JOLT DIGEST
HARVARD JOURNAL OF LAW & TECHNOLOGY

Beyond FLOPs: Shortcomings of FLOPs as a Model Classification Metric in AI Regulation

By Sasha Rosenthal-Larrea and Lucille D. Finn - Edited by Shriya Srikanth

Sasha Rosenthal-Larrea is a partner in Cravath's Corporate Department, where she focuses her practice on advising clients on the most significant intellectual property issues, including with respect to complex licensing and collaborations, patent and copyright licensing strategy, software and artificial intelligence.

Lucille D. Finn is an associate in Cravath's Corporate Department.

Introduction

Regulators in the U.S. and around the world are walking a tightrope between removing roadblocks to artificial intelligence ("AI") innovation and protecting against the potential dangers posed by certain AI systems. Some regulators are focused on mitigating risks to health, safety and fundamental rights, among other systemic risks posed by AI, such as AI-designed biological weapons and AI systems capable of manipulating humans through subliminal techniques. While the Trump administration has made significant moves to remove barriers to AI innovation in the U.S. [1], the EU has continued on the track of comprehensive regulation through the EU AI Act [2], and U.S. state regulators are somewhere in between.

Regardless of the approach to regulation, a threshold question behind every AI policy is: which AI models call for regulation (if any), and which do not? Answering this question requires a deep understanding of the technologies and development processes behind AI, and we are seeing different formulations across jurisdictions and regulatory agencies. There are many ways to measure and differentiate AI models: model capability benchmarks (*i.e.*, performance on standardized AI model evaluations), revenue generation, number of users, deployment use cases, number of model parameters—and the metric of focus in this paper—the total number of floating point operations ("FLOPs"), a measurement of the total amount of computational effort (or "compute") used in the model's training process.

Several regulations with the potential to significantly impact AI development and use are defining categories of AI systems to be regulated by FLOPs. [3] However, these definitions might not be sustainable because new AI architectures are increasing training efficiency, potentially rendering FLOP thresholds less relevant to the regulators' normative goals. Frontier AI labs have trained sophisticated generative AI models using less compute than prior models. Training efficiency innovations will continue to accelerate this trend.

In this article, we will assess the relevance of FLOPs as a regulatory categorization tool and present potential shortcomings and consequences of relying on FLOPs as a metric of a model's sophistication or potential danger. Notably, if regulations apply only to certain models, it is critical that the criteria are clear—a multi-part qualitative test, for example, would create too much uncertainty and chill innovation. But even accepting the need for bright-line rules, careful consideration should be given to the metrics selected to draw these lines. A solid understanding of the implications of a chosen metric is necessary to ensure that it serves the particular regulatory aims and to evaluate how regulations based on that metric may in turn affect technological development.

1. What Are FLOPs and What Do They Measure?

A FLOP is a basic mathematical calculation between two numbers (*e.g.*, addition, multiplication). AI model training and inference occur through a series of FLOPs performed across matrices of billions to trillions of model parameters, and the total number of FLOPs measures the amount of computational work used to train an AI model. The transformer neural network architecture (a popular model architecture used in generative AI), is notoriously FLOP-intensive. [4] Many of the most sophisticated transformer-based models, such as OpenAI's GPT, are created through training processes requiring vast quantities of FLOPs.

2. What Influences the Number of FLOPs?

Major variables that influence the number of FLOPs include: (1) the size of training data sets, (2) the type of training data, (3) the model architecture, (4) the training methodology and (5) the size of the model. All else being equal, a model that is trained on more data will have a higher total number of FLOPs, since every additional piece of training data entails additional calculations during the training process. The type of training data also matters. While the training process for text-generating large language models ("LLMs") already involves running each piece of text-based training data through billions or trillions of mathematical calculations, *i.e.*, FLOPs [5], the number of required calculations for multi-modal models (*i.e.*, AI models designed to process multiple types of training data such as text, images and audio) can be even higher. Moreover, all else being equal, a model that has more parameters will have a higher total number of FLOPs. Model architecture and training methodologies can also significantly reduce or increase the total number of FLOPs used in training.

Some regulators have landed on FLOPs as the metric to use to set a threshold for which models will be regulated, because they deem the number of calculations involved in training to be an approximation of how advanced or sophisticated the model is (*i.e.*, a model's ability to be applied flexibly across different contexts to perform a wider range of tasks). Regulators reason that these generalized capabilities may lead to malfunctions or misuse—such as AI-enabled hacking, biologic attacks and loss of control—making a model capable of posing catastrophic risks. [6]

FLOPs, however, are merely an *indirect* way of *approximating* sophistication. FLOPs measure only one variable of the model development process: computing power. Other factors, such as the diversity or quality of the training data and the methods used in the training process, can significantly impact training efficiency while lowering the number of total FLOPs. [7] This can be observed, for example, in distillation—a technique that involves training a model on the outputs of another model—which can enable training powerful capabilities in a fraction of the FLOPs. Advancements and innovations in AI architectures are also increasing compute efficiency and decreasing the number of FLOPs required to achieve similar model sophistication. In other words, models trained using fewer FLOPs can pose similar risks to models trained using more FLOPs.

Consider a few examples of more efficient model architectures from the last year. Several labs have adopted Mixture of Experts ("MoE") transformer architectures, which dynamically route inputs to a subset of smaller "expert" models rather than a single larger model, and claim to have significantly lower compute costs than dense (non-MoE) models. As another example, Structured State Space Models ("SSMs") perform calculations on combined "states" rather than on each token individually. [8] More recently, Hierarchical Reasoning Models ("HRMs"), an architecture that employs a new method of "latent reasoning",

a process of performing reasoning steps in its internal hidden state space (without reliance on natural language or other observable chain-of-thought outputs), have demonstrated promising performance on complex reasoning tasks with a lower number of model parameters. [9] Researchers at Google DeepMind created “Griffin”, which has demonstrated similar performance to transformers despite being trained on over six times fewer tokens. Its efficiency comes partly from the model basing its output on only a subset of preceding tokens, rather than all preceding tokens. [10] In each of these cases, the total number of calculations in training may be reduced significantly as compared to the standard transformer models that existed at the time of drafting of recent regulations.

These new architectures are likely only the beginning. The machine learning field is rapidly innovating, and it is highly likely that more efficient models trained using fewer FLOPs will compete with, or even supplant, the current leading models trained using more FLOPs, rendering FLOPs thresholds in existing regulations obsolete.

3. Where Do Regulations Stand on Using FLOPs To Classify AI Systems?

Currently, some key regulations use FLOPs as a proxy for the models’ capabilities, and use FLOPs to classify the AI systems subject to oversight. Such regulations include the European Union’s Artificial Intelligence Act (the “E.U. AI Act”) [11], U.S. federal regulations, such as the U.S. Treasury Department rules on outbound investment (the “Treasury Rule”) [12], and U.S. state AI regulations, such as California’s newly effective Transparency in Frontier Artificial Intelligence Act (the “TFAIA”) [13] and New York’s recently enacted Responsible AI Safety and Education Act (the “RAISE Act”). While both the E.U. AI Act and U.S. regulators have adopted FLOPs-based thresholds (using FLOPs as a proxy for model capabilities), the two jurisdictions have adopted different thresholds without elaborating on the rationale behind those thresholds. U.S. state legislators in New York have continued to propose FLOPs thresholds, while other U.S. state regulators have expressed skepticism about FLOPs-based thresholds altogether.

A. E.U. AI Act

The E.U. AI Act is currently the most comprehensive and prescriptive AI regulation to date. The sections of the Act governing “general-purpose AI models that pose systemic risks” took effect in August 2025. [14] A general-purpose AI model falls within the scope of the E.U. AI Act’s significant obligations and penalties if the cumulative amount of computation used for its training is greater than 10^25 FLOPs. [15] The European Commission has explained that it used FLOPs as a metric because “the capabilities of the models above this threshold are not yet well enough understood”, and those models “could pose systemic risks.” [16] Notably, the E.U. AI Act itself acknowledges that FLOPs are not a perfect measure of risk—it expressly allows developers to present arguments that the model does not present systemic risks [17], and the Commission may identify additional models below the threshold to be subject to the regulation if they have high impact capabilities that present systemic risks. [18]

B. U.S. Federal Regulation

The U.S. Department of the Treasury’s rule on outbound investments in Chinese companies active in developing AI and other national security-related technologies, which went into effect on January 2, 2025, uses FLOPs to classify prohibited transactions and transactions that require notification. Specifically, the Treasury Rule prohibits certain transactions involving AI systems trained using greater than (i) 10^25 FLOPs or (ii) 10^24 FLOPs using primarily biological sequence data. Notification to the U.S. Department of the Treasury is required for certain transactions involving AI systems that are trained using greater than 10^23 FLOPs. [19]

Although rescinded by the Trump administration [20], an earlier executive order issued under the Biden administration in 2023 (the “2023 AI Executive Order”) also imposed regulatory requirements for certain types of AI models based on a general threshold of 10^26 FLOPs (and a lower threshold of 10^23 FLOPs for models using primarily biological sequence data). [21] The FLOPs thresholds set in the 2023 AI Executive Order were intended to be temporary and subject to further consultation and revision of technical conditions.

C. U.S. State Regulation

Regulators continue to develop AI guardrails at the state level, but—perhaps due to more recent awareness of the implications of efficiency gains in the training process—states are split in their use of FLOPs thresholds to classify AI models. While many early state-level AI regulations used risk-based or usage-based classifications rather than bright-line FLOPs thresholds, recent legislation from California and New York focused on frontier AI model safety each rely on FLOPs.

In January 2026, California's landmark AI safety law, the TFAIA, entered into force. [22] The TFAIA relies on a FLOPs threshold to define "frontier models" within the scope of the act as AI models (i) trained on a broad data set, (ii) designed for generality of output, (iii) adaptable to a wide range of distinctive tasks and (iv) trained using greater than 10^26 FLOPs. [23] The TFAIA will impose safety and transparency requirements on developers of frontier models, and those that have annual gross revenue of over $500 million in the preceding year will be subject to additional, enhanced transparency obligations and governance requirements (including publishing and maintaining a company-wide "frontier AI framework" and transparency reports containing summaries of catastrophic risk assessments). [24]

The 10^26 FLOPs threshold in the TFAIA remains contentious. Anthropic stated in its endorsement of the TFAIA that 10^26 FLOPs is "an acceptable starting point", but noted "there is always a risk that some powerful models may not be covered." [25] Others have described the FLOPs threshold in the TFAIA as "an arbitrary proxy for capability and risk", in view of the fact that "some smaller-scale models could still pose serious real-world risk, while some large-scale models may present very low risk." [26] Similar to the E.U. AI Act, the TFAIA acknowledges that frontier model definitions may change, stating that "foundation models developed by smaller companies or that are behind the frontier may pose significant catastrophic risk, and additional legislation may be needed at that time." [27] The TFAIA requires the California Department of Technology to assess the "frontier model" definition to ensure that it accurately reflects "technological developments, scientific literature, and widely accepted national and international standards" on or before January 1, 2027. [28] This open-endedness harkens back to the legislative history of the prior California AI bill SB 1047 that was vetoed by California Governor Gavin Newsom exactly one year prior. Newsom noted in his veto statement that smaller, specialized models could pose similar or greater risks, such that the focus on large-scale models defined by FLOPs had the potential of "curtailing the very innovation that fuels advancement in favor of the public good." [29]

New York's RAISE Act, which was signed into law in December 2025 and will become effective in January 2027, similarly imposes safety and reporting obligations on developers of frontier models based on a FLOPs threshold. The RAISE Act defines "frontier models" using a combination of FLOPs and dollar cost of training, covering AI models that are trained using (i) greater than 10^26 FLOPs and (ii) a compute cost exceeding $100 million, or models trained through *distillation* of models meeting (i) and (ii). [30] Notably, the RAISE Act's inclusion of distillation (which was not addressed in California's TFAIA) signals the regulatory environment playing catch-up to emergent AI training innovations. A similar bill proposed in Arizona (HB 4098) includes the same FLOPs, cost, and distillation criteria.

In parallel, Colorado's Consumer Protections for Artificial Intelligence ("SB24-205"), which became effective on February 1, 2026, provides a non-FLOPs alternative for classifying AI systems. [31] SB24-205 defines a covered "high-risk" AI system as one that makes or is a substantial factor in making "a consequential decision", which is limited to decisions with a material impact on a few specific areas: education, employment, financial or lending services, essential government services, healthcare, housing, insurance or legal services. [32] Other states have enacted similar laws that focus on regulating uses, and avoid FLOPs-based definitions, such as in Utah (focused on "high-risk interaction") and Texas (imposing transparency duties for government and healthcare uses, and prohibitions on specific categories of use). Virginia legislators drafted a similar AI bill, HB 2094, that focused on "high-risk" systems making "consequential decision[s]" [33], though the bill was ultimately vetoed by the governor due to "economic growth" considerations. [34]

4. Looking Ahead: FLOPs and the Effect on Innovation

If regulators aim to measure a model's potential risks and enhance the safety and security of AI development, FLOPs are not quite on point. FLOPs are a technical measure of one aspect of a model's training process. They do not address issues such as bias, misuse or potential for causing harm. Regulations using FLOPs to classify AI models will always be both over- and under-inclusive; some high-risk models may fall under the FLOPs threshold, while some low-risk models may inadvertently exceed the FLOPs threshold and be subject to burdensome reporting and other requirements.

Another important unintended consequence of framing regulations in terms of FLOPs is that doing so may arbitrarily steer developers in the direction of innovating away from model types that are compute-heavy (such as transformers) in favor of other more FLOP-efficient models. There may be public policy reasons for wanting to steer developers toward compute-lite models (e.g., energy conservation, promotion of competition), but those policy objectives should be addressed head-on, rather than having it be an unintended consequence of poorly targeted safety regulations.

One could also see a scenario where developers choose to train transformer-based models only up to the FLOPs threshold in order to avoid regulation, leaving them even more error-prone, biased and harmful—issues that could be remediated by more rigorous model training and fine-tuning with an increased number of total FLOPs. In turn, these regulations may perpetuate the exact issues that they are trying to mitigate.

As regulators and innovators continue to navigate policy-making and compliance, increased awareness of the limitations and merits of FLOPs as a model classification metric is critical to shaping the future of this industry.

---

[1] David J. Kappos, Evan Norris & Sasha Rosenthal-Larrea, *Early Trump AI Moves Come in a Complex Regulatory Landscape*, Bloomberg Law (Feb. 20, 2025), https://www.cravath.com/a/web/2F81SLjrcBc399PMRCab8A/a47Kmw/bloomberg-law-early-trump-ai-moves-come-in-a-complex-regulatory-landscape-22025.pdf.

[2] David J. Kappos, et al., *EU AI Act to Enter Into Force*, Cravath (Jul. 2024), https://www.cravath.com/a/web/4ScwYSAHgxi49JrkYJQC6a/9hPcjv/1199267-eu-ai-act-briefing-v20.pdf.

[3] *See, e.g.*, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689) 1; Transparency in Frontier Artificial Intelligence Act (TFAIA), S.B. 53, 2025–2026 Reg. Sess. (Cal. 2025); Responsible AI Safety and Education (RAISE) Act, S. 6953-B, 2025–2026 Reg. Sess. (N.Y. 2025).

[4] David J. Kappos, *et al., How ChatGPT Understands Context: The Power of Self-Attention*, Cravath (Feb. 2024), https://www.cravath.com/a/web/25fvkMDn6Q8MyAtaPpsLf2/8BaHMZ/cravath-tech-explainers-how-chatgpt-understands-context-022024.pdf.

[5] Tom B. Brown, *et al., Language Models are Few Shot Learners*, Arxiv (Jul. 20, 2020), https://arxiv.org/pdf/2005.14165.

[6] TFAIA §1(j).

[7] Sally Beatty, *Tiny But Mighty: The Phi-3 Small Language Models with Big Potential*, Microsoft (Apr. 23, 2024), https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/.

[8] Albert Gu & Tri Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, Arxiv (May 31, 2024), https://arxiv.org/abs/2312.00752.

[9] Guan Wang, *et. al., Hierarchical Reasoning Model*, Arxiv (Aug. 4, 2025), https://arxiv.org/pdf/2506.21734.

[10] Soham De, *et al., Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models*, Arxiv (Feb. 29, 2024), https://arxiv.org/abs/2402.19427.

[11] 2024 O.J. (L 1689) 1.

[12] Provisions Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern, 89 Fed. Reg. 90398, (Nov. 15, 2024) (codified at 31 C.F.R. pt. 850).

[13] TFAIA.

[14] 2024 O.J. (L 1689) 26.

[15] *Id.* at 83.

[16] European Commission, *AI Act Service Desk–Frequently Asked Questions* (last visited Mar. 12, 2026), https://ai-act-service-desk.ec.europa.eu/en/faq.

[17] 2024 O.J. (L 1689) 26. Note that the initial draft of the Act in 2021 did not include such a definition, which was added by the European Parliament in March 2024.

[18] *Id.*

[19] Provisions Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern, 89 Fed. Reg. 90398, (Nov. 15, 2024).

[20] Exec. Order No. 14,148, 90 Fed. Reg. 8237 (Jan. 28, 2025).

[21] Exec. Order No. 14,110 §4.2(b), 88 Fed. Reg. 75191 (Nov. 1, 2023).

[22] TFAIA.

[23] TFAIA §2(f),(i).

[24] TFAIA §2(j).

[25] *Anthropic is Endorsing SB 53,* Anthropic (Sep. 8, 2025), https://www.anthropic.com/news/anthropic-is-endorsing-sb-53.

[26] Aden Hizkias, *Why California's SB 53 Still Gets AI Regulation Wrong*, Chamber of Progress (Jul. 9, 2025), https://progresschamber.org/insights/why-californias-sb-53-still-gets-ai-regulation-wrong/.

[27] TFAIA §1(n).

[28] TFAIA § 227575.14(a).

[29] Gavin Newsom, *Veto Message*, Office of the Governor, (Sep. 29, 2024), https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf.

[30] RAISE Act §6(a).

[31] Concerning Consumer Protections In Interactions with Artificial Intelligence Systems., S.B. 24-205, 74th Gen. Assemb., 2d Reg. Sess. (Colo. 2024).

[32] *Id.* § 6-1-1701 (3).

[33] High-risk artificial intelligence; definitions, development, deployment, and use, civil penalties., H.B. 2094, 2025 Gen. Assemb., Reg. Sess. (Va. 2025).

[34] *Governor's Veto* (2025), https://lis.virginia.gov/bill-details/20251/HB2094/text/HB2094VG.