL5-ATKG] ("[T]he U.S. government should consider a combination of licensing or registration requirements for development and release of AI models above a crucial threshold of capabilities"); Microsoft, *supra* note 1, at 21 ("To achieve safety and security objectives, we envision licensing requirements such as advance notification of large training runs. . . . Microsoft will support the development of a national registry of high-risk AI systems that is open for inspection so that members of the public can learn where and how those systems are in use."); Bengio et al., *supra* note 72, at 844 (identifying measures to mitigate risks from "exceptionally capable future AI systems" and stating that "[g]overnments must be prepared to license their development"). Some argue that licensing regimes are warranted only for the highest-risk AI activities, where there is evidence of sufficient chance of large-scale harm and other regulatory approaches appear inadequate. *See* Anderljung et al., *supra* note 1, at 20–21.

[128] Hadfield et al., *supra* note 1 ("[G]overnments should establish national registries for large generative AI models over a threshold defined by size (number of parameters or amount of compute used for training, for example) and capabilities."); *cf.* Bengio et al., *supra* note 72, at 843 ("Regulators should mandate . . . registration of key information on frontier AI systems and their datasets throughout their life cycle and monitoring of model development.").

[129] *See* Egan & Heim, *supra* note 1, at 3, 7–10; Lennart Heim et al., *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*, OXFORD MARTIN SCHOOL (Mar. 2024), https://www.oxfordmartin.ox.ac.uk/publications/governing-through-the-cloud-the-intermediary-role-of-compute-providers-in-ai-regulation/ [https://perma.cc/9AAZ-QGBP], at 21–23.

[130] *Senators Mitt Romney, Jack Reed, Jerry Moran & Angus S. King, Jr., Framework for Mitigating Extreme AI Risks* (Apr. 16, 2024), https://www.romney.senate.gov/wp-content/uploads/2024/04/AI-Framework_2pager.pdf [https://perma.cc/WL66-4T3W]; *Letter from Senators Mitt Romney, Jack Reed, Jerry Moran & Angus S. King, Jr. to Senators Chuck Schumer, Mike Rounds, Martin Heinrich & Todd Young* (Apr. 16, 2024), https://www.romney.senate.gov/wp-content/uploads/2024/04/240415-AI-Letter-final.pdf [https://perma.cc/3A3K-DKZT].

Conduct for Advanced AI Systems, agreed upon by G7 leaders on October 30, 2023.[131] Compute thresholds could identify when specific obligations are triggered in other proposals, including proposals for: (1) conducting thorough risk assessments of frontier AI models before deployment;[132] (2) subjecting AI development to evaluation-gated scaling;[133] (3) pausing development of frontier AI;[134] (4) subjecting developers of advanced models to governance audits;[135] (5) monitoring advanced models after deployment;[136] and (6) requiring that advanced AI models be subject to information security protections.[137]

B. Why Might Compute Be Relevant Under Existing Law?

Even without a formal compute threshold, the significance of training compute could affect the interpretation and application of existing laws. Courts and regulators may rely on compute as a proxy for how much risk a given AI system poses—alongside other factors such as capabilities, domain, safeguards, and whether the application is in a higher-risk context—when determining whether a legal condition or regulatory threshold has been met. This section briefly covers a few examples. First, it discusses the potential implications for duty of care and foreseeability analyses in tort law. It then goes on to describe how regulatory agencies could depend on training compute as one of several factors in evaluating risk from frontier AI, for example as an indicator of change to a regulated product and as a factor in regulatory impact analysis.

---

[131] *See* Hiroshima Process International Code of Conduct for Advanced AI Systems (Oct. 30, 2023), https://ec.europa.eu/newsroom/dae/redirection/document/99641[https://perma.cc/MP2B-53VJ] (recommending, among others, to take appropriate measures to identify, evaluate, and mitigate risks across the AI lifecycle, identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market, and publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use).

[132] Anderljung et al., *supra* note 1, at 3, 23; Shevlane et al., *supra* note 83, at 1 ("Developers must be able to identify dangerous capabilities (through 'dangerous capability evaluations') and the propensity of models to apply their capabilities for harm (through 'alignment evaluations')."); Markus Anderljung et al., *Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem Under the ASPIRE Framework*, ARXIV (Nov. 15, 2023), https://doi.org/10.48550/arXiv.2311.14711 [https://perma.cc/KUQ7-95XR].

[133] *See* Jide Alaga & Jonas Schuett, *Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers*, ARXIV (Sept. 30, 2023), https://doi.org/10.48550/arXiv.2310.00374 [https://perma.cc/NM9G-U6M6]; *Anthropic's Responsible Scaling Policy*, ANTHROPIC (Sept. 19, 2023), https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf [https://perma.cc/SZX9-LLPU]; *Responsible Scaling Policies (RSPs)*, MODEL EVALUATION & THREAT RESEARCH (METR) (last updated Oct. 26, 2023), https://metr.org/blog/2023-09-26-rsp [https://perma.cc/84NK-ZMXK]; Evan Hubinger, *RSPs Are Pauses Done Right*, AI ALIGNMENT FORUM (Oct. 14, 2023), https://www.alignmentforum.org/posts/mcnWZBnbeDz7KKtjJ/rsps-are-pauses-done-right [https://perma.cc/NVN7-AFPE].

[134] *Pause Giant AI Experiments: An Open Letter*, FUTURE LIFE INST. (Mar. 22, 2023), https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf [https://perma.cc/5YTS-DMXB].

[135] *See generally* Jakob Mökander et al., *Auditing Large Language Models: A Three-Layered Approach*, AI ETHICS (2023), https://doi.org/10.1007/s43681-023-00289-2 [https://perma.cc/DUQ2-7QP3].

[136] Anderljung et al., *supra* note 1, at 27; *see also* Joe O'Brien et al., *Deployment Corrections: An Incident Response Framework for Frontier AI Models,* INST. FOR AI POL'Y & STRATEGY (Sept. 30, 2023), https://doi.org/10.48550/arXiv.2310.00328 [https://perma.cc/LA7Z-KSBG], at 23–25.

[137] Luke Muehlhauser, *12 Tentative Ideas for US AI Policy*, OPEN PHILANTHROPY (Apr. 17, 2023), https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/ [https://perma.cc/YZ77-4X3Y]. Some cybersecurity requirements have already been established by the EU AI Act, *supra* note 4, art. 15 ("High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle. . . . The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.") and Art. 55 ("[P]roviders of general-purpose AI models with systemic risk shall . . . ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model."). The Executive Order on AI mandates reporting on physical and cybersecurity measures but does not require specific measures. Exec. Order on AI, *supra* note 2, § 4.2(a) (requiring "[c]ompanies developing or demonstrating an intent to develop potential dual-use foundation models" to report the "physical and cybersecurity protections taken to assure the integrity of that training process against sophisticated threats" and "the physical and cybersecurity measures taken to protect [] model weights").

The application of existing laws and ongoing development of common law, such as tort law, may be particularly important while AI governance is still nascent[138] and may operate as a complement to regulations once developed.[139] However, courts and regulators will face new challenges as cases involve AI, an emerging technology of which they have no specialized knowledge, and parties will face uncertainty and inconsistent judgments across jurisdictions. As developments in AI unsettle existing law[140] and agency practice, courts and agencies might rely on compute in several ways.

For example, compute could inform the duty of care owed by developers who make voluntary commitments to safety.[141] A duty of care, which is a responsibility to take reasonable care to avoid causing harm to another, can be conditioned on the foreseeability of the plaintiff as a victim or be an affirmative duty to act in a particular way; affirmative duties can arise from the relationship between the parties, such as between business owner and customer, doctor and patient, and parent and child.[142] If AI companies make general commitments to security testing and cybersecurity, such as the voluntary safety commitments secured by the Biden administration,[143] those commitments may give rise to a duty of care in which training compute is a factor in determining what security is necessary. If a lab adopts a responsible scaling policy that requires it to have protection measures based on specific capabilities or potential for risk or misuse,[144] a court

---

[138] *Cf.* Gary E. Marchant, *Governance of Emerging Technologies as a Wicked Problem*, 73 VANDERBILT L. REV. 1861, 1875 (2020) (discussing liability as one of several governance options for emerging technologies in the context of gene drives and noting its importance "when government regulations do not exist").

[139] *See generally* Mary L. Lyndon, *Tort Law and Technology*, 12(137) YALE J. ON REGUL. 137 (1995). However, tort liability for software defects has been quite limited. *See* Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 41–42 and accompanying text (2019) ("Tort liability for software failures is a rarity. . . . Courts uniformly dismiss claims of software defect, often because there is no physical injury at stake, but also for a broad range of other disqualifying reasons. And even when the plaintiff alleges an eligible injury, it remains exceedingly difficult to prove whether the software caused the injury, and whether that cause was due to some defect intrinsic to the software."); Jacob Kreutzer, *Somebody Has to Pay: Products Liability for Spyware*, 45 AM. BUS. L.J. 61, 74 (2008) ("The few defective software cases brought as tort claims have generally been dismissed as only involving economic damages."). For a review of how tort law could be applied to AI-related harms, see Gabriel Weil, *Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence*, SSRN 21–44 (June 6, 2024), https://dx.doi.org/10.2139/ssrn.4694006 [https://perma.cc/HCB7-7GG8].

[140] *Cf.* European Commission, Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics (Feb. 19, 2020), 12–16, https://op.europa.eu/en/publication-detail/-/publication/4ce205b8-53d2-11ea-aece-01aa75ed71a1 [https://perma.cc/8EY6-ADSM] (discussing how AI challenges existing legal frameworks); B.J. Ard, *Making Sense of Legal Disruption*, 2022(4) WIS. L. REV. FORWARD 42, 46–47 (2022) ("Countless law review articles have invoked disruption to describe the process whereby new technologies unsettle existing law and force courts and lawmakers to reexamine legal doctrine."); Margot E. Kaminski, *Authorship, Disrupted: AI Authors in Copyright and First Amendment Law*, 51 U.C. DAVIS L. REV. 589, 589–90 (2017) (collecting examples).

[141] *Cf.* Vincent R. Johnson, *Cybersecurity, Identity Theft, and the Limits of Tort Liability*, 57 S. C. L. REV. 255, 278–80 (2005) (discussing voluntary assumption of duty in the context of data protection).

[142] *See generally* W. Jonathan Cardi, *The Hidden Legacy of* Palsgraf*: Modern Duty Law in Microcosm*, 91 BOS. UNIV. L. REV. 1873 (2011) (surveying state law).

[143] The White House has obtained voluntary commitments from several companies to better understand and address risks from AI. THE WHITE HOUSE, BIDEN-HARRIS ADMINISTRATION SECURES VOLUNTARY COMMITMENTS FROM LEADING ARTIFICIAL INTELLIGENCE COMPANIES TO MANAGE THE RISKS POSED BY AI (July 21, 2023) [hereinafter VOLUNTARY COMMITMENTS], https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ [https://perma.cc/ZA8A-8KHR] (announcing Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI); THE WHITE HOUSE, BIDEN-HARRIS ADMINISTRATION SECURES VOLUNTARY COMMITMENTS FROM EIGHT ADDITIONAL ARTIFICIAL INTELLIGENCE COMPANIES TO MANAGE THE RISKS POSED BY AI (Sept. 12, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ [https://perma.cc/6KTZ-MXYD]; THE WHITE HOUSE, BIDEN-HARRIS ADMINISTRATION ANNOUNCES NEW AI ACTIONS AND RECEIVES ADDITIONAL MAJOR VOLUNTARY COMMITMENT ON AI (July 26, 2024), https://www.whitehouse.gov/briefing-room/statements-releases/2024/07/26/fact-sheet-biden-harris-administration-announces-new-ai-actions-and-receives-additional-major-voluntary-commitment-on-ai/ [https://perma.cc/8278-7GYB] (announcing Apple).

[144] *Responsible Scaling Policies (RSPs)*, *supra* note 133; *Anthropic's Responsible Scaling Policy*, *supra* note 133; OpenAI, *Preparedness*

might consider training compute as one of several factors in evaluating the potential for risk or misuse.

A court might also consider training compute as a factor when determining whether a harm was foreseeable. More advanced AI systems, trained with more compute, could foreseeably be capable of greater harm, especially in light of scaling laws discussed in Section I.C that make clear the relationship between compute and performance. It may likewise be foreseeable that a powerful AI system could be misused[145] or become the target of more sophisticated attempts at exfiltration, which might succeed without adequate security.[146] Foreseeability may in turn bear on negligence elements of proximate causation and duty of care.

Compute could also play a role in other scenarios, such as in a false advertising claim under the Lanham Act[147] or state and federal consumer protection laws. If a business makes a claim about its AI system or services that is false or misleading, it could be held liable for monetary damages and enjoined from making that claim in the future (unless it becomes true).[148] While many such claims will not involve compute, some may; for example, if a lab publicly claims to follow a responsible scaling policy, training compute could be relevant as an indicator of model capability and the corresponding security and safety measures promised by the policy.

Regulatory agencies may likewise consider compute in their analyses and regulatory actions. For example, the Environmental Protection Agency could consider training (and inference) compute usage as part of environmental impact assessments.[149] Others could treat compute as a proxy for threat to national or public security. Agencies and committees responsible for identifying and responding to various risks, such as the Interagency Committee on Global Catastrophic Risk[150] and Financial Stability Oversight Council,[151] could consider compute in their evaluation of risk from frontier AI. Over fifty federal agencies were directed to take specific

---

Framework (Beta), OPENAI (Dec. 18, 2023), https://cdn.openai.com/openai-preparedness-framework-beta.pdf [https://perma.cc/725B-LVK5]; Anca Dragan et al., *Introducing the Frontier Safety Framework*, GOOGLE DEEPMIND (May 17, 2024), https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/ [https://perma.cc/VHT8-JQ2Q]; *Google DeepMind's Frontier Safety Framework, Version 1.0*, GOOGLE DEEPMIND (May 17, 2024), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf [https://perma.cc/H9FA-M6SD].

[145] *Cf.* Weil, *supra* note 139, at 35 (noting that "crimes committed against third parties using licensed advanced AI may well give rise to liability if there is some factual basis in the record supporting the claim that the misuse was foreseeable").

[146] *See generally* Sella Nevo et al., *Securing AI Model Weights*, RAND (May 30, 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html [https://perma.cc/WA5E-4RMN].

[147] Section 43(a) of the Lanham Act, 15 U.S.C. § 1125 (2018).

[148] For an overview of federal and state consumer protection laws, see Consumer Rights and the Law, JUSTIA (last reviewed Oct. 2024), https://www.justia.com/consumer/consumer-protection-law/ [https://perma.cc/JJJ3-G9N3]; *False Advertising Under Consumer Protection Laws*, JUSTIA (last reviewed Oct. 2024), https://www.justia.com/consumer/deceptive-practices-and-fraud/false-advertising/ [https://perma.cc/U4US-FN9T]; Gregory Klass, *False Advertising Law*, in OXFORD HANDBOOK OF THE NEW PRIVATE LAW 391 (Andrew S. Gold et al. eds., 2020) (providing an overview of false advertising law, duties to consumers and competitors, and remedies).

[149] For a discussion of environmental impacts, see OECD, *supra* note 42; Strubell et al., *supra* note 42; van Wynsberghe, *supra* note 42.

[150] The committee was established under the Global Catastrophic Risk Management Act, which mandates interagency assessment of global catastrophic risk, reporting on global catastrophic and existential risk every ten years, and development and validation of strategies to ensure health, safety, and welfare in case of catastrophe. Global Catastrophic Risk Management Act of 2022 in National Defense Authorization Act for Fiscal Year 2023, H.R. 7776, 117th Cong. §§ 7301–7309 (2022).

[151] *See* FIN. STABILITY OVERSIGHT COUNCIL, ANNUAL REPORT 2023, (2023), https://home.treasury.gov/system/files/261/FSOC2023AnnualReport.pdf [https://perma.cc/Q7HU-CP7R]; *2024 Conference on Artificial Intelligence & Financial Stability*, U.S. DEP'T TREASURY (June 6–7, 2024), https://home.treasury.gov/policy-issues/financial-markets-financial-institutions-and-fiscal-service/financial-stability-oversight-council/2024-conference-on-artificial-intelligence-financial-stability [https://perma.cc/4S8R-6L5M].

actions to promote responsible development, deployment, federal use of AI, and regulation of industry, in the government-wide effort established by Executive Order 14,110[152]—although these actions are now under review.[153] Even for agencies not directed to consider compute or implement a preliminary compute threshold, compute might factor into how guidance is implemented over time.

More speculatively, changes to training compute could be used by agencies as one of many indicators of how much a regulated product has changed, and thus whether it warrants further review. For example, the Food and Drug Administration might consider compute when evaluating AI in medical devices or diagnostic tools.[154] While AI products considered to be medical devices are more likely to be narrow AI systems trained on comparatively less compute, significant changes to training compute may be one indicator that software modifications require premarket submission. The ability to measure, report, and verify compute[155] could make this approach particularly compelling for regulators.

Finally, training compute may factor into regulatory impact analyses, which evaluate the impact of proposed and existing regulations through quantitative and qualitative methods such as cost-benefit analysis.[156] While this type of analysis is not necessarily determinative, it is often an important input into regulatory decisions and necessary for any "significant regulatory action."[157] As agencies develop and propose new regulations and consider how those rules will affect or be affected by AI, compute could be relevant in drawing lines that define what conduct and actors are affected. For example, a rule with a higher compute threshold and narrower scope may be less significant and costly, as it covers fewer models and developers. The amount of compute used to train models now and in the future may be not only a proxy for threat to national security (or innovation, or economic growth), but also a source of uncertainty, given the potential for emergent

---

[152] *See* Exec. Order on AI, *supra* note 2; LAURA HARRIS & CHRIS JAIKARAN, CONG. RSCH. SERV. R47843, HIGHLIGHTS OF THE 2023 EXECUTIVE ORDER ON ARTIFICIAL INTELLIGENCE FOR CONGRESS (Apr. 3, 2024).

[153] Exec. Order No. 14,179, § 5(a), 90 Fed. Reg. 8741 (Jan. 23, 2025).

[154] *Cf. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)—Discussion Paper and Request for Feedback*, FOOD & DRUG ADMIN., REGULATIONS.GOV (Apr. 1, 2019), https://www.regulations.gov/document/FDA-2019-N-1185-0001 [https://perma.cc/QF6F-73XH] (noting the need to "maintain reasonable assurance of safety and effectiveness . . . while allowing the software to continue to learn and evolve over time to improve patient care"). However, compute was not specifically mentioned in subsequent draft guidance. FOOD & DRUG ADMIN., MARKETING SUBMISSION RECOMMENDATIONS FOR A PREDETERMINED CHANGE CONTROL PLAN FOR ARTIFICIAL INTELLIGENCE/MACHINE LEARNING (AI/ML)-ENABLED DEVICE SOFTWARE FUNCTIONS, (2023), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial [https://perma.cc/YC2V-SG9K].

[155] Girish Sastry et al., *Computing Power and the Governance of Artificial Intelligence*, CTR. FOR THE GOVERNANCE OF AI (Feb. 14, 2024), https://www.governance.ai/research-paper/computing-power-and-the-governance-of-artificial-intelligence [https://perma.cc/TF2K-W2G8], at 4, 27–28.

[156] *See* OFF. INFO. & REG. AFF., CIRCULAR A-4, REGULATORY IMPACT ANALYSIS: A PRIMER (Aug. 15, 2011), https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/inforeg/inforeg/regpol/circular-a-4_regulatory-impact-analysis-a-primer.pdf [https://perma.cc/46J9-ZSQX]; OECD, *Regulatory Impact Assessment*, *in* OECD BEST PRACTICE PRINCIPLES FOR REGULATORY POLICY (2020), https://doi.org/10.1787/7a9638cb-en [https://perma.cc/2QDW-WTKG].

[157] Exec. Order No. 14,094, § 1(b), 3 C.F.R. 14094 (2024) (amending Exec. Order No. 12,866, § 3(f) to define "[s]ignificant regulatory action" to include actions likely to result in a rule with an annual effect on the economy of $200 million or more or with the potential to "adversely affect in a material way the economy, a sector of the economy, productivity, competition, jobs, the environment, public health or safety, or State, local, territorial, or tribal governments or communities"). The Office of Management and Budget provides guidance on regulatory analysis in the Circular A-4. *See* OFF. MGMT & BUDGET, CIRCULAR A-4: REGULATORY ANALYSIS (Sept. 17, 2003), https://obamawhitehouse.archives.gov/omb/circulars_a004_a-4/ [https://perma.cc/MM4G-XU3E]; *see also* OFF. MGMT & BUDGET, *Draft Circular A-4: Regulatory Analysis* (Apr. 6, 2023).

capabilities.

C. Where Should the Compute Threshold(s) Sit?

The choice of compute threshold depends on the policy under consideration: what models are the intended target, given the purpose of the policy? What are the burdens and costs of compliance? Can the compute threshold be complemented with other elements for determining whether a model falls within the scope of the policy, in order to more precisely accomplish its purpose?

Some policy proposals would establish a compute threshold "at the level of FLOP used to train *current* foundational models."[158] While the training compute of many models is not public, according to estimates, the largest models today were trained with 1e25 FLOP or more, including at least one open-source model, Llama 3.1 405B.[159] This is the initial threshold established by the EU AI Act. Under the Act, general-purpose AI models are considered to have "systemic risk," and thus trigger a series of obligations for their providers, if found to have "high impact capabilities."[160] Such capabilities are presumed if the *cumulative* amount of training compute, which includes all "activities and methods that are intended to enhance the capabilities of the model prior to deployment, such as pre-training, synthetic data generation and fine-tuning," exceeds 1e25 FLOP.[161] This threshold encompasses existing models such as Gemini Ultra and GPT-4, and it can be updated upwards or downwards by the European Commission through delegated acts.[162] During the AI Safety Summit held in 2023, the U.K. Government included current models by defining "frontier AI" as "highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models" and acknowledged that the definition included the models underlying ChatGPT, Claude, and Bard.[163]

Others have proposed an initial threshold of "*more* training compute than already-deployed

---

[158] Egan & Heim, *supra* note 1, at 9 (emphasis added); *see also* Comment on ANPRM *supra* note 1, at 16 ("[P]lacing a compute threshold at roughly the training compute budget of today's frontier models could be an appropriate initial threshold.").

[159] *See Large-Scale AI Models*, EPOCH (July 31, 2024), https://epochai.org/data/large-scale-ai-models [https://perma.cc/QD8M-TSCG]; *Introducing Llama 3.1: Our Most Capable Models to Date*, META, https://ai.meta.com/blog/meta-llama-3-1/ [https://perma.cc/7CXL-47EM].

[160] EU AI Act, *supra* note 4, art. 51(1).

[161] *Id.*, Recital 111, art. 51(2) ("A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^25."); Luca Bertuzzi, *AI Act: EU Policymakers Nail Down Rules on AI Models, Butt Heads on Law Enforcement,* EURACTIV (Dec. 7, 2023), https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-policymakers-nail-down-rules-on-ai-models-butt-heads-on-law-enforcement/ [https://perma.cc/6KNQ-DU4S] ("[A]utomatic categorisation as 'systemic' for models that were trained with computing power above 10^25 floating point operations."). The threshold that found the agreement of the EU institutions might have been reduced from a prior compute threshold higher than 1e26 FLOP. *See* Luca Bertuzzi, *AI Act: EU Commission Attempts to Revive Tiered Approach Shifting to General Purpose AI*, EURACTIV (Nov. 20, 2023), https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-commission-attempts-to-revive-tiered-approach-shifting-to-general-purpose-ai/ [https://perma.cc/6EW2-TEGX].

[162] EU AI Act, *supra* note 4, art. 51(3) ("The Commission shall adopt delegated acts in accordance with Article 97 to amend the thresholds listed in paragraphs 1 and 2 of this Article, as well as to supplement benchmarks and indicators in light of evolving technological developments, such as algorithmic improvements or increased hardware efficiency, when necessary, for these thresholds to reflect the state of the art.").

[163] U.K., Department for Science, Innovation and Technology, *AI Safety Summit: Introduction*, GOV.UK (last updated Oct. 31, 2023), https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html [https://perma.cc/ZXM8-L6XZ], at 4.

systems,"[164] such as 1e26 FLOP[165] or 1e27 FLOP.[166] No known model currently exceeds 1e26 FLOP training compute, which is roughly five times the compute used to train GPT-4.[167] These higher thresholds would more narrowly target future systems that pose greater risks, including potential catastrophic and existential risks.[168] President Biden's Executive Order on AI[169] and recently-vetoed California Senate Bill 1047[170] are in line with these proposals, both targeting models trained with *more* than 1e26 OP or FLOP.

Far more models would fall within the scope of a compute threshold set *lower* than current frontier models. While only two models exceeded 1e23 FLOP training compute in 2017, over 200 models meet that threshold today.[171] As discussed in Section II.A, compute thresholds operate as a trigger for additional scrutiny, and more models falling within the ambit of regulation would entail a greater burden not only on developers, but also on regulators.[172] These smaller, general-purpose models have not yet posed extreme risks, making a lower threshold unwarranted at this time.[173]

While the debate has centered mostly around the establishment of a single training compute threshold, governments could adopt a *pluralistic* and *risk-adjusted* approach by introducing

---

[164] *See* Hadfield et al., *supra* note 1 ("Given the dramatic shift in capabilities demonstrated by OpenAI's GPT-4, the threshold should be set near and slightly above the capabilities of this model."); Egan & Heim, *supra* note 1, at 7; *see also* Anderljung et al., *supra* note 1, at 30.

[165] Jeff Alstott, *Preparing the Federal Response to Advanced Technologies*, Testimony before the U.S. Senate Committee on Homeland Security and Governmental Affairs, Subcommittee on Emerging Threats and Spending Oversight, RAND (2023), https://www.rand.org/content/dam/rand/pubs/testimonies/CTA2900/CTA2953-1/RAND_CTA2953-1.pdf [https://perma.cc/YA5E-LHNS], at 3; *see also* Nicolas Moës & Frank Ryan, *Heavy Is the Head That Wears the Crown: A Risk-Based Tiered Approach to Governing General Purpose AI*, FUTURE SOCIETY (Sept. 2023), https://thefuturesociety.org/wp-content/uploads/2023/09/heavy-is-the-head-that-wears-the-crown.pdf [https://perma.cc/DCG3-KGLH], at 51–53 & tbl.4 (proposing a tiered system for governance of general purpose model that uses a compute threshold of 1e26 FLOP for prohibiting development); CTR. FOR AI POL'Y, Responsible Advanced AI Act, § 3(u) (Apr. 2024), https://assets.caip.org/caip/RAAIA%20(April%202024).pdf [https://perma.cc/TA6V-F7ST] (proposing tiers of AI models according to how likely they are to generate major security risks, with initial criteria that would classify a model trained on at least 1e26 FLOP as a "high-concern AI system").

[166] Jason Matheny, *Artificial Intelligence: Challenges and Opportunities for the Department of Defense*, Testimony before the U.S. Senate Committee on Armed Services, Subcommittee on Cybersecurity, RAND (Apr. 19, 2023), https://doi.org/10.7249/CTA2723-1 [https://perma.cc/RA9Z-FHWC], at 2 (proposing a 1e27 OP threshold for reporting a training run).

[167] *See Notable AI Models*, *supra* note 22 (estimating the training compute for GPT-4 as 2.1e25).

[168] *See supra* notes 70–71 (collecting sources on the potential risk of current and future models).

[169] Exec. Order on AI, *supra* note 2, §§ 4.1(c)(iii) & 4.2(b)(i).

[170] S.B. 1047, 2023–2024 Reg. Sess. (Cal. 2024) § 3 (as enrolled, Sept. 3, 2024).

[171] *Large-Scale AI Models*, *supra* note 159.

[172] Egan & Heim, *supra* note 1, at 7 ("Setting the threshold to capture and monitor the compute of all AI models would not be beneficial, as it would capture too much information to be useful while imposing a significant imposition on industry. Such risks could instead be managed through other safeguards."). Nonetheless, some have proposed a moratorium on development of models that exceed 1e24. Miotti & Wasil, *supra* note 1, at 11 ("[W]e believe an initial moratorium threshold of 10^24 FLOP would be an appropriate starting point."); Jolyn Khoo & Nik Samoylov, Submission to the High-level Advisory Body on Artificial Intelligence's Call for Papers on Global AI Governance, by the Office of the UN Secretary-General's Envoy on Technology, Campaign for AI Safety (Oct. 18, 2023), https://www.campaignforaisafety.org/submission-to-the-high-level-advisory-body-on-artificial-intelligences-call-for-papers-on-global-ai-governance-by-the-office-of-the-secretary-generals-envoy-on-technology/ [https://perma.cc/97TP-YBQJ] (proposing a prohibition on training models with over 1e24 FLOP, with the potential to revise that threshold "as new safety research is published and if models become smaller").

[173] The capabilities of small language models have been growing significantly, in some cases matching the capabilities of much larger models. *See, e.g.*, Misha Bilenko, *Introducing Phi-3: Redefining What's Possible with SLMs*, MICROSOFT (Apr. 23, 2024), https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/ [https://perma.cc/2LTZ-JUGL]; Marah Abdin et al., *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*, ARXIV (Aug. 30, 2024), https://doi.org/10.48550/arXiv.2404.14219 [https://perma.cc/Y5R2-GEPN]. As discussed in Part II.E, compute thresholds may complement metrics used to target other risks, such as those from small language models.

multiple compute thresholds that trigger different measures or requirements according to the degree or nature of risk. Some proposals recommend a tiered approach that would create fewer obligations for models trained on less compute. For example, the Responsible Advanced Artificial Intelligence Act of 2024 would require pre-registration and benchmarks for lower-compute models, while developers of higher-compute models must submit a safety plan and receive a permit prior to training or deployment.[174] Multi-tiered systems may also incorporate a higher threshold beyond which no development or deployment can take place, with limited exceptions, such as for development at a multinational consortium working on AI safety and emergency response infrastructure[175] or for training runs and models with strong evidence of safety.[176]

Domain-specific thresholds could be established for models that possess capabilities or expertise in areas of concern and models that are trained using less compute than general-purpose models.[177] A variety of specialized models are already available to advance research, trained on extensive scientific databases.[178] As discussed in Part I.D, these models present a tremendous opportunity, yet many have also recognized the potential threat of their misuse to research, develop, and use chemical, biological, radiological, and nuclear weapons.[179] To address these risks, President Biden's Executive Order on AI, which set a compute threshold of 1e26 FLOP to trigger reporting requirements, set a substantially lower compute threshold of 1e23 FLOP for models trained "using primarily biological sequence data."[180] The Hiroshima Process International Code of Conduct for Advanced AI Systems likewise recommends devoting particular attention to offensive cyber capabilities and chemical, biological, radiological, and nuclear risks, although it does not propose a compute threshold.[181]

While domain-specific thresholds could be useful for a variety of policies tailored to specific risks, there are some limitations. It may be technically difficult to verify how much biological sequence data (or other domain-specific data) was used to train a model.[182] Another

---

[174] CTR. FOR AI POL'Y, *supra* note 165, § 3(u) (classifying models based on security risk, with "low concern" defined as those trained on less than 1e24 FLOP, "medium concern" as those trained on at least 1e24 but less than 1e25, and "high concern" as those trained on at least 1e26 FLOP); *see also* Moës & Ryan, *supra* note 165, at 73–94 (proposing various measures, including reporting, registration, Know-Your-Customer measures, and auditing, for general-purpose models according to training compute, grouped into Type-I models trained on at least 1e21 FLOP, Type-II models trained on at least 1e23 FLOP, and potentially prohibited models trained on over 1e26 FLOP).

[175] *See* Miotti & Wasil, *supra* note 1, at 9–10.

[176] CTR. FOR AI POL'Y, *supra* note 165, § 9.

[177] *See* Exec. Order on AI, *supra* note 2, § 4.2(b)(i); Comment on ANPRM, *supra* note 1, at 15 (noting that, for models in certain domains, such as biosecurity and cybersecurity, "thresholds will be more static, and capture an absolute level of risk" and "the development of protective measures could render a particular threshold obsolete").

[178] *See* Nicole Maug et al., *Biological Sequence Models in the Context of the AI Directives*, EPOCH (Apr. 9, 2024), https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives [https://perma.cc/E5FW-7KRS] (discussing models trained on biological sequence data); *Notable AI Models*, *supra* note 22 (compiling models across several domains).

[179] *See supra* notes 85–87 and accompanying text; Exec. Order on AI, *supra* note 2, § 3(k)(i); DEP'T OF HOMELAND SEC., DEPARTMENT OF HOMELAND SECURITY REPORT ON REDUCING THE RISKS AT THE INTERSECTION OF ARTIFICIAL INTELLIGENCE AND CHEMICAL, BIOLOGICAL, RADIOLOGICAL, AND NUCLEAR THREATS 8–19 (Apr. 26, 2024), https://www.dhs.gov/sites/default/files/2024-06/24_0620_cwmd-dhs-cbrn-ai-eo-report-04262024-public-release.pdf [https://perma.cc/J2HT-BPDT] ("The increased proliferation and capabilities of AI tools . . . may lead to significant changes in the landscape of threats to U.S. national security over time, including by influencing the means, accessibility, or likelihood of a successful CBRN attack").

[180] Exec. Order on AI, *supra* note 2, § 4.2(b)(i).

[181] Hiroshima Process International Code of Conduct for Advanced AI Systems, *supra* note 131, at 3.

[182] *See generally* Dami Choi et al., *Tools for Verifying Neural Models' Training Data*, ARXIV (July 2, 2023), https://doi.org/10.48550/arXiv.2307.00682 [https://perma.cc/YY9F-2UEW] (introducing a verification tool while also highlighting that verifying training data is challenging and requires access to snapshots and checkpoints of the model training).

challenge is specifying how much data in a given domain causes a model to fall within scope, particularly considering the potential capabilities of models trained on mixed data.[183] Finally, the amount of training compute required may be so low that, over time, a compute threshold is not practical.

When choosing a threshold, regulators should be aware that capabilities might be substantially improved through *post-training enhancements*, and training compute is only a general predictor of capabilities. The absolute limits are unclear at this point; however, current methods can result in capability improvements equivalent to a 5- to 30-times increase in training.[184] To account for post-training enhancements, a governance regime could create a *safety buffer*, in which oversight or other protective measures are set at a lower threshold.[185] Along similar lines, *open-source models* may warrant a lower threshold for at least some regulatory requirements, since they could be further trained by another actor and, once released, cannot be moderated or rescinded. [186]

D. Does a Compute Threshold Require Updates?

Once established, compute thresholds and related criteria will likely require updates over time.[187] Improvements in algorithmic efficiency could reduce the amount of compute needed to train an equally capable model,[188] or a threshold could be raised or eliminated if adequate protective measures are developed or if models trained with a certain amount of compute are demonstrated to be safe.[189] To further guard against future developments in a rapidly evolving field, policymakers can authorize regulators to update compute thresholds and related criteria.[190]

---

[183] Maug et al., *supra* note 178 (noting that, since Executive Order requires the model to be trained "primarily" on biological sequence data to be subject to the lower compute threshold, "[m]odels trained on less than 1e26 FLOPs could potentially incorporate all known protein sequences while evading oversight by not being primarily biological").

[184] Davidson et al., *supra* note 21, at tbl.1, 4–5 (summarizing post-training enhancements and their corresponding compute-equivalent gain).

[185] *Id.* at 22–23. For more on post-training enhancements, see *supra* note 21 (collecting references).

[186] NAT'L TELECOMMS. & INFO. ADMIN., U.S. DEP'T OF COM., DUAL-USE FOUNDATION MODELS WITH WIDELY AVAILABLE MODEL WEIGHTS 8 (July 2024); Anderljung et al., *supra* note 1, at 36.

[187] *See, e.g.*, Egan & Heim, *supra* note 1, at 3, 9 (noting that a "threshold would need to be dynamic and subject to periodic reassessments by government."); Christoph Winter & Charlie Bullock, *The Governance Misspecification Problem* (Inst. for L. & AI, Working Paper No. 3-2024), https://law-ai.org/wp-content/uploads/2024/10/Governance-misspecification-1.pdf [https://perma.cc/9N5J-69SW] (observing that "any well-specified legal rule that uses a compute threshold is likely to be rendered both overinclusive and underinclusive soon after being implemented"); Hooker, *supra* note 53, at 20–23; *The Limits of Thresholds: Exploring the Role of Compute-Based Thresholds for Governing the Risks of AI Models*, COHERE FOR AI (July 2024), https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf [https://perma.cc/46EU-W7E3], at 14 (recommending "dynamic rather than static thresholds"); Comment on ANPRM, *supra* note 1, at 16 ("thresholds will be a constantly moving target"); Helen Toner & Timothy Fist, *Regulating the AI Frontier: Design Choices and Constraints*, CTR. FOR SEC. & EMERGING TECH. (Oct. 26, 2023), https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/ [https://perma.cc/V8GL-3EB4] (observing that "targeting frontier AI regulation purely based on compute thresholds (e.g., stipulating that any AI model that was trained with a certain level of compute is a 'frontier model') is unlikely to work as a complete solution over the longer term"); Moës & Ryan, *supra* note 165, at 48 (predicting that compute thresholds will be an adequate stop-gap measure of capabilities for the next two years, but will "certainly need to be augmented in the relatively near future with more accurate benchmarks").

[188] *See supra* notes 61–67 and accompanying text and sources (on algorithmic innovation).

[189] *See* Comment on ANPRM, *supra* note 1, at 15 (noting that "the development of protective measures could render a particular threshold obsolete"); *cf.* Lennart Heim & Leonie Koessler, *Training Compute Thresholds: Features and Functions in AI Regulation*, ARXIV 21–23 (Aug. 6, 2024), https://doi.org/10.48550/arXiv.2405.10799 [https://perma.cc/9FQH-BWFN].

[190] Particularly in light of the Supreme Court decision to overturn *Chevron* deference in *Loper Bright*, Congress must be clear in granting authority and discretion to an agency. *Cf.* Loper Bright Enters. v. Raimondo, No. 22-4751, 2024 WL 3208360 (U.S. June 28, 2024).

Several policies, proposed and enacted, have incorporated a dynamic compute threshold. For example, President Biden's Executive Order on AI authorized the Secretary of Commerce to update the initial compute threshold set in the order, as well as other technical conditions for models subject to reporting requirements, "as needed on a regular basis" while establishing an interim compute threshold of 1e26 OP or FLOP.[191] Similarly, the EU AI Act provides that the 1e25 FLOP compute threshold "should be adjusted over time to reflect technological and industrial changes, such as algorithmic improvements" and authorizes the European Commission to amend the threshold and "supplement benchmarks and indicators in light of evolving technological developments."[192] The California Senate Bill 1047 would have created the Frontier Model Division within the Government Operations Agency and authorized it to "update both of the [compute] thresholds in the definition of a 'covered model' to ensure that it accurately reflects technological developments, scientific literature, and widely accepted national and international standards and applies to artificial intelligence models that pose a significant risk of causing or materially enabling critical harms."[193]

Regulators may need to update compute thresholds rapidly. Historically, failure to quickly update regulatory definitions in the context of emerging technologies has led to definitions becoming useless or even counterproductive.[194] In the field of AI, developments may occur quickly and with significant implications for national security and public health, making responsive rulemaking particularly important. In the United States, there are several statutory tools to authorize and encourage expedited and regular rulemaking.[195] For example, Congress could expressly authorize interim or direct final rulemaking, which would enable an agency to shift the comment period in notice-and-comment rulemaking to take place after the rule has already been promulgated, thereby allowing them to respond quickly to new developments.[196]

Policymakers could also require a periodic evaluation of whether compute thresholds are achieving their purpose to ensure that it does not become over- or under-inclusive. While establishing and updating a compute threshold necessarily involves prospective *ex ante* impact assessment, in order to take precautions against risk without undue burdens, regulators can learn much from retrospective *ex post* analysis of current and previous thresholds.[197] In a survey

---

[191] *See* Exec. Order on AI, *supra* note 2, § 4.2(b).

[192] *See* EU AI Act, *supra* note 4, Recital 111 and art. 51(3).

[193] S.B. 1047, 2023–2024 Reg. Sess. (Cal. 2024) § 3 (as enrolled, Sept. 3, 2024).

[194] *See generally* Winter & Bullock, *supra* note 187.

[195] *See generally* KEVIN J. HICKEY, CONG. RSCH. SERV., R45336, AGENCY DELAY: CONGRESSIONAL AND JUDICIAL MEANS TO EXPEDITE AGENCY RULEMAKING (Oct. 5, 2018).

[196] *Id*. Even absent Congressional authorization, an agency may forego notice-and-comment procedures when it "for good cause finds" that those procedures "are impracticable, unnecessary, or contrary to the public interest." 5 U.S.C. § 553(b)(3)(b). Agencies regularly rely on this good cause exception, but the resulting rule may be challenged on procedural grounds. *See generally* Kyle Schneider, *Judicial Review of Good Cause Determinations Under the Administrative Procedure Act*, 73 STAN. L. REV. 237 (2021); JARED P. COLE, CONG. RSCH. SERV., R44356, THE GOOD CAUSE EXCEPTION TO NOTICE AND COMMENT RULEMAKING: JUDICIAL REVIEW OF AGENCY ACTION (Jan. 29, 2016); Connor Raso, *Agency Avoidance of Rulemaking Procedures*, 67 ADMIN. L. REV. 1, 83–93 (2015).

[197] For more on retrospective regulatory analysis, see generally Lori S. Bennear & Jonathan B. Wiener, *Institutional Roles and Goals for Retrospective Regulatory Analysis*, 12(3) J. BENEFIT-COST ANALYSIS 466 (2021); Cary Coglianese, *Moving Forward with Regulatory Lookback*, 30 YALE J. REGUL. ONLINE 57 (2012); Cass R. Sunstein, *The Regulatory Lookback*, 94 BOS. UNIV. L. REV. 579 (2014); Joseph E. Aldy, Learning from Experience: An Assessment of the Retrospective Reviews of Agency Rules and the Evidence for Improving the Design and Implementation of Regulatory Policy, Report to the Admin. Conf. of the U.S. (Nov. 17, 2014); Reeve T. Bull, *Building a*

conducted for the Administrative Conference of the United States, "[a]ll agencies stated that periodic reviews have led to substative [sic] regulatory improvement at least some of time. This was more likely when the underlying evidence basis for the rule, particularly the science or technology, was changing."[198] While the optimal frequency of periodic review is unknown, the study found that U.S. federal agencies were more likely to conduct reviews when provided with a clear time interval ("at least every X years").[199]

Several further institutional and procedural factors could affect whether and how compute thresholds are updated. In order to effectively update compute thresholds and other criteria, regulators must have access to expertise and talent through hiring, training, consultation and collaboration, and other avenues that facilitate access to experts from academia and industry.[200] Decisions will be informed by the availability of data, including scientific and commercial data, to enable ongoing monitoring, learning, analysis, and adaptation in light of new developments. Decision-making procedures, agency design, and influence and pressures from policymakers, developers, and other stakeholders will likewise affect updates, among many other factors.[201] While more analysis is beyond the scope of this Article, others have explored procedural and substantive measures for adaptive regulation[202] and effective governance of emerging technologies.[203]

Framework for Governance: Retrospective Review and Rulemaking Petitions, 67 ADMIN. L. REV. 265 (2015).

[198] Lori S. Bennear & Jonathan B. Wiener, Periodic Review of Agency Regulation, Report to the Admin. Conf. of the U.S. 47 (June 7, 2021) [hereinafter Periodic Review], https://www.acus.gov/sites/default/files/documents/ACUS%20-%20Periodic%20Review%20-%20Periodic%20Review%20of%20Agency%20Regulation%202021%2006%2007%20final%20%281%29.pdf [https://perma.cc/X7W3-WUBX]; *see also* Lori S. Bennear & Jonathan B. Wiener, *Pursuing Periodic Review of Agency Regulation*, REGUL. REV. (Nov. 9, 2021), https://www.theregreview.org/2021/11/09/bennear-wiener-periodic-review/ [https://perma.cc/S7JL-GAJU].

[199] *Id.*

[200] For examples of recent White House efforts, *see* Exec. Order on AI, *supra* note 2, § 10.2; *Bring Your AI Skills to the U.S.*, AI.GOV, https://ai.gov/immigrate/ [https://perma.cc/3VX5-KPCC].

[201] *See generally* Periodic Review, *supra* note 198 (on data collection, agency policies and procedures for review, the role of stakeholders, and more); Jacob Gersen, *Designing Agencies*, *in* RESEARCH HANDBOOK ON PUBLIC CHOICE AND PUBLIC LAW 333 (Daniel A. Farber & Anne Joseph O'Connell eds., 2010) (on agency design generally); PREVENTING REGULATORY CAPTURE: SPECIAL INTEREST INFLUENCE AND HOW TO LIMIT IT (Daniel Carpenter ed., 2013) (on regulatory capture); Rachel E. Barkow, *Insulating Agencies: Avoiding Capture Through Institutional Design*, 89 TEX. L. REV. 15 (2010) (on regulatory capture).

[202] *See generally* Lori S. Bennear & Jonathan B. Wiener, *Built to Learn: From Static to Adaptive Environmental Policy*, in A BETTER PLANET: FORTY IDEAS FOR A SUSTAINABLE FUTURE 353, 356 (Daniel C. Esty ed., 2019) (discussing measures such as "processes for data collection, analysis, review, and potential policy changes," periodic review, and creation of a safety board or investigative body to prepare to learn from a crisis); Lori S. Bennear & Jonathan B. Wiener, *Adaptive Regulation: Instrument Choice for Policy Learning over Time* (Feb. 12, 2019), available at https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/Regulation%20-%20adaptive%20reg%20-%20Bennear%20Wiener%20on%20Adaptive%20Reg%20Instrum%20Choice%202019%2002%2012%20clean.pdf [https://perma.cc/QMA5-UL2K]; Lawrence E. McCray et al., *Planned Adaptation in Risk Regulation: An Initial Survey of US Environmental, Health, and Safety Regulation*, 77(6) TECH. FORECASTING & SOC. CHANGE 951 (2010); Irina Brass & Jesse H. Sowell, *Adaptive Governance for the Internet of Things: Coping with Emerging Security Risks*, 5 REGUL. & GOVERNANCE 1092 (2021); Jesse H. Sowell, *A Conceptual Model of Planned Adaptation (PA)*, in DECISION MAKING UNDER DEEP UNCERTAINTY: FROM THEORY TO PRACTICE 289 (Vincent A.W.J. Marchau et al. eds., 2019); *Governance Innovation Ver.2: A Guide to Designing and Implementing Agile Governance*, JAPAN MINISTRY OF ECONOMY, TRADE AND INDUSTRY (2021), 59–110 https://www.meti.go.jp/press/2021/07/20210730005/20210730005-2.pdf [https://perma.cc/F8Z2-4JUK] (discussing the design and implementation of "agile governance"); CREATING ADAPTIVE POLICIES: A GUIDE FOR POLICY-MAKING IN AN UNCERTAIN WORLD (Darren Swanson & Suruchi Bhadwal eds., 2009).

[203] *See, e.g.*, Gary E. Marchant & Yvonne A. Stevens, *Resilience: A New Tool in the Risk Governance Toolbox for Emerging Technologies*, 51 U.C. DAVIS L. REV. 233 (2017). *See generally* Matthijs M. Maas, *Aligning AI Regulation to Sociotechnical Change*, in THE OXFORD HANDBOOK OF AI GOVERNANCE 358 (Justin B. Bullock et al. eds., 2022); Hadassah Drukarch et al., *An Iterative Regulatory Process for Robot Governance*, 5 DATA & POL'Y e8 (2023).

Some have proposed defining compute thresholds in terms of *effective* compute,[204] as an alternative to updates over time. Effective compute could index to a particular year (similar to inflation adjustments) and thus account for the role that algorithmic progress (e.g., 1e25 of 2023-level effective compute).[205] However, there is not an agreed upon way to more precisely define and calculate effective compute, and the ability to do so depends on the challenging task of calculating algorithmic efficiency, including choosing a performance metric to anchor on. Furthermore, effective compute alone would fail to address potential changes in the risk landscape, such as the development of protective measures.

E. What Are the Advantages and Limitations of a Training Compute Threshold?

Compute has several properties that make it attractive for policymaking: it is (1) correlated with capabilities and thus risk, (2) essential for training, with thresholds that are difficult to circumvent without reducing performance, (3) an objective and quantifiable measure, (4) capable of being estimated before training (5) externally verifiable after training, and (6) a significant cost during development and thus indicative of developer resources. However, training compute thresholds are not infallible: (1) training compute is an imprecise indicator of potential risk, (2) a compute threshold could be circumvented, and (3) there is no industry standard for measuring and reporting training compute.[206] Some of these limitations can be addressed with thoughtful drafting, including clear language, alternative and supplementary elements for defining what models are within scope, and authority to update any compute threshold and other criteria in light of future developments.

First, training compute is correlated with model capabilities and associated risks. Scaling laws predict an increase in performance as training compute increases, and real-world capabilities generally follow (Section I.C). As models become more capable, they may also pose greater risks if they are misused or misaligned (Section I.D). However, training compute is not a precise indicator of downstream capabilities. Capabilities can seemingly emerge abruptly and discontinuously as models are developed with more compute,[207] and the open-ended nature of foundation models means those capabilities may go undetected.[208] Post-training enhancements such as fine-tuning are often not considered a part of training compute, yet they can dramatically improve performance and capabilities with far less compute. Furthermore, not all models with dangerous capabilities require large amounts of training compute; low-compute models with capabilities in certain domains, such as biology or chemistry, may also pose significant risks, such

---

[204] Hernandez & Brown, *supra* note 63, at 13 ("The conception we find most useful is if we imagine how much more efficient it is to train models of interest in 2018 in terms of floating-point operations than it would have been to 'scale up' training of 2012 models until they got to current capability levels. . . . We considered many other conceptions we found less helpful."); Comment on ANPRM, *supra* note 1, at 16–17, app. A.

[205] *See* Comment on ANPRM, *supra* note 1 at 16–17, app. A (suggesting a compute threshold above 1e25 of "2022-level effective compute").

[206] For further discussion of some of these advantages and limitations, see generally Heim & Koessler, *supra* note 189; Lennart Heim & Leonie Koessler, *Training Compute Thresholds: Features and Functions in AI Regulation*, ARXIV 21–23 (Aug. 6, 2024), https://doi.org/10.48550/arXiv.2405.10799 [https://perma.cc/7Z5J-E67H].

[207] *See supra* notes 72–78 and accompanying text (collecting sources on emergent capabilities).

[208] Ganguli et al., *supra* note 47, at 4, 6–8.

as biological design tools that could be used for drug discovery or the creation of pathogens worse than any seen to date.[209] The market may shift towards these smaller, cheaper, more specialized models,[210] and even general-purpose low-compute models may come to pose significant risks. Given these limitations, a training compute threshold cannot capture all possible risks; however, for large, general-purpose AI models, training compute can act as an initial threshold for capturing emerging capabilities and risks.

Second, compute is necessary throughout the AI lifecycle, and a compute threshold would be difficult to circumvent. There is no AI without compute (Section I.A). Due to its relationship with model capabilities, training compute cannot be easily reduced without a corresponding reduction in capabilities, making it difficult to circumvent for developers of the most advanced models. Nonetheless, companies might find "creative ways" to account for how much compute is used for a given system in order to avoid being subject to stricter regulation.[211] To reduce this risk, some have suggested monitoring compute usage below these thresholds to help identify circumvention methods, such as structuring techniques or outsourcing.[212] Others have suggested using compute thresholds alongside additional criteria, such as the model's performance on benchmarks, financial or energy cost, or level of integration into society.[213] As in other fields, regulatory burdens associated with compute thresholds could encourage regulatory arbitrage if a policy does not or cannot effectively account for that possibility.[214] For example, since compute can be accessed remotely via digital means, data centers and compute providers could move to less-regulated jurisdictions.

Third, compute is an objective and quantifiable metric that is relatively straightforward to measure. Compute is a quantitative measure that reflects the number of mathematical operations performed. It does not depend on specific infrastructure and can be compared across different sets

---

[209] *See, e.g.*, Lohn & Musser, *supra* note 39, at 21 (noting that "not all progress requires record-breaking levels of compute" and, for instance, "AlphaFold is revolutionizing aspects of computational biochemistry and only required a few weeks of training on 16 TPUs" and "current top performing image classifier only needed two days to train on 512 TPUs"); Urbina et al., *supra* note 80, at 189–191; Sandbrink, *supra* note 87; Matteucci et al., *supra* note 1.

[210] For instance, Hugging Face CEO Clem Delangue predicted that "in 2024, most companies will realize that smaller, cheaper, more specialized models make more sense for 99% of AI use-cases." Clem Delangue, LINKEDIN (Oct. 10, 2023), https://www.linkedin.com/posts/clementdelangue_my-prediction-in-2024-most-companies-will-activity-7117498531942146048-BIDD [https://perma.cc/4CPV-37DV]; *cf.* David Grangier et al., *Specialized Language Models with Cheap Inference from Limited Domain Data*, ARXIV (Oct. 31, 2024), https://doi.org/10.48550/arXiv.2402.01093 [https://perma.cc/B2PV-7XBS] (studying training small, specialized models with different budget considerations).

[211] *See* Toner & Fist, *supra* note 187; *see also* NEEL GUHA ET AL., THE AI REGULATORY ALIGNMENT PROBLEM 4 (2023), https://hai.stanford.edu/sites/default/files/2023-11/AI-Regulatory-Alignment.pdf [https://perma.cc/86L9-7RQK] (noting that "regulations based on threshold criteria may create incentives for strategic evasion [such as] developing multiple models below the compute threshold and combining their outputs"); Comment on Proposed Rule, *supra* note 1, at 7 (discussing "techniques inspired by ensembling, blending, mixture-of-experts, or switch transformers to string together models that individually fall below the compute threshold"); *cf.* Neel Guha et al., *AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing*, 92 GEO. WASH. L. REV. 1473, 1538 (2024) (discussing evasion generally).

[212] Lennart Heim & Janet Egan, Comment Letter on Proposed Rule to Implement Additional Export Controls 3, 9–10 (Dec. 15, 2023), https://cdn.governance.ai/Accessing_Controlled_AI_Chips_via_Infrastructure-as-a-Service.pdf [https://perma.cc/79ML-2SBK].

[213] *See* Toner & Fist, *supra* note 187.

[214] Anderljung et al., *supra* note 1, at 31 ("A regulatory regime for frontier AI could prove counterproductive if it incentivises AI companies to move their activities to jurisdictions with less onerous rules."); *see also* Brian Nussbaum, *Offshore: The Coming Global Archipelago of Corrosive AI*, LAWFARE (June 14, 2023), https://www.lawfaremedia.org/article/offshore-the-coming-global-archipelago-of-corrosive-ai [https://perma.cc/M732-4XRV].

of hardware and software.[215] By comparison, other metrics, such as algorithmic innovation and data, have been more difficult to track.[216] Whereas quantitative metrics like compute can be readily compared across different instances, the qualitative nature of many other metrics makes them more subject to interpretation and difficult to consistently measure. Compute usage can be measured internally with existing tools and systems; however, there is not yet an industry standard for measuring, auditing, and reporting the use of computational resources.[217] That said, there have been some efforts toward standardization of compute measurement.[218] In the absence of a standard, some have instead presented a common framework for calculating compute, based on information about the hardware used and training time.[219]

Fourth, compute can be estimated ahead of model development and deployment. Developers already estimate training compute with information about the model's architecture and amount of training data, as part of planning before training takes place. The EU AI Act recognizes this, noting that "training of general-purpose AI models takes considerable planning which includes the upfront allocation of compute resources and, therefore, providers of general-purpose AI models are able to know if their model would meet the threshold before the training is completed."[220] Since compute can be readily estimated before a training run, developers can plan a model with existing policies in mind and implement appropriate precautions during training, such as cybersecurity measures.

Fifth, the amount of compute used could be externally verified after training. While laws that use compute thresholds as a trigger for additional measures could depend on self-reporting, meaningful enforcement requires regulators to be aware of or at least able to verify the amount of compute being used. A regulatory threshold will be ineffective if regulators have no way of knowing whether a threshold has been reached. For this reason, some scholars have proposed that developers and compute providers be required to report the amount of compute used at different stages of the AI lifecycle.[221] Compute providers already employ chip-hours for client billing,

---

[215] Hooker, *supra* note 53, at 12; *see also* Sastry et al., *supra* note 155, 4, 27–28.

[216] Amodei & Hernandez, *supra* note 29 ("Algorithmic innovation and data are difficult to track, but compute is unusually quantifiable, providing an opportunity to measure one input to AI progress."); *see also* Hernandez & Brown, *supra* note 63.

[217] Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*, ARXIV 35–36 (Apr. 20, 2020), https://doi.org/10.48550/arXiv.2004.07213 [https://perma.cc/UJ7M-HQ64] ("The absence of standards for measuring the use of computational resources reduces the value of voluntary reporting and makes it harder to verify claims about the resources used in the AI development process."); Krystal Jackson et al., *Compute Accounting Principles Can Help Reduce AI Risks*, TECH POL'Y PRESS (Nov. 30, 2022), https://techpolicy.press/compute-accounting-principles-can-help-reduce-ai-risks/ [https://perma.cc/NE6W-QCU7]; Hooker, *supra* note 53, at 18 & app. A.

[218] Brundage et al., *supra* note 217, at 35–36 (highlighting the MLPerf benchmark suite, a working group at the Transaction Processing Performance Council, and a proposal that one or more AI labs voluntarily estimate the compute involved in a single project and report the method for wider adoption).

[219] *See generally Estimating Training Compute*, *supra* note 23; Amodei & Hernandez, *supra* note 29.

[220] EU AI Act, *supra* note 4, at Recital 112; Anderljung et al., *supra* note 1, at 36 & n.82 (noting that compute is largely determinable *ex ante* "from the planned specifications of the training run"); Koessler et al., *supra* note 114, at 3 ("Training compute is a very imperfect proxy for risk, but can easily be measured and forecasted early on in the development process").

[221] *See* Mulani & Whittlestone, *supra* note 71 (suggesting that developers share several compute-related metrics before, during, and after training and deployment, including the amount of compute used, the training time required, the quantity and variety chips used, a description of the networking of the compute infrastructure, and the physical location and provider of the compute); *see also* U.K., Department for Science, Innovation and Technology, *Emerging Processes for Frontier AI Safety* (Oct. 27, 2023), https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety [https://perma.cc/NZ6K-E6E5] (outlining potential safety practices and noting that model reporting and information sharing could provide

which could be used to calculate total computational operations,[222] and the centralization of a few key cloud providers could make monitoring and reporting requirements simpler to administer.[223] Others have proposed using "on-chip" or "hardware-enabled governance mechanisms" to verify claims about compute usage.[224]

Sixth, training compute is an indicator of developer resources and capacity to comply with regulatory requirements, as it represents a substantial financial investment.[225] For instance, Sam Altman reported that the development of GPT-4 cost "much more" than $100 million.[226] Researchers have estimated that Gemini Ultra cost $70 million to $290 million to develop.[227] A regulatory approach based on training compute thresholds can therefore be used to subject only the most resourced AI developers to increased regulatory scrutiny, while avoiding overburdening small companies, academics, and individuals. Over time, the cost of compute will most likely continue to fall, meaning the same thresholds will capture more developers and models. To ensure that the law remains appropriately scoped, compute thresholds can be complemented by additional metrics, such as the cost of compute or development. For example, the vetoed California Senate Bill 1047 was amended to include a compute cost threshold, defining a "covered model" to include one trained with over 1e26 OP, only if the cost of that training compute exceeded $100,000,000 at the start of training.[228]

At the time of writing, many consider compute thresholds to be the best option currently available for determining which AI models should be subject to regulation, although the limitations of this approach underscore the need for careful drafting and adaptive governance. When considering the legal obligations imposed, the specific compute threshold should correspond to the nature and extent of additional scrutiny and other requirements and reflect the fact that compute is only a proxy for, and not a precise measure of, risk.

"compute details (including the maximum the organisation plans to use, as well as information about its location and who provides it)" and "[e]xpected compute requirements for running the model during deployment"); O'Brien et al., *supra* note 136, app. I at 42–43 (suggesting that compute providers report "about certain aspects of development and deployment, such as AI compute usage per customer").

[222] Egan & Heim, *supra* note 1, at 19 ("Compute providers can easily access data related to total compute usage, such as the number of chip hours and the type of chip."); *see also* Heim et al., *supra* note 129 at 27–28.

[223] Heim et al., *supra* note 129, at 14, 20–21.

[224] *See* Onni Aarne et al., *Secure, Governable Chips*, CTR. FOR A NEW AM. SEC. (Jan. 2024), https://www.cnas.org/publications/reports/secure-governable-chips [https://perma.cc/3JKJ-3PHJ], at 7–10, 12 (describing chips able to "make a wide range of 'verifiable claims,' such as the amount of compute used to train an AI model"); Gabriel Kulp et al., *Hardware-Enabled Governance Mechanisms*, RAND (Jan. 18, 2024), https://doi.org/10.7249/WRA3056-1 [https://perma.cc/GE3P-UXNX], at viii (discussing hardware-enabled mechanisms as a complement to export controls); *see also* Lennart Heim, *Considerations and Limitations for AI Hardware-Enabled Mechanisms*, BLOG.HEIM.XYZ (Mar. 10, 2024), https://blog.heim.xyz/considerations-and-limitations-for-ai-hardware-enabled-mechanisms/ [https://perma.cc/CA9X-KZW3] (describing some limitations of hardware-enabled mechanisms); Shavit, *supra* note 74, at 6, 8–9 ("Ideally, chips could remotely report their logs, with on-chip firmware and remote attestation being sufficient to guarantee that those logs were truthfully reported.").

[225] *See supra* note 39 and accompanying text (collecting sources on the cost of training AI models).

[226] Massachusetts Institute of Technology & Imagination in Action, *Breakthrough Potential of AI*, YOUTUBE, at 6:36 (recorded Apr. 13, 2023), https://www.youtube.com/watch?v=T5cPoNwO7II [https://perma.cc/L6FU-MUZB].

[227] *Machine Learning Trends*, EPOCH (last updated Jan. 13, 2025), https://epochai.org/trends [https://perma.cc/8V3P-BTBY] (reporting a 90% confidence interval for the total amortized cost, including hardware, electricity, and staff compensation).

[228] S.B. 1047, 2023–2024 Reg. Sess. (Cal. 2024) § 3 (as enrolled, Sept. 3, 2024).

F. How Do Compute Thresholds Compare to Capability Evaluations?

A regulatory approach that uses a capabilities-based threshold or evaluation may seem more intuitively appealing and has been proposed by many.[229] There are currently two main types of capability evaluations: benchmarking and red-teaming.[230] In benchmarking, a model is tested on a specific dataset and receives a numerical score. In red-teaming, evaluators can use different approaches to identify vulnerabilities and flaws in a system, such as through prompt injection attacks to subvert safety guardrails. Model evaluations like these already serve as the basis for responsible scaling policies, which specify what protective measures an AI developer must implement in order to safely handle a given level of capabilities. Responsible scaling policies have been adopted by companies like Anthropic, OpenAI, and Google, and policymakers have also encouraged their development and practice.[231]

Capability evaluations can complement compute thresholds. For example, capability evaluations could be required for models exceeding a compute threshold that indicates that dangerous capabilities might exist. They could also be used as an alternative route to being covered by regulation. The EU AI Act adopts the latter approach, complementing the compute threshold with the possibility for the European Commission to "take individual decisions designating a general-purpose AI model as a general-purpose AI model with systemic risk if it is found that such model has capabilities or an impact equivalent to those captured by the set threshold."[232]

Nonetheless, there are several downsides to depending on capabilities alone. First, model capabilities are difficult to measure.[233] Benchmark results can be affected by factors other than capabilities, such as benchmark data being included during training[234] and model sensitivity to small changes in prompting.[235] Downstream capabilities of a model may also differ from those

---

[229] *See, e.g.*, Microsoft, *supra* note 1, at 14, 21; Bengio et al., *supra* note 72, at 844 (identifying the need for "policies that automatically trigger when AI hits certain capability milestones."); Anderljung et al., *supra* note 1, at 30 ("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."). For a discussion of capability thresholds and their relationship to risk, see generally Koessler et al., *Risk Thresholds for Frontier AI*, CTR. FOR THE GOVERNANCE OF AI (June 20, 2024), https://www.governance.ai/research-paper/risk-thresholds-for-frontier-ai [https://perma.cc/3RSR-USW4].

[230] For an overview of different methods, see *Challenges in Evaluating AI Systems*, ANTHROPIC (Oct. 4, 2023), https://www.anthropic.com/research/evaluating-ai-systems [https://perma.cc/FHC7-2QA8].

[231] *See generally Anthropic's Responsible Scaling Policy*, *supra* note 133; OpenAI, *Preparedness Framework* (Beta), *supra* note 144; *Google DeepMind's Frontier Safety Framework, Version 1.0*, *supra* note 144.

[232] *See* EU AI Act, *supra* note 4, at Recital 111, art. 51–52, Annex XIII (authorizing the Commission to designate general-purpose AI models with systemic risk considering other factors, including tools and benchmarks for assessing high-impact capabilities).

[233] *See, e.g.*, Anwar et al., *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*, ARXIV (Sep. 6, 2024), https://doi.org/10.48550/arXiv.2404.09932 [https://perma.cc/L77F-UK9W]; Elliot Jones et al., *Under the Radar? Examining the Evaluation of Foundation Models*, ADA LOVELACE INST. (July 25, 2024), https://www.adalovelaceinstitute.org/report/under-the-radar/ [https://perma.cc/SS7T-8BMY]; Anka Reuel et al., *Open Problems in Technical AI Governance*, ARXIV (July 20, 2024), https://doi.org/10.48550/arXiv.2407.14981 [https://perma.cc/XN8H-KNTS].

[234] *See* Kun Zhou et al., *Don't Make Your LLM an Evaluation Benchmark Cheater*, ARXIV (Nov. 3, 2023), https://doi.org/10.48550/arXiv.2311.01964 [https://perma.cc/Q4CB-PBPU] (discussing "benchmark leakage," in which test data or relevant data has been included in the pre-training corpus).

[235] *See, e.g.*, Abel Salinas & Fred Morstatter, *The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance*, ARXIV (Apr. 1, 2024), https://doi.org/10.48550/arXiv.2401.03729 [https://perma.cc/2FS7-74RD] (measuring the impact of prompt variation on LLMs' predictions and accuracy); Moran Mizrahi et al., *State of What Art? A Call for Multi-Prompt LLM Evaluation*, ARXIV (May 6, 2024), https://doi.org/10.48550/arXiv.2401.00595 [https://perma.cc/8WAM-EG37]; Melanie Sclar et al., *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design*, ARXIV (July 1, 2024), https://doi.org/10.48550/arXiv.2310.11324 [https://perma.cc/ULL6-YPVG].

during evaluation due to changes in dataset distribution.[236] Some threats, such as misuse of a model to develop a biological weapon, may be particularly difficult to evaluate due to the domain expertise required, the sensitivity of information related to national security, and the complexity of the task.[237] For dangerous capabilities such as deception and manipulation, the nature of the capability makes it difficult to assess,[238] although some evaluations have already been developed.[239] Furthermore, while evaluations can point to what capabilities do exist, it is far more difficult to prove that a model does not possess a given capability. Over time, new capabilities may even emerge and improve due to prompting techniques, tools, and other post-training enhancements.

Second, and compounding the issue, there is no standard method for evaluating model capabilities.[240] While benchmarks allow for comparison across models, there are competing benchmarks for similar capabilities; with none adopted as standard by developers or the research community, evaluators could select different benchmark tests entirely.[241] Red-teaming, while more in-depth and responsive to differences in models, is even less standardized and provides less comparable results. Similarly, no standard exists for when during the AI lifecycle a model is evaluated, even though fine-tuning and other post-training enhancements can have a significant impact on capabilities. Nevertheless, there have been some efforts toward standardization, including the U.S. National Institute of Standards and Technology beginning to develop guidelines and benchmarks for evaluating AI capabilities, including through red-teaming.[242]

---

[236] *See* Dario Amodei et al., *supra* note 72, at 16–20; Aleksandr Podkopaev & Aaditya Ramdas, *Tracking the Risk of a Deployed Model and Detecting Harmful Distribution Shifts*, ARXIV 2 (May 6, 2022), https://doi.org/10.48550/arXiv.2110.06177 [https://perma.cc/Y8EV-23F9] ("[A] model deployed in the real world inevitably encounters variability in the input distribution, a phenomenon referred to as *dataset shift*"); Carlos Mougan et al., *Explanation Shift: How Did the Distribution Shift Impact the Model?*, ARXIV 1 (Sept. 7, 2023), https://doi.org/10.48550/arXiv.2303.08081 [https://perma.cc/58LA-JR6D] ("As input data distributions evolve, the predictive performance of machine learning models tends to deteriorate"); Sean Kulinski & David I. Inouye, *Towards Explaining Distribution Shifts*, ARXIV 1 (June 20, 2023), https://doi.org/10.48550/arXiv.2210.10275 [https://perma.cc/2U35-R449].

[237] *See* U.S. AI SAFETY INST., MANAGING MISUSE RISK FOR DUAL-USE FOUNDATION MODELS (July 2024), https://doi.org/10.6028/NIST.AI.800-1.ipd [https://perma.cc/M89P-7TX5], at 2–3, 5–6; *Challenges in Evaluating AI Systems*, *supra* note 230; *Challenges in Red Teaming AI Systems*, ANTHROPIC (June 12, 2024), https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems [https://perma.cc/3CZX-473T].

[238] *See, e.g.*, Everett Thornton Smith et al., Comment Letter on NTIA AI Accountability Policy Request for Comment, CTR. FOR THE GOVERNANCE OF A.I. (June 12, 2023), https://cdn.governance.ai/GovAI_Response_to_the_NTIA_AI_Accountability_Policy_Request_for_Comment.pdf [https://perma.cc/5G6V-Q6L8]; *see also* Alaga & Schuett, *supra* note 133, at 4 ("We are aware of evaluations for power-seeking behavior and efforts to develop evaluations for deception, situational awareness, and manipulation. We are unaware of evaluations for other capabilities, such as the ability to exploit vulnerabilities in software systems or develop weapons.").

[239] *Cf.* Alexander Meinke, Bronson Schoen, Jérémy Scheurer et al., *Frontier Models Are Capable of In-Context Scheming*, APOLLO RESEARCH (Dec. 5, 2024), https://www.apolloresearch.ai/research/scheming-reasoning-evaluations [https://perma.cc/JM5B-BJKP]; Kristina Suchotzki & Matthias Gamer, *Detecting Deception With Artificial Intelligence: Promises and Perils*, 28 TRENDS COGNITIVE SCI. 481 (2024).

[240] STAN. INST. FOR HUMAN-CENTERED A.I., ARTIFICIAL INTELLIGENCE INDEX REPORT 2024 (2024), https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf [https://perma.cc/A2T6-Y3GZ], at 17.

[241] *See* Mostafa Dehghani et al., *The Benchmark Lottery*, ARXIV 30 (July 14, 2021), https://doi.org/10.48550/arXiv.2107.07002 [https://perma.cc/CA53-ENY8] (showing that the ranking of models can be drastically altered based on the choice of the subset of the benchmark considered, and introducing the notion of "benchmark lottery" to describe the fragility of the benchmarking process); *see also* Inioluwa Deborah Raji et al., *AI and the Everything in the Whole Wide World Benchmark*, ARXIV 7–9 (Nov. 26, 2021), https://doi.org/10.48550/arXiv.2111.15366 [https://perma.cc/STM6-DEYV]; Jones et al., *supra* note 233.

[242] U.S. National Institute of Standards and Technology, *Test, Evaluation & Red-Teaming*, https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence/test [https://perma.cc/3MFK-UWHZ]; U.S. National Institute of Standards and Technology, NIST AI 800-1, *Managing Misuse Risk for Dual-Use 4 Foundation Models* (July 2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf.

Third, it is much more difficult to externally verify model evaluations. Since evaluation methods are not standardized, different evaluators and methods may come to different conclusions, and even a small difference could determine whether a model falls within the scope of regulation. This makes external verification simultaneously more important and more challenging. In addition to the technical challenge of how to consistently verify model evaluations, there is also a practical challenge: certain methods, such as red-teaming and audits, depend on far greater access to a model and information about its development. Developers have been reluctant to grant permissive access,[243] which has contributed to numerous calls to mandate external evaluations.[244]

Fourth, model evaluations may be circumvented. For red-teaming and more comprehensive audits, evaluations for a given model may reasonably reach different conclusions, which allows room for an evaluator to deliberately shape results through their choice of methods and interpretation. Careful institutional design is needed to ensure that evaluations are robust to conflicts of interest, perverse incentives, and other limitations.[245] If known benchmarks are used to determine whether a model is subject to regulation, developers might train models to achieve specific scores without affecting capabilities, whether to improve performance on safety measures or strategically underperform on certain measures of dangerous capabilities.

Finally, capability evaluations entail more uncertainty and expense. Currently, the capabilities of a model can only reliably be determined *ex post*,[246] making it difficult for developers to predict whether it will fall within the scope of applicable law. More in-depth model evaluations such as red-teaming and audits are expensive and time-consuming, which may constrain small organizations, academics, and individuals.[247]

Capability evaluations can thus be viewed as a complementary tool for estimating model risk. While training compute makes an excellent initial threshold for regulatory oversight, as an objective and quantifiable measure that can be estimated prior to training and verified after, capabilities correspond more closely to risk. Capability evaluations provide more information and can be completed after fine-tuning and other post-training enhancements, but are more expensive, difficult to carry out, and less standardized. Both are important components of AI governance but

---

[243] *See generally* Stephen Casper et al., *Black-Box Access Is Insufficient for Rigorous AI Audits*, *in* FACCT '24: PROCS. 2024 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 2254 (June 2024).

[244] *See, e.g.*, *id.*; *Theories of Change for AI Auditing*, APOLLO RSCH. (Nov. 13, 2023), https://www.apolloresearch.ai/blog/theories-of-change-for-ai-auditing [https://perma.cc/W5MN-V744]; Lee Sharkey et al., *A Causal Framework for AI Regulation and Auditing*, APOLLO RSCH. 1 (2023) https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/65a6f1389754fc06cb9a7a14/1705439547455/auditing_framework_web.pdf [https://perma.cc/45WU-YFNF]; Anderljung et al., *supra* note 1, at 3, 23; Anderljung et al., *supra* note 132; Mökander et al., *supra* note 135; Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, *in* FACCT '20: PROCS. 2020 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 33 (2020).

[245] *See* Casper et al., *supra* note 243, at 2261–62, 2271–72, app. D; Anderljung et al., *supra* note 132, at 3–4; Raji et al., *supra* note 241, at 35; Sasha Costanza-Chock et al., *Who Audits the Auditors? Recommendations From a Field Scan of the Algorithmic Auditing Ecosystem*, *in* FACCT '22: PROCS. 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 1571 (2022); Victor Ojewale et al., *Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling*, ARXIV 8–10 (Mar. 14, 2024), https://doi.org/10.48550/arXiv.2402.17861 [https://perma.cc/9PCJ-VMFB].

[246] Anderljung et al., *supra* note 1, at 35 ("At present, there is no rigorous method for reliably determining, ex ante, whether a planned model will have broad and sufficiently dangerous capabilities.")

[247] Jones et al., *supra* note 233; Laura Galindo et al., *Open Loop US Program on Generative AI Risk Management: AI Red Teaming and Synthetic Content Risk*, META (2024), https://www.usprogram.openloop.org/site/assets/files/1/openloop_us_phase1_report_and_annex.pdf [https://perma.cc/X2H3-4FEB], at 22.

serve different roles.

IV. CONCLUSION

More powerful AI could bring transformative changes in society. It promises extraordinary opportunities and benefits across a wide range of sectors, with the potential to improve public health, make new scientific discoveries, improve productivity and living standards, and accelerate economic growth. However, the very same advanced capabilities could result in tremendous harms that are difficult to control or remedy after they have occurred. AI could fail in critical infrastructure, further concentrate wealth and increase inequality, or be misused for more effective disinformation, surveillance, cyberattacks, and development of chemical and biological weapons.

In order to prevent these potential harms, laws that govern AI must identify models that pose the greatest threat. The obvious answer would be to evaluate the dangerous capabilities of frontier models; however, state of the art model evaluations are subjective and unable to reliably predict downstream capabilities, and they can take place only after the model has been developed with a substantial investment.

This is where training compute thresholds come into play. Training compute can operate as an initial threshold for estimating the performance and capabilities of a model and, thus, the potential risk it poses. Despite its limitations, it may be the most effective option we have to identify potentially dangerous AI that warrants further scrutiny. However, compute thresholds alone are not sufficient. They must be used alongside other tools to mitigate and respond to risk, such as capability evaluations, post-market monitoring, and incident reporting. Further research avenues could develop better governance via compute thresholds:

1. What amount of training compute corresponds to future systems of concern? What threshold is appropriate for different regulatory targets, and how can we identify that threshold in advance? What are the downstream effects of different compute thresholds?

2. Are compute thresholds appropriate for different stages of the AI lifecycle? For example, could thresholds for compute used for post-training enhancements or during inference be used alongside a training compute threshold, given the ability to significantly improve capabilities at these stages?

3. Should domain-specific compute thresholds be established, and if so, to address which risks? If domain-specific compute thresholds are established, such as in President Biden's Executive Order 14,110, how can competent authorities determine if a system is domain-specific and verify the training data?

4. How should compute usage be reported, monitored, and audited?

5. How should a compute threshold be updated over time? What is the likelihood of future frontier systems being developed using less (or far less) compute than is used today? Does growth or slowdown in compute usage, hardware improvement, or algorithmic efficiency warrant an update, or should it correspond solely to an increase in capabilities? Relatedly, what kind of

framework would allow a regulatory agency to respond to developments effectively (e.g., with adequate information and the ability to update rapidly)?

6. How could a capabilities-based threshold complement or replace a compute threshold, and what would be necessary (e.g., improved model evaluations for dangerous capabilities and alignment)?

7. How should the law mitigate risks from AI systems that sit below the training compute threshold?