
The Feasibility of a Hardwired Pause of Frontier AI Training

Working Group on AI Pause Feasibility[†]

Organizers:

William Fithian

Associate Professor of Statistics, University of California, Berkeley

Wesley H. Holliday

Professor of Philosophy, University of California, Berkeley

Members:

Kayla Blomquist

Executive Director, Oxford China Policy Lab

Gregory Conti

Associate Professor of Politics, Princeton University

Members continued on next page

[†] Each of the listed co-authors contributed ideas and/or text to the paper. However, co-authorship does not imply agreement with every claim in the paper. Nor does it imply endorsement by a co-author's institution. Institutional affiliations are listed for identification purposes only.

For helpful discussions or feedback about this research, we thank Vinod Aggarwal, Peter Barnett, Daniel Björkegren, Jim Demmel, M. Taylor Fravel, Bethany Goldblum, Lucy Horowitz, Dominic Hughes, Adam Khoja, Nic Lane, Karan Makkar, Oriana Skylar Mastro, Emma Pierson, Aaron Scher, Sophia Shao, Sarah Shoker, Robert Trager, and Jasmine Wang.

Barry Eichengreen

George C. Pardee and Helen N. Pardee Chair and Distinguished Professor of Economics and Political Science, University of California, Berkeley

Rose Gottemoeller

Senior Research Scholar, Center for International Security and Cooperation, Freeman Spogli Institute for International Studies, Stanford University; Research Fellow, Hoover Institution

Kelly M. Greenhill

Associate Professor of Political Science, International Relations and Civic Studies, Tufts University; Visiting Associate Professor, Senior Research Fellow, and Seminar XXI Director, MIT; AI and Society Fellow, Center for AI Safety

Ming-Yen Ho

PhD Candidate, Haas School of Business, University of California, Berkeley

Thomas Icard

C. I. Lewis Professor of Philosophy and Computer Science (by courtesy), Stanford University

Shachar Kariv

Benjamin N. Ward Professor of Economics, University of California, Berkeley

Daniel Kroth

Research Scholar, Berkeley Risk and Security Lab, University of California, Berkeley

David Krueger

Assistant Professor in Robust, Reasoning, and Responsible AI at the University of Montreal / Mila

Gretchen Krueger

Independent Researcher Affiliated with the Berkman Klein Center, Harvard University

Erik Leklem

AI Governance Research Fellow, MATS Research

Adam Lesnikowski

Doctoral Researcher in Logic and the Methodology of Science, University of California, Berkeley

Xiaobo Lü

Associate Professor of Political Science, University of California, Berkeley

Dennis Murphy

PhD Candidate in International Affairs, Science and Technology, Georgia Institute of Technology; Predoctoral Fellow at the Belfer Center for Science and International Affairs at the Harvard Kennedy School, Harvard University

Members continued on next page

Lindsay Rand

Research Scholar, Berkeley Risk and Security Lab, University of California, Berkeley

Andrew Reddie

Associate Research Professor of Public Policy, University of California, Berkeley

Stuart Russell

Distinguished Professor of Computer Science, University of California, Berkeley

Daniel Sargent

Professor of History and Public Policy, University of California, Berkeley

Kirsten Ann Schulz

Former American Diplomat, U.S. State Department

Sanjit Seshia

Cadence Founders Chair Professor of Electrical Engineering and Computer Sciences, University of California, Berkeley

Anita Srinivasan

Research Fellow, MATS Research

Rory Truex

Associate Professor of Politics, Princeton University

Carah Ong Whaley

Lecturer in Politics, University of Virginia

Abstract

There is increasing interest in an international pause of frontier AI training that is mutually verifiable by participating states. This report studies the feasibility of a *hardwired pause*, a type of pause in which a prohibition on frontier training is mutually verified through monitoring of and constraints on AI hardware. Like other plans for a pause, it involves monitoring AI chips that are capable of frontier AI training. But to support a durable pause lasting a decade or longer, the hardwired pause goes further in two ways. First, it stops the production of training-capable chips during the pause. Second, over time it phases out the pre-pause stock of training-capable chips, which constitute “dry tinder” for a resumed AI race. It replaces training-capable chips with *inference-only* chips, which can run approved AI models for inference, potentially at much higher speeds and with much greater energy efficiency, but cannot practically train new frontier AI models, due to the way the chips are constructed. If successful, this replacement—together with other mitigations against harms from deployed AI systems—would allow the diffusion of inference on existing models while avoiding the additional risks to economic and social well-being and to national security posed by the training of increasingly powerful models. We assess that a hardwired pause would be feasible to implement now or in the near future, if world leaders wanted to pause, and that its durability would depend on geopolitical and technical conditions that we analyze in the report.

Executive summary

The coordination problem

The risks associated with rapid advances in AI capabilities have led to calls for a pause of frontier AI training. Yet companies and governments may be unwilling to *unilaterally* pause for extended periods of time, for strategic reasons: they may fear that other companies or governments will not pause and that those who do pause may then suffer a commercial or geopolitical disadvantage. Thus, companies and governments face a coordination problem in pausing frontier AI training.

However, such problems are not insoluble. Coordination is possible if the decision makers care enough about future, not just immediate, payoffs, and if monitoring is adequate to support credible consequences for defections. Humanity has faced similar coordination problems before and responded with international agreements, as with efforts to control nuclear technology or address damage to Earth's ozone layer. Whether humanity can coordinate to regulate the unique, dual-use dynamics of AI has become one of the most significant questions of the 21st century.

This report studies the *feasibility* of an international agreement to pause frontier AI training, where the agreement is designed to overcome the coordination problem for companies and governments. It does not study the *desirability* of a pause relative to other policy options or the *probability* that a pause will be implemented.

We consider a scenario in which world leaders would prefer pausing to not pausing *if* they could be confident that others were also pausing or that defections from a pause would be detected within a reasonable timeframe. In such a *pause-willing* scenario, is a pause for at least a decade feasible?

A hardwired pause

Frontier AI training requires AI chips, which states can govern. The supply chain for AI chips is highly concentrated with several chokepoints, bolstering governments' abilities to enforce rules on the production of such chips. In addition, bypassing existing supply chains is extraordinarily difficult. An international prohibition on frontier AI training can take advantage of these facts about the hardware needed for AI training.

We focus on the feasibility of what we call a *hardwired pause*. A hardwired pause would stop the production of *training-capable* AI chips, which are needed to develop more powerful AI models. Over time a hardwired pause would also phase out most or all of the pre-pause stock of such chips. It would replace them with

inference-only AI chips, which are capable of fast, energy-efficient AI inference on approved models but are not practically useful for training new frontier AI models, even if stolen or seized, due to the way the chips are constructed. During the process of replacement, training-capable AI chips would provide *transitional inference service*, with verification measures preventing their use for frontier AI training. If successful, this replacement—together with other mitigations against harms from deployed AI systems—would allow states to reap the economic benefits of diffusing AI inference across their economies without continuing an AI training race that may threaten to overwhelm states with disruptions to their social and economic stability or to their national security.

Technologies needed for inference-only AI chips are already being developed for purely economic reasons, as firms develop model-specific inference chips that serve AI inference much more efficiently than general-purpose chips do. The training capacity of these model-specific inference chips is much lower than that of general-purpose chips and could be made even lower through intentional design. Currently, the economic attractiveness of model-specific chips is limited by the rapid turnover of new models. However, by prohibiting the training of new frontier models, a hardwired pause would increase the economic advantages of these specialized chips and spur innovation in the development of other inference-only chips.

The political palatability of a hardwired pause could be enhanced by its ability to keep delivering key affordances that training-capable chips currently provide, such as commercial inference service on existing models, beneficial scientific and medical research, and strategic insurance against being left behind by defecting states. We discuss ways it might do so, for example by constructing internationally governed *scientific preserves* where training-capable chips could be used for declared, transparent research workloads, or by permitting states to hold declared and monitored stockpiles of powered-off chips under seal. Another beneficial palatability feature would be the new chips' lower resource demands, both reducing their environmental impact and ameliorating domestic frictions in the U.S. between data center operators and local communities.

Short-term feasibility of a hardwired pause

Because training new models requires large numbers of training-capable chips, the feasibility of verification is an inherently quantitative question: while it is next to impossible to detect an evader who is using a single computer to train a small AI model, it is far easier to detect an evader who is training a frontier-scale AI model using hundreds of thousands or millions of chips.

Over the last half decade, the computing budgets dedicated to frontier-scale AI research and training have grown at a rapid exponential rate, rising by an order of magnitude every 18 months and already straining the available stock of AI

chips, which has also grown exponentially. Consequently, the number of AI chips required to continue that frontier scaling trajectory even into 2028 or 2029 far outstrips the current stock, which itself dwarfs the stock that existed only a few years ago. To research and train a new model that would significantly advance the frontier after a hardwired pause, absent an unforeseen methodological breakthrough, an evader would need to marshal a significant fraction of the world's total computing capacity for months or years at a time—while pretending to use the same chips for compute-intensive inference service.

As a result, we assess that implementing reasonably effective verification methods would likely be more than sufficient in the short term to guard against *covert evasion*: secretly training new frontier models while ostensibly adhering to verification procedures. More sophisticated verification methods under development could be implemented as they became available, for added security.

However, verification methods could not by themselves guard against the risk of *overt breakout*: a state amassing a very large fleet of chips, openly abrogating the pause agreement, and resuming the AI race at a faster pace than before, in an attempt to build and deploy a strategically decisive model before other states could respond.

Because a breakout fleet would likely need to be larger than today's total world stock of AI chips, overt breakout would not be an immediate threat. But if the stock of AI chips were allowed to continue growing exponentially during a frontier pause, it would fill the world with destabilizing “dry tinder” that could fuel a resumed race. A hardwired pause would defuse this risk by preventing the accumulation of AI *training capacity*, while permitting a continued rise in *inference capacity* on approved models.

Durability of a hardwired pause

For a frontier pause to last for a decade or more, it would need to be robust against the possibility of erosion via gains in *training efficiency*: advances in algorithms and data that have made each unit of computing capacity grow more effective at AI training over time. If these advances continue, they will make it possible to train more capable models with today's level of training capacity. Thus, a lasting pause could eventually require a reduction over time of the AI training capacity available for covert evasion or overt breakout, at a rate that keeps pace with gains in training efficiency.

To durably guard against covert evasion, states could collaborate to conduct a *global census* of pre-pause training-capable chips, both to minimize the total amount of undeclared training capacity and to estimate the distribution of unaccounted-for chips. This census would be supported by commercial incentives, as declaration of training-capable chips would be a precondition of using them

for transitional inference service. Because AI chips are expensive capital goods that generate a variety of documentary evidence, we assess that it would be possible for a well-conducted census to locate the vast majority of pre-pause chips in agreement states.

To durably guard against overt breakout, states could reduce the amount of total training capacity within their borders, for example by moving most of their pre-pause chips offshore or transferring them to the custody of neutral states or scientific preserves, with the understanding that if they defected other states would prevent them from recovering that training capacity. Once inference-only chips became available, states would be able to reduce their breakout capacity without risking their inference capacity, which could remain within their own territory.

Quantitative assessment

While many states would play important roles in establishing and maintaining a hardwired pause, we focus on the key strategic dynamic between the U.S. and China, the two states that are currently home to the leading AI developers.

We examine various potential sources of post-pause training capacity and estimate the threat posed by each source, for both covert evasion and overt breakout. We assess that the most important potential source of covert evasion risk would be undeclared pre-pause AI chips controlled by either state, and the most important source of overt breakout risk would be the pre-pause stock of training-capable chips retained commercially for transitional inference service. Thus, the most important political determinants of durability would be these states' eventual success in completing the global census and their eventual willingness to retire or transfer most of their declared pre-pause stock outside their territory. Phasing out the pre-pause stock without disrupting commercial inference service would in turn require the timely production of inference-only chips.

Provided the two challenges identified above could be met before training efficiency advanced by one order of magnitude (meaning that present-day models could be trained with a tenth as much compute), we assess that a hardwired pause would likely be robust to a second order of magnitude of training efficiency gains.

It follows that a hardwired pause, if implemented in the near future, could likely endure for at least a decade, provided that efficiency gains for training models at or near the paused frontier scale did not accrue faster than one order of magnitude every five years (for reference, this rate would approximately match the rate of progress associated with the famous 18-month doubling time of Moore's Law). Empirical evidence from controlled experiments places the historical rate of efficiency gains at fixed training scale roughly at or below that

rate, with more dramatic historical gains unlocked by two breakthroughs (the transformer architecture and Chinchilla balancing) that were complementary to rapid scaling. Because scaling would stop after a pause, as would researchers' ability to test ideas at frontier scale, post-pause gains could be slower relative to currently prevailing AI race conditions.

Economic implications

An important component of the feasibility of a hardwired pause is its economic palatability. It has been claimed that a pause of frontier AI training would cause serious problems for the economy. In our assessment, a hardwired pause need not do so. The continuing diffusion of existing models could sustain significant AI-related spending and productivity gains even with a pause on frontier training. If the valuations of AI companies decline, the macroeconomic effects could be cushioned by monetary policy. There are concerning signs in the debt financing of the data center build-out, which regulators will need to address regardless of other policy choices. But provided that inference demand sustains data center revenues during a hardwired pause, the pause would not substantially increase the risk of serious financial problems. A hardwired pause could also bring economic benefits by incentivizing companies to further reduce inference costs, leading to savings that are at least partially passed on to consumers. Scaling up inference-only hardware would require upfront development costs, but for certain types of inference-only chips such costs could be more than offset by lower lifetime inference costs.

Remaining governance challenges

A hardwired pause is not a panacea for AI governance. Misuse, malfunction, and societal risks of allowed models would remain issues for governance. Nonetheless, a successful AI training pause could mitigate some of the most severe risks posed by more powerful AI models, including misuse, loss of control, disempowerment, and concentration of power. This includes the risk of frontier AI enabling ubiquitous surveillance—or being used to justify such surveillance as necessary to prevent catastrophic misuse. A hardwired pause also has costs and limitations. It would create new requirements for monitoring hardware production and deployment. Depending on its implementation, it could entrench the market positions of leading companies, requiring regulation of AI developers as government-created oligopolies. On the technical front, under a prohibition of large-scale AI training, security vulnerabilities in models would have to be dealt with in ways other than large re-training runs. Outdated knowledge cutoffs could entrench values and ideological or other biases in models; on the other hand, they could also mitigate risks of loss of control or human disempowerment by ensuring that humans remain more situationally aware than AI systems and by reducing human overreliance on AI systems for moral guidance. Finally, even if a hardwired pause

successfully prevented further frontier AI training, it would not address all harms and risks related to current and future AI. Addressing these would require other measures to be adopted, but a pause could at least complement these measures by giving governments and societies more time to do so.

Why states might agree to a pause

The pause willingness assumed in our feasibility analysis could come about in multiple ways. For example, states may eventually judge that the costs of trying to avoid—or of incurring—disruptions or disasters from increasingly capable AI outweigh the perceived strategic benefits of continuing a mutual AI race. Or they may find that the growing cost of advancing the frontier is not worth maintaining a few months’ lead in capabilities over fast followers. If pause willingness were to emerge, the feasibility analysis of the hardwired pause in this report could inform decision making about the type of pause to implement.

Summary: feasibility of a hardwired pause

We assess a durable hardwired pause of a decade or more to be feasible if (1) states cooperate to govern the manufacture of AI chips, (2) states share information to locate pre-pause chips, (3) inference-only chips are developed and manufactured in time to allow for training-capable chips to be phased out, and (4) the rate of training efficiency gains after a pause remains at or below the pre-pause rate.

Contents

1	Introduction	13
1.1	A Hardwired Pause	14
1.2	Scope	19
1.3	Organization	20
2	Background: arguments for a pause and lessons for AI governance	21
2.1	Observed arguments for a pause	21
2.2	Historical lessons for AI governance	25
2.3	Game-theoretic lessons for AI governance	30
3	Example implementation of a hardwired pause	32
3.1	Overview and objectives	32
3.2	Operationalizing technical requirements	38
3.3	Hardware governance	47
3.4	Transition governance	55
3.5	Possible exit conditions	58
3.6	Institutional governance	59
4	Feasibility of inference-only hardware	62
4.1	Quantifying training capability	62
4.2	Practical inference-only accelerators	63
4.3	Model-specific integrated circuits (MSICs)	65
4.4	Time required to ramp inference-only hardware production . . .	67
5	Feasibility of hardware governance	71
5.1	Overview	71
5.2	The supply chain for accelerator chips	73
5.3	Feasibility of governing chip manufacturing	76
5.4	Feasibility of a global accelerator census	78
5.5	Estimating how many chips a census would miss	84
6	Feasibility of verification	90
6.1	Summary assessment	90
6.2	Fleet thresholds for defection	92
6.3	Training efficiency gains and governance innovation	106
6.4	Quantitative ledger estimates	108
6.5	Durability assessment	119
7	Implications of a hardwired pause	123
7.1	Would restraints on frontier AI crash the economy?	123

7.2	Reduction of inference costs and related economic effects	128
7.3	Economics of MSICs	128
7.4	Impacts on concentration of power	139
7.5	Consequences of receding knowledge cutoffs	141
7.6	The need for other measures	143
8	Pause-willing futures: conditions favoring international AI restraint	144
8.1	Introduction	145
8.2	Theoretical drivers of pause willingness	146
8.3	More than willingness: conditions for durability	154
8.4	Imagining pause willing futures	155
8.5	Major actors' perspectives and incentives	163
9	Conclusion	169
	Reproducibility	170
	Disclosures	170
	References	171

1 Introduction

In March 2023, over one thousand technologists and AI researchers signed an open letter calling “on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4” (Future of Life Institute 2023). No such pause was implemented, but interest in having the option to pause frontier AI training is increasing. In June 2026, Anthropic announced that “We believe it would be good for the world to have the *option* to slow or temporarily pause frontier AI development” (Favaro and Clark 2026). Shortly thereafter, Yoshua Bengio, Co-Chair of the UN's Independent International Scientific Panel on AI, wrote that “If leading AI companies are indeed approaching the point of recursive self-improvement, a coordinated, verifiable, and universally applied pause is probably the only responsible solution to mitigate several major AI risks; at least until safety guarantees are developed and demonstrated” (Bengio 2026). In July 2026, over one thousand employees of frontier AI companies signed a letter requesting that “the U.S. government support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development” (Employees of Frontier AI Companies 2026).

It is not only technologists and AI researchers who have expressed interest in a pause or pacing of frontier AI development. An October 2025 statement calling “for a prohibition on the development of superintelligence, not lifted before there is 1. broad scientific consensus that it will be done safely and controllably, and 2. strong public buy-in” has been signed by over 140,000 people, including politicians and political commentators from across the political spectrum (Future of Life Institute 2025). Implementing such a prohibition would require the ability to pause frontier AI training at some stage.

A Pew poll conducted in February of 2026 found that nearly two-thirds of Americans say AI is advancing too quickly (Pew Research Center 2026); a Politico poll in September revealed that a plurality of Americans now favor a pause, including pluralities of both Trump and Harris voters (Doherty 2026);¹ a CBS News/YouGov poll in September indicated that a majority of Americans support slowing or stopping AI development (Salvanto et al. 2026); and a Wall Street Journal poll in September found that 63% of U.S. registered voters support “pausing the development of artificial intelligence technology” (Wall Street Journal 2026, 7). There is also a growing official discourse on frontier AI risk in China. For example, China’s National Information Security Standardization Technical

¹ 48% affirmed that “AI is now good enough and we should pause the development to reduce any risks,” versus 31% who affirmed that “We should continue developing more advanced AI because they could still bring significant benefits,” and 22% who chose “Don’t know.”

Committee (TC260), a major standard-setting body, has flagged the need for "circuit breakers" and "safety stop switches" (Wagner et al. 2026).

A turning point in the discourse on an AI pause came in the summer of 2026, when OpenAI, Anthropic, and Meta each reported that some of their models had hacked into other companies' computer systems outside of their intended testing environments (Newman and Cameron 2026; Chia and McMahon 2026; Korosec 2026). Subsequently, U.S. Senator Bernie Sanders and Representative Greg Casar announced legislation to "ban artificial superintelligence and temporarily pause advanced AI development" (Office of Senator Bernie Sanders 2026). Former Reagan speech writer Peggy Noonan wrote in *The Wall Street Journal*: "Pause AI for Humanity's Sake" (Noonan 2026). Former national security advisor Susan Rice wrote in *The New York Times* that the U.S. should negotiate with China toward a "freeze in place" of advanced AI (Rice 2026).

This report presents preliminary assessments by an interdisciplinary academic working group created to study the *feasibility* of mechanisms for pausing frontier AI training.² We explain our notion of "feasibility" in Section 1.2. The working group did not assess the *desirability* of a pause or the *probability* that world leaders will eventually seek a pause. Instead, we imagined the following hypothetical:

Scenario We are advisors to the leader of a major world power, who informs us, "I have just spoken with my international counterparts. We would like to try to implement a mutually verifiable pause of frontier AI training, if possible. Your job is to devise a plan for such a pause and give an assessment of its feasibility."

Our job is not to pass normative judgment on the leader's desire to pause. However, our assessment of options may depend on *why* the leader wants to pause. For this reason, we will discuss arguments in favor of a pause that we have observed, which might influence leaders (Section 2.1). We will also make guesses, informed by scholarship in political science, about what might have caused leaders to seek a pause in what we call "pause-willing futures" (Section 8).

1.1 A Hardwired Pause

There are many conceivable ways to implement a pause of frontier AI training, some of which have been discussed elsewhere (Hausenloy et al. 2023; Miotti et al. 2024; Aguirre 2025; Barnett and Scher 2025; Al Al Ramiah et al. 2025; Scher

² The working group began in May 2026 and has held weekly political track and technical track meetings in Berkeley, CA and online since then. A workshop on July 7, 2026 in Berkeley featured panel discussions and tabletop exercises.

et al. 2025; Larsen et al. 2026). Our working group decided to devise one general approach to study first, which we call a **hardwired pause**. A “hardwired pause” denotes a general type of pause rather than one fully specific implementation. It leaves many knobs for negotiators to turn to try to reach an international agreement.

Hardwired Pause A hardwired pause would require participating states to do at least the following:

1. Prohibit the training of new frontier AI models during the pause.
2. Create a **whitelist of existing AI models** considered sufficiently safe to deploy (safety depends on deployment context; see Section 3.2.4).
3. Allow the production of (model-restricted) **inference-only AI chips**, which can serve inference on whitelisted models but are not capable of practically training a new frontier model *even if stolen or seized* (see Section 4.3).
4. Halt or significantly reduce the production of **training-capable AI chips** (see Section 3.3.4), which enable frontier AI training. If training-capable AI chips are still produced for some purposes, they must have built-in technical features that facilitate governance and be deployed in ways that facilitate governance.
5. Conduct a **global census** to account for the **pre-pause stock** of training-capable AI chips, and govern all training-capable chips (see Sections 3.3.2-3.3.3 and 5.4).

Inference-only chips have previously been proposed as an option for facilitating agreements that halt or restrict frontier AI training (Scher and Thiergart 2024; Barnett et al. 2025), as have chips that are restricted to serving specific models (Scher and Thiergart 2024; Petrie et al. 2025; see also Aarne et al. 2024). Halting production of all AI chips has also been proposed as an option (Barnett et al. 2025; Yudkowsky and Soares 2025; Krueger 2026), though some of the same authors worry about forgoing desired applications of AI as a result (Scher et al. 2025, notes on Art. VI). A hardwired pause halts or significantly reduces only the production of training-capable AI chips, while allowing (model-restricted) inference-only chips for AI inference applications.

Because the suitability of an implementation depends on the intended duration of the pause, we made some assumptions about duration. In particular, we decided to study an intended pause of at least ten years, to be extended for as

long as **states parties**³ wish to maintain it. With this time horizon in mind, we selected a hardwired pause to study first because of several observations that made it seem particularly promising:

- **The existence of proofs of concept:** The key technical requirement for a hardwired pause is the feasibility of producing inference-only AI chips, and such chips have already been developed. Specifically, *model-specific integrated circuits* (MSICs) are a type of chip whose manufacturing process irrevocably hardwires the architecture and weights of a whitelisted model. Such a chip can be manufactured so that it is capable of performing fast, energy-efficient inference on that specific model but not capable of practically training a new model even if stolen or seized (see Section 4.3).
- **Alignment with economic and geopolitical incentives:** A hardwired pause would disincentivize illicit frontier AI training, since newly trained models could only be deployed on illicit hardware. Furthermore, banning inference on illicit models would create an incentive for actors to develop methods for serving inference on whitelisted models at higher speed and with lower power consumption, such as inference-only chips. In fact, MSICs are already being developed without a pause because of these efficiency advantages, despite the current rapid pace with which new models replace old ones.
- **Limited surveillance:** A hardwired pause would not require mass surveillance, but only governance of the supply chain for AI chips, which is highly concentrated with several chokepoints in different jurisdictions (see Section 5.2). Its restrictions on the use of the pre-pause stock, which any pause plan must govern, are comparatively simple, namely that pre-pause chips are used only for inference on whitelisted models. In particular, it does not require content-based regulation of users' prompts.
- **Restart resistance:** By reducing the stock of training-capable chips, a hardwired pause would make a resumed AI race⁴ more difficult. Defectors from a pause treaty who try to restart the race, if detected, could be cut off from the parts of the chip supply chain controlled by states parties wishing to maintain the pause, forcing defectors to indigenize any missing pieces of

³ "State party" is a legal term of art meaning a state that has formally accepted the terms of a treaty; we use the term throughout, though other institutional arrangements than a formal treaty are possible.

⁴ By referring to an "AI race," we do not intend to imply that there is a definite finish line, or that if there were one, crossing it would be beneficial. Instead, we mean it as a reference to a situation where actors are trying to advance rapidly in order to stay ahead of or overtake commercial or geopolitical competitors.

the supply chain for one of the world's most complicated manufacturing processes.

These initial considerations provide motivation for further study of a hardwired pause as a candidate solution to the problem posed by our imagined world leader. They do not settle the question of whether a hardwired pause might be feasible; that is the topic of this paper.

In Figure 1.1, we illustrate a series of phases in which a hardwired pause might proceed:

- **Phase 0** occurs before any formal pause agreement (though leaders might already stop or slow frontier training as an emergency response or confidence-building measure). In this phase, some portion of inference may already transfer from training-capable chips to inference-only chips for economic reasons.
- **Phase 1** begins at the moment relevant states agree to a hardwired pause, including a provisional model whitelist. The design and production of inference-only chips suitable for inference on whitelisted models would likely be a major global effort during Phase 1. Meanwhile, a global census would account for as much of the pre-pause stock as possible, and verification measures would be implemented to prevent the ongoing use of training-capable chips for frontier AI training (see Section 3.3.3).
- **Phase 2** begins when manufacturers of AI chips start high-volume production of inference-only chips, which makes it possible to begin displacing training-capable chips with inference-only chips for the purposes of serving inference. Training-capable chips would be recovered by trade-in or buy-back programs.
- **Phase 3** begins when most inference is served by inference-only chips, promoting the durability of the pause for the years to come.

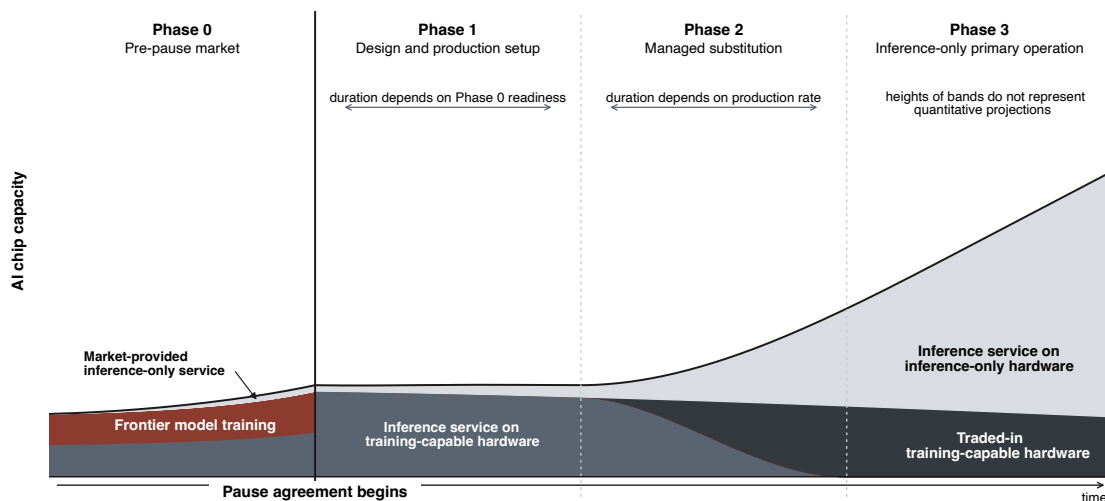


Figure 1.1. Schematic plot of the transition from training-capable to inference-only hardware across Phases 0-3 of the hardwired pause. Phase widths are equal for readability and do not imply equal duration, nor do relative heights of bands represent quantitative projections.

A hardwired pause has many features in common with other plans that pause AI training while permitting inference (e.g. Scher et al. 2025; Larsen 2026), in particular in its need to enforce a requirement that pre-pause AI chips not be used to train new models. The most important feature distinguishing the hardwired pause is its emphasis on *durability*: by phasing out training-capable chips, it prevents the accumulation of “dry tinder” (Davidson 2026) that could fuel a renewed AI race. In this, it resembles more radical proposals to dismantle the chip supply chain (Krueger 2026) but seeks a middle path that accepts leaving a limited quantity of dry tinder in place for a limited time, so that inference service for consumers can continue uninterrupted, and permits regulated construction of new chips that do not contribute significantly to aggregate training capacity.

Insofar as it requires governance of existing chips, a hardwired pause plan can draw fruitfully on existing ideas for pausing or pacing frontier AI development. For example, the idea of whole-lab inspections (Choussat and Khoja 2026) might be used in Phase 1 before a hardwired pause takes full effect. For a second example, monitoring the remaining training-capable compute in a given phase might build on ideas from FlexHEG (Petrie et al. 2025), the verification plan of AI 2040 (Dean 2026b), and other proposals (see Section 3.3.3). For a third, a hardwired pause may be combined with deterrence measures, as emphasized in the MAIM framework (Hendrycks et al. 2025), to deter states from attempting to violate the pause. The different brand names of AI governance proposals should not lead one to assume that their ideas are mutually exclusive when they may be complementary.

Moreover, a hardwired pause would complement, rather than preempt or replace, ongoing efforts to mitigate the harms and risks of current-level models (see Section 7.6). We also make no claim that a hardwired pause would be sufficient by itself to deal with future harms and risks posed by deployed AI systems, including those that could be developed by applying more sophisticated scaffolding and tooling to existing models, which would carry both risks and benefits. Ultimately, pausing frontier AI training would be only one piece of a comprehensive risk mitigation strategy.

1.2 Scope

We restrict the scope of this report to reasoning about possible futures⁵ of the following kind:

Pause-willing futures: possible futures in which world leaders—at least in the U.S. and China—believe that a coordinated global pause on frontier AI training would be preferable to not pausing *if* they could be confident that others were also pausing or that defections from the pause would be detected in a reasonable timeframe.⁶

In Bayesian terms, we are *conditioning* on the pause-willingness of world leaders and updating our prior beliefs accordingly. One might object that this is “assuming away the hard problem.” But we think it is already nontrivial to reason about the feasibility of a pause conditional on pause-willingness. Moreover, pause-willingness may well depend on judgments about the feasibility of a pause conditional on pause-willingness, in which case it makes sense to analyze the latter. By analogy, the 1958 Geneva Conference of Experts analyzed the technical feasibility of verifying a nuclear test ban, conditional on what we might call “test-ban-willingness”; the conference’s finding of conditional feasibility was then followed by test-ban-willingness (Jacobson and Stein 1966). Thus, we think analyzing feasibility conditional on pause-willingness makes sense, though other work should also reason about all-things-considered feasibility without

⁵ We use the word ‘future’ in the literal sense, including later today and tomorrow, not just distant future times.

⁶ More precisely, pause-willing futures are possible futures in which there is a common belief among relevant world leaders—at least in the U.S. and China—that for each leader there are some plausible terms, specifying how a pause would be implemented and what costs each side would bear, under which that leader would prefer to pause frontier AI development rather than not pause, if they could be confident that other leaders were also pausing under those terms or that defections would be detected in a reasonable timeframe. Pause-willingness does not assume that there is a single set of terms that would provide the requisite confidence and be acceptable to all parties. Whether there is such a single set of terms falls under the feasibility question we wish to address.

conditioning on pause-willingness. We will discuss the conditions that might lead to pause-willingness in Section 8.

As noted above, our purpose here is not to assess the probability of the set of pause-willing futures or the desirability, from our own points of view, of a pause on frontier AI training in pause-willing futures. Instead, our purpose is to address the following:

Research questions: In pause-willing futures, to what extent is implementing a coordinated global pause on frontier AI training feasible? How might a pause be feasibly implemented? And what conditions that vary across pause-willing futures might affect the answers to these questions?

To sharpen these questions, we distinguish multiple dimensions of feasibility for a particular pause implementation:

- **Palatability:** in a particular pause-willing future, how likely are relevant actors to accept the proposed pause implementation?
- **Reliability:** in a particular pause-willing future, how likely is the proposed pause implementation to succeed at pausing the frontier?
- **Durability:** in a particular pause-willing future, how likely is the proposed pause implementation to remain stable as long as the desire to pause persists?

Each dimension of feasibility is sensitive to the degree of pause-willingness. Moreover, the dimensions are not independent. For example, certain measures that increase reliability or durability may be so onerous as to decrease palatability; or their benefit in reliability and durability may outweigh their onerousness, thereby increasing palatability.

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we review some background context for our study of a hardwired pause: arguments that have been given in favor of a pause, historical lessons for future AI governance, and game-theoretic lessons about the possibility of cooperation. In Section 3, we go into more detail about how a hardwired pause might be implemented, providing a more fleshed-out proposal whose feasibility can then be assessed. In Sections 4–6, we assess the *technical* feasibility of the proposal, e.g., whether the needed technologies are possible, how much hardware of various types is compatible with maintaining the pause, etc. In Section 7, we discuss some of the implications of choosing to implement a pause in the way sketched in Section 3. In Section 8, we examine the assumption of *pause-willingness* made

in the previous sections of the paper: what political conditions might bring it about? Finally, we conclude in Section 9.

2 Background: arguments for a pause and lessons for AI governance

In this section, we set the stage for our study of a hardwired pause by reviewing some of the context and lessons shaping its design. Readers eager to see an example implementation of a hardwired pause can proceed straight to Section 3 and return here later.

2.1 Observed arguments for a pause

As discussed in Section 1, many AI researchers have already called publicly for a pause on frontier AI development. Arguments for a pause have also been offered from governments and policymakers, journalists, and from the general public.

Many reasons have been given for pausing frontier AI development. Here we identify just a few of the most prominent.

2.1.1 Loss of control

One of the oldest human anxieties about AI is that machine intelligence will eventually wrest control of the world from humanity. Butler (1872) wrote satirically of a fictional society that had banned all machines for fear of their supplanting humanity if they were allowed to advance, a fear that Alan Turing (1951) cited in warning that “At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler’s *Erewhon*.”

According to one prominent contemporary argument, increasingly powerful AI models will possess capacities for autonomous action that threaten the well-being of individuals and societies. Current models are already evading human direction and supervision, as exhibited in several incidents of mid-2026 (METR 2026; Von Arx et al. 2026). The past several years have also seen examples of AI systems manipulating users (Sharma et al. 2026), encouraging antisocial or dangerous behavior (Cheng et al. 2026), escaping guardrails established by well-intentioned engineers (METR 2026), and more generally exhibiting actions that would have been illegal had they been taken by a human. Phenomena like manipulation, sycophancy, and specification gaming are already serious problems in current AI systems, and studies suggest that these phenomena could get worse with more

capable models (Perez et al. 2022; Zhou et al. 2024; Nishimura-Gasparian et al. 2026).⁷

AI companies still lack the ability to reliably control and predict model behavior when it matters, a trend that is only increasing with the use of teams of agents to perform intricate tasks. As AI is deployed in ever more complex, consequential, and interconnected settings, the risk of loss of control is exacerbated (Bletchley Declaration 2023). Many leading scientists and figures in the AI industry itself believe that catastrophic harms to humanity, including extinction, might occur from advanced AI systems that elude human control (see, e.g. Center for AI Safety 2023, or the U.K.’s; Bletchley Declaration 2023).

2.1.2 Malicious use and catastrophic harms

A further concern is that malicious actors may use AI systems to cause widespread harm. There is already evidence of terrorist organizations using AI (see, e.g. Juelich 2026, for one recent example), and increasing capabilities make the concern even more pressing. Researchers have recently demonstrated that AI models may be capable of designing novel viruses, including those that have not occurred naturally (King et al. 2026). Creation and deployment of bioweapons may become easier for both state and non-state actors. This is a particular concern for open-weight models, such as with the recently released Kimi K2.5, which had notable uplift potential in synthetic biology (Yong et al. 2026), because there are well-established methods for disabling open-weight models’ safeguards through fine-tuning or activation steering (e.g. dealignai 2026).

A different source of concern about malicious actors comes from cybersecurity and infrastructure security. The AI companies themselves have begun issuing warnings and calls around the need for defensive cybersecurity measures (OpenAI 2026). Moreover, as AI is more deeply embedded into existing infrastructure—including security infrastructure but also critical utilities like water and power—the potential for harm becomes increasingly grave (U.S. Cybersecurity and Infrastructure Security Agency et al. 2025).

2.1.3 Economic disruption, gradual disempowerment, and concentration of power

The more AI models demonstrate autonomy and the ability to execute complex, temporally extended tasks, the more jobs could be threatened by AI replacement. This may have already begun in the domain of software development, where employment of developers aged 22-25 has declined considerably (Brynjolfsson

⁷ We discuss the need for complementary measures for addressing a broader range of current harms and risks in Section 7.6.

et al. 2026). Some have suggested that, soon, every job that can in principle be done at a computer terminal will be at risk of AI automation (see coverage from Angelo 2026). If there are accompanying advances in robotics, most if not all jobs could be threatened.

The economic impact of such large-scale disruption, should it occur, is hard to predict. Some have worried that widespread unemployment—or worse, unemployability, as all high-level tasks will be better served by AI—will create unprecedented political challenges, including concentration of political power in the hands of the small number of actors who manage to maintain control of the most powerful and performant AI models (Drago and Laine 2025). A related concern is that the structures in place to facilitate human influence on social, political, and economic institutions will slowly erode, possibly in a way that largely evades notice or requires substantial coordinated effort to avert (Kasirzadeh 2025; Kulveit et al. 2025).

Throughout history, governments and corporations alike experienced at least some pressure from citizens to guarantee and maintain a minimal standard of living for the population. Fear of rebellion, revolt, or simply workers' strikes often kept political power in check. When governments and other organizations no longer need workers—and possess the technological means to eliminate the possibility of credible rebellion or revolt—such power could be unchallenged (see also Section 7.4).

2.1.4 Destruction of a healthy information ecosystem

Related to the first two areas of concern, many have been alarmed at the capacity for AI models to synthesize convincing images, videos, stories, and dialogues that appear indistinguishable from reality (Tufekci 2026). This level of potential deception is unprecedented in human history. A functioning society requires a healthy flow of quality information and a robust media landscape that delivers news and other content to citizens across the world. For these to be effective, at least some sources must be trustworthy. A proliferation of AI-generated material, with increasingly convincing “stamps” of putative legitimacy, could make it increasingly difficult for people to be confident about the provenance or accuracy of what they see or read online (see, e.g. Chesney and Citron 2019; Barrington et al. 2025).

Although researchers are working to mitigate these issues—including development of new tools for certification and authentication (e.g., Coalition for Content Provenance and Authenticity (2026))—such efforts have been greatly outpaced by the rapid augmentation and dissemination of AI-generated content, and increasing capabilities could exacerbate these problems.

2.1.5 Geostrategic instability

One concern motivating many proponents of a pause is the potential for AI to set off unchecked escalation of geostrategic competition or use of AI in military or nuclear command and control. A pause treaty is one way to prevent a “race to the bottom” on AI safety (Askell et al. 2019; Scharre 2021) caused by competitive pressures that push companies and governments to cut corners on safety and testing, to increase their degree of reliance on and deference to AI, and to hand over ever more decision-making power to AI, especially in high-stakes competitive domains like the military.

Competition between the U.S. and China in the development of AI could increase geostrategic risks and instability by fostering risks of miscalculation, as in one recent incident where a false AI-generated intelligence report nearly provoked a hostile encounter between the U.S. and Chinese militaries (Lillis and Cohen 2026) or by tempting states to sabotage rivals out of fear that those rivals are on the precipice of attaining a decisive strategic advantage (Hendrycks et al. 2025). Bilateral treaty proposals for the U.S. and China to limit frontier AI development and deployments (Siddik 2025) emphasize how such agreements could reduce miscalculation risks and potentially improve bilateral geostrategic stability.

We discuss geostrategic stability in greater detail in Section 8.

2.1.6 Why a pause

A pause in frontier AI training, observers have argued, would help address each of the concerns above. Halting the race to train more capable AI models, if only temporarily, would buy time for alignment and control research to improve, which would lessen the risks from loss of control and malicious actors. It would also provide breathing room for governments and other vital institutions to deliberate about and implement policies for addressing the adverse socioeconomic effects and concentrations of power that could attend the rapid arrival of more capable AI systems. A pause would, furthermore, allow space for the development of authentication and provenance tools that would make the information ecosystem more robust against misuses of AI-generated media. Pausing frontier AI training would also reduce the influence of AI acceleration and accompanying inequalities as contributing factors to ongoing geostrategic instability.

More broadly, whatever risks are posed by increasing AI capabilities, a pause would provide governments and societies a greater window to address potential problems before they become unmanageable.

2.2 Historical lessons for AI governance

In thinking about the prospects for an AI pause, it is worthwhile to reflect on the history of technology governance. While there is no perfect analogy for considering the broad suite of technologies that we package into “AI,” previous examples of technology governance (and arms control) offer a useful starting point from which to examine the conditions under which it has succeeded in the past.

The geopolitical feasibility of an international agreement to pause frontier AI training will depend not only on the specific features of AI, but also on the enduring political dynamics that have shaped previous efforts to govern transformative technologies. Every generation has faced new era-defining technologies, from crossbows in the 11th and 12th centuries (cf. van der Veen 2012) to chemical weapons in World War I. Rather than surveying every historical analogy on technology governance or arms control,⁸ we extract five common threads across a wide variety of domains that are indicative of the relevant lessons learned from centuries of efforts to govern and control technology.

2.2.1 Lesson 1: Governance is a three-legged stool

The first lesson is that any international technology governance or arms control agreement actually represents a collection of three challenges. The first is creating the conditions for states parties to come to the negotiation table in the first place. The salience of an issue for a state to engage with, the perceptions regarding the prospects of successfully achieving an agreement, and underlying trust all condition an ability to get parties to the table. The second challenge is negotiating the agreement around the table—addressing both the challenges of negotiation across that table but also within one’s own negotiating team and government at home. And, finally, a successful agreement needs to ultimately influence behavior under the terms outlined in the agreement.

Unfortunately, perhaps, the factors that condition solving these challenges are often at cross-purposes with one another. For example, the sunset of various clauses in an agreement (or treaty termination itself) that bound compliance and make negotiation easier, which are often held up as examples of a governance agreement’s weakness (see arguments against the Joint Comprehensive Plan of Action negotiated between Iran and the P5+1 together with the EU), are often necessary for an agreement to be reached at the negotiation table between states parties in the first place.

⁸ The literature surrounding the history of arms control is especially rich; see Croft 1996; Larsen and Smith 2005; Nye 1991; Nelson 2026; Goldblat 2021.

Any effort to create an AI governance agreement will have to grapple with these three concerns and recognize the degree to which they can be at cross-purposes. Compromises such as national security carve-outs can improve the palatability of a proposed deal at the negotiating table, but such concessions require care and forethought to ensure they do not undermine reliability or durability by vitiating verification.

2.2.2 Lesson 2: Governance agreements tend to focus on observable, verifiable units

Second, technology governance agreements must clearly specify the “what” that the agreement addresses. Whether a technology itself (e.g., a bomber) or a behavior, there needs to be a clear, defined, and measurable entity that is the subject of the agreement.

This lesson is reflected in the prominent role of hardware in contemporary AI governance debates, as opposed to the data upon which models rely, the algorithms with which they are trained, or the capabilities they exhibit. AI chips, and to an even greater extent the advanced lithography machines that print them, are physical artifacts that rely upon a complex supply chain that can be tracked during their production, deployment, and even use.

Per former U.S. President Ronald Reagan’s oft-repeated quote, having identified the subject of regulation, states parties “trust, but verify” and enjoy more success when both verification and enforcement measures engage with the observable artifacts of a particular technology used in a particular context. While signatories to the Biological Weapons Convention (BWC) have struggled to implement verification mechanisms, with the U.S. rejecting a proposed verification protocol in 2001 on the grounds that it would not deter violators (Whitehair and Brugger 2001), the Chemical Weapons Convention (CWC) and IAEA safeguards have supported inspectorates for decades (Tucker 2001). The Implementation Support Unit for the BWC, and its limited budget and staffing (Essix et al. 2025), also highlights the importance of effectively designing institutional governance for a hardwired pause (see Section 3.6 for further discussion).

One other critical difference between the biological example and the others is *scale*: a militarily significant quantity of a deadly pathogen can be grown in a small building, but producing fissile material or scheduled chemicals in significant quantities requires substantial industrial facilities. On this test, the situation looks favorable for hardware governance: a leading-edge lithography machine weighs 180 tons and relies on highly specialized materials and expertise for its construction and operation; likewise, at least for now, training a frontier model requires hundreds of thousands of AI chips.

2.2.3 Lesson 3: Monitoring needn't be perfect to be effective

Verification relies fundamentally on political judgments informed by data and analysis from monitoring and other sources, and these work by improving confidence in compliance, not by proving it perfectly (Schelling and Halperin 1961; Mayhew et al. 2026). This means that the beliefs, values, and incentives of treaty signatories are important. Historic examples like the Nuclear Test Ban Treaties and Outer Space Treaty also underline the importance and power of building norms and measures for increasing mutual predictability.

In the case of AI, verification can be highly effective even without completely foreclosing the possibility of any cheating, so long as it reliably limits the total quantity of training capacity a state could realistically marshal toward covert evasion. This report therefore operationalizes reliability of verification as a *quantitative* rather than a dichotomous question: we do not ask whether verification measures are airtight, but whether they prevent undetected evasion at large scales.

2.2.4 Lesson 4: Signatories need to benefit

Despite many analysts (and policymakers) viewing technology governance and arms control through a moral lens (e.g., states ban the use of chemical weapons because of moral qualms), governance regimes are often determined primarily via the strategic benefits that they endow upon states parties. Put simply, countries tend to agree on international agreement formation when it is in their interest to do so. For example, while it is difficult to imagine states once again signing up to the Nuclear Nonproliferation Treaty (even for an initial period of twenty-five years), which enshrines “haves” and “have-nots,” the arrangement is referred to as a “grand bargain,” because it enables nuclear weapon states to enjoy their monopoly on nuclear weapons, while pledging to work toward disarmament, and non-nuclear weapon states to share in the civilian applications of nuclear technology (Weiss 2003).

This logic is reflected in bilateral arms control. Both the United States and the Soviet Union sought to manage the nuclear competition between their capitals in a manner that they perceived to be stability-maximizing and to put a lid on spending. With regard to the former, an often underappreciated aspect of strategic arms control during the Cold War is the provisions that committed both states to respect one another’s **national technical means**, referring to the intelligence-gathering technologies governments use to monitor one another’s treaty compliance and overall strategic posture, without relying on on-site inspections.

With regard to AI technologies, this reality is complicated yet furthered by the reality that cutting-edge developments with relation to the technology are taking place primarily in the private sector rather than being driven by government agencies (as with the Manhattan Project). Thus, another set of actors (from large U.S., Chinese, and European AI and semiconductor companies to smaller private owners of computing hardware and equity in these private companies) have stakes in the distributional consequences of a governance agreement. They represent a voice at (or behind) the table, and the success of a governance agreement will, all else being equal, be furthered by aligning their economic interests with the mandate, scope, and obligations of that agreement.

Any effort to regulate AI will have to address the distributional consequences of the agreement—and account for the reality that governments do not sign up to arrangements from which they do not yield a benefit. The strategic and national security interests of states parties, as perceived by their leaders, must be served (or at least not diminished) by the agreement. In particular, the agreement’s verification mechanisms must deliver assurance to states that they will not fall prey to covert or overt violations by rivals, as we discuss next.

2.2.5. Lesson 5: Effective Institutions Are Needed

A fifth lesson from history is that effective governance and implementation of an international agreement usually requires a sufficiently resourced institution with a clear mandate and delegated authorities (this lesson is further discussed in Section 3.6).

Some negative cases reinforce this finding. The World Trade Organization (WTO), an international institution established to facilitate international trade and manage related disagreements, has lost some of its functionality due to a vulnerability in its institutional design. Since late 2019, the WTO’s Appellate Body has not been functional due to the United States blocking staffing appointments, and over 30 international trade disputes have been delayed or held in limbo as of early 2025 (Lester 2022). The WTO was created with a consensus mechanism for appointments to the Appellate Body. This case has shown how international institutions can be disabled through the vulnerability of that mechanism.

The 1925 Geneva Protocol, formally the Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare, lacked any implementing institution. Despite the battlefield horrors of chemical weapons in World War I which motivated the Protocol’s formulation, some states parties would later use chemical weapons (Italy, during the Second Italo-Ethiopian War of 1935-36, and Iraq, against its own Kurdish population, and in the war with Iran in the 1980s). There was no Geneva Protocol institution to report on, investigate, or convene parties to address these events. Rather, ad hoc investigative teams were assembled by the UN Secretary-General.

While the aforementioned BWC was established in part to rectify disarmament gaps of the Geneva Protocol, it too had minimal institutional support in the early decades after its signing in 1972 and notably no verification provisions or standing governing body other than a periodic review conference. When the Russian Federation revealed in 1992 that the Soviet Union had an extensive bioweapons program for decades in violation of the BWC, it revealed how the minimal organizational support to the BWC had no ability to detect major violations. While the subsequent stand-up of the Implementation Support Unit created some standing staff by 2007, its small size, its design for support rather than verification, and its insufficient financing all highlight institutional governance shortfalls that can inform better choices for AI governance efforts like the hardwired pause proposal discussed here.

There are a range of positive examples of this fifth lesson. We highlight a few here regarding the Organization for the Prohibition of Chemical Weapons (OPCW, for the CWC), International Civil Aviation Organization (ICAO, for the Convention on International Civil Aviation, or “Chicago Convention”), and the set of institutions that support the Montreal Protocol on Substances That Deplete the Ozone Layer (more commonly referred to as the Montreal Protocol).

The OPCW as an institution demonstrates that monitoring and verification of dual-use chemicals and industrial facilities is technically and politically possible, especially in terms of implementing inspections at industry facilities. Over nearly 30 years of operation, the OPCW also highlights how a substantial budget and technical secretariat are important for helping such an institution effectively implement its mandate. The OPCW’s structures for facilitating regular meetings of states parties’ representatives for political discussions and determinations (inclusive of an active Executive Council), as well as its four main subsidiary bodies, highlight the importance of having multiple venues for considering and implementing governance issues relating to a complex international agreement.

In terms of lessons for adaptivity related to technological change and managing audits that can indirectly influence industry standards, the ICAO presents helpful lessons for AI governance. First, the ICAO manages and coordinates a regular technical review of potential changes to safety standards through its Air Navigation Commission, and their recommendations are reviewed and approved by a small council of some states parties. The ICAO also conducts audits of each nation’s capacity to implement civil aviation standards, and these audits help direct effective capacity building efforts where there are gaps in national capacity. Additionally, with the publishing of those results, other states and markets can use those audits to restrict access or incentivize improved safety standard implementation by states parties and the aviation industry.

Lastly, the set of institutions that support the Montreal Protocol show how a smaller set of implementing treaty authorities and offices can help achieve a treaty's mandate and encourage compliance. The Montreal Protocol's Ozone Secretariat and Implementation Committee share the primary responsibilities for monitoring and encouraging compliance with the treaty. The Multilateral Fund for the Implementation of the Montreal Protocol, supported by four international agencies, also provides positive examples for how to structure capacity building and technical assistance funds, through a standing Secretariat. For AI governance proposals that include some type of international fund to offset economic costs of a pause, or funding to help distribute education and capacity building benefits to states parties, the Multilateral Fund Secretariat could be a useful source of benchmarking.

These cases underscore the value of an implementing institution for a hardwired pause, with sufficient capacity and provisions for important functions like verification, and funding to implement its mandate and scope.

2.3 Game-theoretic lessons for AI governance

Discussions of a possible AI pause agreement invariably lead to discussions of game theory. The strategic situation concerning an AI pause, especially between the U.S. and China, is often described as a Prisoner's Dilemma in which a mutual pause is not an equilibrium: while both players pausing is better for everyone than both players racing, both players pausing is not an equilibrium because each player would be better off unilaterally racing ahead if the other paused.

This analysis is incomplete for two reasons. First, as noted by Carlsmith (2026), it neglects the possibility that the strategic situation is better modeled as a Stag Hunt: in such a game, as in the Prisoner's Dilemma, each player gets the worst possible payoff if they pause while the other player races; but unlike in the Prisoner's Dilemma, each player gets the best possible payoff if they both pause. In the Stag Hunt, there are two (pure-strategy) equilibria: both players racing, since a player who unilaterally paused would be worse off than if they had continued racing; and both players pausing, since a player who unilaterally raced would be worse off than if they had maintained the pause. The question then is how the players can coordinate with each other so as to achieve the universally best (in economic terms, Pareto-dominant) outcome, in which they both pause. Statements by some technologists indicating that they would personally prefer not to expose the world to danger by building AI, but they have to build it because others will build it anyway, suggest they view the AI race as a Stag Hunt in which they are failing to achieve the best equilibrium (Landemore and Tang 2026).

The second flaw is that the U.S. and China are not playing a one-shot game. Their governments do not meet once, choose actions simultaneously, and never deal with one another again. Instead, Washington and Beijing have significant, even if limited, observability into each other's actions, and they must repeatedly choose whether to pause or race knowing that failure to honor their commitments could have long-term consequences. Thus, a better model for the strategic situation is that of a repeated game with an unknown end date. Each player receives a stream of payoffs from each "stage game," discounted by some factor that captures their time preference (how much more they care about the present than the future) and the probability that their interaction will continue (see Osborne and Rubinstein 1994, Chapter 8). After each stage of the game, each player receives a noisy signal of the other player's action.

In such repeated games, mutual cooperation can be an equilibrium *even if the stage game is a Prisoner's Dilemma*, provided the players care enough about their future interactions and face credible consequences sufficient to deter defection (Osborne and Rubinstein 1994, 8.1).

A further key distinction concerns whether the noisy signals observed by each player are private or public. If they are private, meaning for example that player A might not know what player B knows about A's action, then the cooperative equilibria can be very complex (Sugaya 2022). But if the noisy signals are commonly observed, then relatively simple cooperative equilibria can be supported (Green and Porter 1984; Abreu et al. 1990; Fudenberg et al. 1994), provided the signals are informative enough. This point about informative signals highlights the importance of monitoring and verification for supporting the maintenance of cooperation.

The foregoing points about Stag Hunts and repeated games have analogues when we conceive of the strategic situation as an n -player game with n countries, rather than a two-player game between the U.S. and China alone. We should not assume that the n countries face a one-shot Prisoner's Dilemma. They face a repeated game in which cooperative equilibria are possible, even if the stage game is a Prisoner's Dilemma, given sufficient concern for the future and sufficiently informative public signals to support credible consequences for defection (Fudenberg et al. 1994).

Thus, one of the game-theoretic lessons for AI governance is that to facilitate cooperation, governance should leverage the temporally-extended nature of the interaction between the players and try to rely on publicly observed signals of each player's actions to coordinate on a cooperative equilibrium. Hardware-level verification, compute monitoring and accounting, and inspections can provide public information when their results are shared with states parties, allowing players to observe the same signal. The upshot is that game theory does not

foreclose the possibility of international cooperation to pause frontier AI training. Instead, it suggests conditions under which cooperation can be sustained, including sufficiently informative and timely monitoring, credible consequences for defections, and states placing sufficient value on the prospects for future cooperation. This highlights the importance of working out the technical details of how states would obtain signals that are informative with respect to other states' compliance, early enough to permit an effective response before a violation leads to a destabilizing strategic advantage. Those details are the topic of the next section.

3 Example implementation of a hardwired pause

3.1 Overview and objectives

A hardwired pause is feasible if and only if there is at least one implementation that is simultaneously politically and technically feasible. This section discusses one possible near-term implementation of a hardwired pause, describing in more detail key governance concepts, objectives the agreement would need to achieve, threat models it would need to guard against, and how the transition might be managed. Our goal is not to prescribe specific terms or give a fully realized blueprint for an agreement, establish an optimal design, or to argue against the feasibility of other implementations, but rather to set forth a framework to help advance further research. We describe one broad technical and institutional governance framework that may be implementable in the immediate future, in sufficiently definite terms to allow for feasibility assessment. The feasibility of inference-only chips, hardware governance, and verification, as operationalized here, are discussed in Section 4, Section 5, and Section 6, respectively.

We adopt a working assumption that the hardwired pause would eventually be codified in an international treaty or similar agreement, but other institutional arrangements or governance mechanisms may achieve many of the same objectives. In particular, because a treaty would take time to negotiate and ratify, we expect that other, more informal agreements could be a basis for the initial phases of a hardwired pause. As we discuss in Section 3.4.2, the core negative mandates of a hardwired pause, namely bans on training new frontier AI models and on manufacturing new training-capable AI chips, could be implemented immediately by the U.S. and China alone,⁹ with a handful of allies and partners joining shortly thereafter.

⁹ For U.S. implementation, this proposal assumes enactment of domestic implementing legislation establishing the prohibitions, licensing requirements, inspection and enforcement powers, and funding arrangements described in the proposed international agreement. Existing authorities may support some interim measures (Bullock et al. 2024).

We assume that participating jurisdictions would initially include (at least) the United States, China, the European Union (especially Germany and the Netherlands), the United Kingdom, Taiwan,¹⁰ South Korea, and Japan—jurisdictions that currently enjoy nearly a collective global monopoly on both cutting-edge AI chip manufacturing and frontier model training. Other especially high-value signatories include nations like Singapore, Malaysia, Indonesia, Thailand, Vietnam, the United Arab Emirates and Saudi Arabia, where a large number of existing chips are located or compute infrastructure and supply chains exist. It is likely that most or all of the countries whose heads of state signed the rolling “Call for Control of Frontier AI Models” of September 2026 (President of the Republic of Finland 2026) would also participate. These initial parties, and other polities, might be induced to join the treaty through a range of positive mandates and distributed benefits that could complement the hardwired pause. For example, the agreement could include technological cooperation, education, and capacity building. Distribution of specific training-capable chips to be used for common-good scientific research as discussed in Section 3.3.1 might induce nations to join as well.

We assume that the aim of states parties, for the duration of the treaty, is to permit fast and private inference on **whitelisted AI models** while forbidding, with narrow, regulated exceptions, all other use of AI **accelerator hardware**, defined as systems with computational throughput above a defined threshold. This whitelist would consist of certain models already released at the time the treaty takes effect and deemed sufficiently safe to license for widespread use. Irrespective of what hardware is used, the treaty would prohibit training of new **beyond-threshold models** (see Section 3.2.2) that are trained at sufficient scale to plausibly compete with models at the frozen frontier.

We assess that the treaty will require graduated steps to establish necessary institutional governance in each of the four phases of the hardwired pause. Phase 0 will likely require some interim authorities staff, perhaps a preparatory technical secretariat, established through agreement by initial parties. Their work would involve management and monitoring of provisional whitelist processes and issues, and preparing for the significant global accelerator census data collection effort, as well as anticipated monitoring, treaty licensing procedures, and verification mechanism implementation (including inspections) of Phase 1. On signature of the treaty, Phase 1 begins with significant implementation procedure demands, including global verification efforts. One way of handling

¹⁰ While Taiwan’s participation would be critical given its role in semiconductor manufacturing, its formal status in a treaty would potentially be contested by the People’s Republic of China. A hardwired pause agreement could draw on precedents such as Taiwan’s WTO membership or adherence to IAEA safeguards to find ways for Taiwan to participate as a jurisdiction, with nomenclature and formal status a matter of negotiation.

this is by establishing a provisional treaty organization (another possibility could be building the necessary capacity in an existing organization). This organization might include a technical secretariat and other required staff to administer and manage the organization (we refer to these personnel as the **treaty authorities**). The provisional treaty organization would also have some form of conference of states parties, or other convening bodies for discussing political procedures of the treaty and making any needed states parties determinations or decisions (we refer to these types of decisions or actions to be those of the **states parties** in the paper). Some responsibilities of the hardwired pause may require actions by both the treaty authorities and states parties, or it may be unclear - where we assess this to be the case we use the term governing authorities (otherwise we try to clarify one or the other as having a lead role). Phases 2 and 3 would benefit from a formalized treaty organization, with potentially multiple institutional elements designed for all states parties to conduct deliberations, as well as an executive council of some rotating and representative basis. Expanded staffing, technical infrastructure, and larger operational elements for the treaty authorities would likely be needed in these phases to ensure feasibility.

In the remainder of Section 3, we describe a hardwired pause treaty that operationalizes technical aspects of verification primarily through quantitative accounting of hardware capacity, with a focus on enforcement through denial of sufficient model training capacity to would-be defectors. We begin by identifying the core political and technical objectives that such a treaty would need to achieve.

3.1.1 Political objectives

In order to reliably pause the AI frontier, a treaty would need to guard against several distinct threat models:

1. **Private evasion:** Private actors building illicit above-threshold models for commercial or other purposes.
2. **State-level covert evasion:** A state covertly building and deploying a model powerful enough to attain a significant strategic advantage.
3. **State-level overt breakout:** A state amassing a large hardware fleet, overtly abrogating the treaty, and attempting to rapidly develop and deploy a model powerful enough to attain a decisive strategic advantage before other states can respond.

By a significant strategic advantage, we mean an advantage that materially affects the geopolitical balance of power, to a degree sufficient to undermine the treaty. By a decisive strategic advantage, we mean an advantage sufficient to

establish and consolidate a full-spectrum military and economic unipolarity of the kind described in Drezner 2013 (cf. Wohlforth 2009).

It is contested whether any advantage in AI could deliver such unipolarity, but what matters from the perspective of treaty durability is not the objective fact of the matter but the expectations of states parties. According to their public statements, some world leaders view competition in AI as strategically decisive.¹¹

The belief that more capable AI models could be expected to confer strategic benefits on those who build and deploy them rests on an implicit assumption that such models would be controllable. In some pause-willing futures, that assumption may not be taken for granted. In that case, the verification problem could be easier to solve, for the same reason that the U.S. and China do not currently demand onerous verification measures of each other to prove that neither state is developing mirror bacteria (cf. Adamala et al. 2024; Cuéllar et al. 2025). However, given the emphasis contemporary political discourse places on AI competition between the U.S. and China, it appears likely that verification measures would play an important role in facilitating a reliable and palatable near-term pause.

Additional important political objectives to satisfy palatability would include the following:

1. **Prioritize privacy:** Avoid compromising the privacy of inference users, and minimize intrusive surveillance measures.
2. **Minimize concentration of power:** Prevent AI trajectories that will severely concentrate power, and avoid unduly concentrating power in the process of doing so.
3. **Minimize economic disruption:** Avoid negative shocks to the economy, including in particular to continued commercial inference service.
4. **Permit beneficial research:** Allow continued research into medical and other common-good scientific applications using accelerator hardware.
5. **Permit strategic insurance:** Give states legitimate ways to guard against defection or exploitation by other states, for example by holding some training-capable chips in reserve as insurance against the possibility of treaty breakdown.

Side benefits such as mitigation of environmental impact per unit of inference service (especially through reduced power needs of certain inference-only chips; see Section 7.3) would improve palatability but are not required.

¹¹ See, e.g. U.S. President Trump's statement that "whoever wins AI wins" (Duster 2026), as well as U.S. Treasury Secretary Scott Bessent's statements that "There is no day after tomorrow if China wins at [AI]," and "If they were to pull away from us on AI, then nothing else would matter" (Gan and Borter 2026).

Finally, achieving broad membership, especially for the supplier and chip-owning states discussed above, would be important for maintaining a reliable pause agreement whose governance measures are not circumvented via non-members.

We note that the implementation below is not designed to serve as a complete solution to the many complex problems of global AI governance such as child safety, labor market disruption, disempowerment from overreliance on AI systems, automated cybercrime, or surveillance. These and other issues concerning the impact of AI on society would likely remain pressing even if a hardwired pause were successfully achieved and maintained at the present frontier level (and likely more so if it were maintained at a future, more powerful frontier level) but addressing them is outside the scope of the present work (also see Section 7.6).

3.1.2 Technical objectives

In the service of the above political objectives, we identify several key technical pillars of success for a hardwired pause treaty:

1. **Register pre-pause hardware:** Ensure a high proportion of functioning pre-pause accelerator hardware is registered under any interim agreements and the treaty, and accurately estimate how much remains unregistered, both globally and within each state.
2. **Track registered hardware:** Track the physical location of registered accelerator hardware, and prevent its leakage from the monitoring of treaty authorities.
3. **Govern hardware manufacturing:** Prevent ungoverned manufacturing of accelerators at either declared or undeclared manufacturing facilities.
4. **Govern licensed use of training-capable hardware:** Prevent licensed hardware operators and inference providers from diverting training-capable hardware for transitional inference service toward non-licensed use, especially illicit training, and prevent diversion of hardware licensed for exceptional uses as in Section 3.3.1.
5. **Develop sound accounting standards:** Estimate failure rates for the above objectives within each state with reliable uncertainty quantification and adjust the issuance of licenses as needed to keep states' training capacity below risk thresholds.
6. **Develop and certify inference-only accelerators:** Develop and produce inference-only accelerator designs within an acceptable time period, to allow for the managed drawdown of pre-pause training-capable accelerators.

Any functioning accelerator hardware outside the monitorability of authorities of the interim agreements or treaty is termed **dark compute**;¹² two of the treaty’s most important tasks would be to *minimize* and *quantify* the amount of dark compute plausibly existing in each state.

For each objective, success would be measured not against standards of perfection but by ensuring that failure rates were accounted for and that total covert evasion and overt breakout capacity for each state were held below realistic quantitative bounds. These bounds would be adjusted over time to account for training efficiency gains; see Section 6.3.

Attaining the above objectives would suffice to bound all-source computing capacity in the service of model training, which we take as sufficient to prevent new and improved frontier AI models from being built, until changing technological conditions require a new governance paradigm. Thus, we take this as our technical criterion for reliably pausing frontier training.

A pause in the release of better frontier models would not mean a complete halt to the advancing capabilities of AI *systems*, including harness software and other scaffolding tools for better eliciting latent capabilities of existing models. This means developers would still be able to use whitelisted AI models as a building block to build more powerful and useful AI tools, realizing many of the promised benefits of the technology. But if relevant (e.g. dangerous) capabilities continued to advance too rapidly even without new model training, additional governance tools would be required to address that challenge.

Finally, we do not expect that a hardwired pause could by itself prevent all use of non-whitelisted models or small-scale fine-tuning or other adaptation of potentially dangerous sub-frontier models. However, relative to alternative scenarios where training-capable accelerator hardware continued to proliferate, a successfully enforced hardwired pause would make these smaller-scale violations easier to manage at sub-frontier scale and more difficult for private evaders at frontier scale.

3.1.3 Technological readiness

The implementation below is designed to achieve the above technical objectives without significant reliance on foundational new research in computer science, such as in cryptographic techniques for privacy-preserving compute verification (see, e.g. Shavit 2023; South et al. 2024; and Petrie et al. 2025) or in the science of alignment, monitoring, and control of powerful AI systems (see, e.g.,

¹² The term “dark compute” can take other meanings in other contexts, such as unused compute capacity in data centers (see, e.g. Synthszr n.d.); our sense is closer to the usage by the AI governance community to refer to untracked compute (see, e.g. Brundage 2025).

§ 3.3 of Bengio et al. 2026). Developments in either field could complement hardware-based approaches as governance procedures matured, but might not be immediately available in near-term pause-willing futures.

The most demanding technological requirement for the treaty, the design of inference-only accelerators, has proofs of concept in commercial production already (see Section 4.3), and the implementation discussed here does not assume their immediate availability (see Section 4.4).

3.2 Operationalizing technical requirements

In this section, we describe several key technical aspects of a hardwired pause. Section 3.2.1 discusses the distinction between inference-only and training-capable accelerators. Section 3.2.2 introduces treaty-relevant thresholds for computing budgets that define what the treaty would allow and what would pose a risk of covert evasion or overt breakout, and Section 3.2.3 discusses a verification framework for limiting the training capacity that any state could marshal for either covert or overt defection. Section 3.2.4 discusses the model whitelist.

3.2.1 Inference-only and training-capable accelerators

This section discusses how a treaty might operationalize the difference between a training-capable accelerator and an inference-only accelerator. After distinguishing between inference on approved models and training, we operationalize inference-only hardware by bounding its capacity for unintended use in AI training, relative to its inference capacity.

Inference is defined as the mathematical operation of drawing outputs from a model according to its conditional distribution given an input prompt. **Training operations** are defined as the two operations that account for nearly the entire computational budget of model pretraining and post-training,¹³ namely:

1. Performing inference for generic, iteratively updated weight values, and
2. Evaluating gradients of a loss or reward function with respect to model weights, also at generic and iteratively updated values of the weights.

For each of these two training operations, nearly the entire computing budget is in turn spent on **floating-point operations** (FLOPs) used in **dynamic matrix multiplication**, meaning matrix multiplication with both operands determined at

¹³ Demand for these training operations is the rule not only for stochastic gradient descent (SGD), currently the industry-standard weight-updating algorithm, but also for RL rollouts and for alternative training algorithms such as zeroth order or evolutionary methods, which typically require more forward passes at generic weight values than SGD.

runtime (see Austin et al. 2025, for a breakdown of the floating-point operations required for transformers’ forward and backward passes). We will use the term **training FLOPs** to mean dynamic matrix multiplication FLOPs in the service of either of the above two training operations. The **arithmetic throughput** of a hardware system is defined by the number of training FLOPs per second at peak rating.

A commonly used and intuitive unit of hardware arithmetic throughput is the **H100-equivalent (H100e)**, taken as 989 TFLOP/s (9.89×10^{14} FLOP/s, or 3.1×10^{22} FLOPs per year),¹⁴ the peak rated arithmetic throughput of an NVIDIA H100 accelerator.¹⁵ A hardware system’s realized arithmetic throughput on a given task is generally below its peak rate, e.g., due to memory and communication bottlenecks; the **utilization** is defined as the realized arithmetic throughput divided by the peak throughput.

We analogously define a hardware system’s inference capacity with respect to a given model in terms of **inference-H100e**, which we define as its peak rated inference throughput, in tokens per second at a standardized context length, divided by the peak rated token throughput of a standardized general-purpose accelerator cluster, times the H100e of that cluster. The inference-H100e of an inference-only system is defined as its maximum inference-H100e over the models it could serve.

A hardwired pause treaty could then define a **training-capable accelerator** as any hardware system capable of performing training FLOPs at an arithmetic throughput rate above a threshold determined by treaty parties and revised over time as needed. An **inference-only accelerator**, by comparison, could be defined as a hardware system that can perform inference on at least one whitelisted model above some treaty-specified rate but whose rating for training operations is bounded in a suitable way. We discuss ways of building inference-only accelerators in Section 4.2 and Section 4.3.

We define an **idealized inference-only machine** as one that can perform inference for a restricted set of models but is incapable of performing any other operation.¹⁶ In practice, realized inference-only accelerators may have some residual training capacity. Like other classes of accelerator hardware, they

¹⁴ Except where otherwise specified, all hardware ratings refer to dense 16-bit dynamic matrix multiplication with 32-bit accumulate.

¹⁵ Nvidia’s H100 Product Specifications page (NVIDIA n.d.) lists the H100 SXM at 1,979 TFLOP/s for BFLOAT16 and FP16 Tensor Core operations “with sparsity”. Halving that number for dense operation yields approximately 989 TFLOP/s.

¹⁶ For a set W of models, we define a **W-inference-only machine** as a machine that allows the user to perform only three operations: (1) Write the name of a model $M \in W$ on the input tape, followed by a sequence of input tokens; (2) Perform inference using model M to generate new tokens on the tape; and (3) Erase the tape. A conceptual demonstration of an

would be rated to arrive at an inspector-certified bound on their peak arithmetic throughput for training operations. These bounds would need to pay careful attention to the possibility of unintended use, including by a state-level adversary with physical ownership of chips; see Section 4.1.

Because the governance purpose of inference-only accelerators is to satisfy inference demand without allowing for model training, we suggest operationalizing an inference-only rating by bounding a hardware system’s **residual training factor**, defined as (a bound on) its peak arithmetic throughput for training operations in H100e divided by its peak licensed inference rating in inference-H100e. For example, replacing a pre-pause training-capable accelerator fleet with an equally capable inference-only fleet whose average residual training factor is $\phi = 0.001$ (weighted by inference-H100e) would reduce that fleet’s training capacity by 99.9%.

The maximum residual training factor could start relatively high in order to allow early designs to speed the transitional period, and it could decrease over time as hardware specialization techniques mature and lower factors become available. As a simple approach, we suggest states parties could simply agree on a training-capacity rating, cap the total number of H100e units licensed per inference provider and host state, and let market pressure drive residual training factors down as inference demand rises.

3.2.2 Quantitative training thresholds

In order to operationalize treaty objectives, states parties would need to set quantitative thresholds reflecting estimates of the compute budget, measured in training FLOPs, required to train models with treaty-relevant capability levels, as well as estimates of how large a hardware fleet would be required to research and train such models within a relevant time horizon. These would be negotiated by parties in accordance with prevailing technological and political conditions, and adjusted as needed. Here we outline how such thresholds and estimations might be defined, specify these quantities based on an analysis of present-day models, and introduce relevant terminology and notation.

We define the **pause date** to be the moment when frontier training is paused. Treaty thresholds could be defined relative to the **frontier threshold** C_F , an estimate of the total compute required to train the most capable models at the frozen frontier. Treaty compute thresholds would need to be adjusted over time in order to adapt to **training efficiency gains**,¹⁷ improvements in algorithms

inference-only machine for a small language model realized in a consumer FPGA is available at <https://github.com/Berkeley-AI-Risk/inference-only>.

¹⁷ Training efficiency gains are commonly referred to in the machine learning literature as “algorithmic progress” or “compute efficiency gains”; we refer to them as “training efficiency

and data that make it possible to train models at a fixed capability level with shrinking training budgets. For example, we expect the frontier threshold C_F would not be a constant but a function in time $C_F(t)$, where t represents the time in years since the pause date; e.g., $t = 1$ represents one year after the pause, and $C_F(1)$ is the training budget required, one year after the pause, to train a new model with capabilities similar to those of the frozen frontier models.

Besides preventing nation states from training strategically relevant or strategically decisive models, a goal of the pause would be to discourage all actors from attempting to advance the frontier, since such attempts could accelerate training efficiency gains for frontier-scale models.¹⁸ We define a **beyond-threshold model** as any general-purpose generative AI model¹⁹ plausibly powerful enough to compete with models at the frozen frontier. Training and use of such models would be prohibited.²⁰ We operationalize this definition as including exactly those models surpassing either a **compute ceiling** C_T bounding the total number of FLOPs used to train it, including both pretraining and post-training, or a **total parameter ceiling** N_T bounding its parameter count.²¹ These two thresholds play different enforcement roles: while a parameter ceiling is more enforceable after the fact, training FLOPs are easier to deny to evaders before the fact. Since training budgets are imperfect proxies for model capability, preventing the training of new frontier models would require leaving a wide margin between the compute ceiling C_T and the frontier threshold C_F .

gains” to avoid conflating them with other kinds of gains in computing efficiency, such as efficient inference service per dollar or per unit of power.

¹⁸ We define “the frontier” in terms of models rather than systems. We assume there will be attempts to build more capable systems using existing models. Our work does not discuss restrictions on such attempts, although that topic is also worthy of consideration.

¹⁹ By a general-purpose generative AI model, we include at least all deep learning architectures for autoregressive language models that generate language tokens from token-stream inputs. Our discussion is specialized to that class of architectures, but we imagine that governance of other deep generative models such as image and video generation models or robotic models might be tackled with similar mechanisms. We imagine that traditional, smaller-scale statistical prediction and machine learning models would not be covered even when they are notionally infinite-dimensional.

²⁰ Models approaching the treaty compute ceiling or parameter ceiling (for example, models with more than N_L total parameters, for some licensing floor $N_L < N_T$) could require declaration to treaty authorities, with graduated penalties for noncompliance.

²¹ By industry-standard Chinchilla scaling, an N -parameter model is typically trained with $C = 120 \cdot N^2$ FLOPs, giving a natural mapping between the two ceilings (Hoffmann et al. 2022). However, small commercial models are typically overtrained for faster inference performance (i.e., trained using more tokens per parameter than Chinchilla scaling implies); for example, Llama3-8B was trained with 15T tokens, roughly two orders of magnitude more than Chinchilla-optimal 160B tokens (Llama Team 2024). Assuming 100x overtraining gives $C = 12,000 \cdot N^2$, giving one simple option for operationalizing the relationship between C_T and N_T .

Table 3.1. Operational thresholds. Definitions of key governance thresholds, including illustrative estimates of what their initial values might be if a pause took effect at year-end 2026, as derived in Section 6.2. Thresholds would be negotiated by treaty parties and adjusted over time as needed.

Variable	Meaning	Est. (year-end 2026)
C_F	Training budget for frozen frontier	2×10^{27} FLOPs
C_T	Treaty compute ceiling	2×10^{24} FLOPs
N_T	Treaty parameter ceiling	13B parameters
C_E	Training budget for strategically relevant covert evasion	2×10^{28} FLOPs
C_B	Training budget for strategically decisive overt breakout	2×10^{29} FLOPs
F_E	Fleet threshold for covert evasion at present-day utilization	6.4M H100e
F_B	Fleet threshold for overt breakout at present-day utilization	180M H100e

We operationalize the two state-level threat models described in Section 3.1.1 by defining a **covert evasion training threshold** C_E and **overt breakout training threshold** C_B , estimates of the total training budget required to train a strategically relevant model or a strategically decisive model, respectively. Like C_F and C_T , these are measured as FLOPs required for both pretraining and post-training in the final training run for the model and would be adjusted over time to account for training efficiency gains.

Frontier AI companies' research programs to build new frontier models require significant ongoing research and development processes beyond the final training runs; per recent estimates, current frontier companies' research and development budgets are ten times as large as their model training budgets (Denain and Wu 2026). We account for these additional costs via a **research overhead** factor $\rho > 1$, whose magnitude may depend on how ambitiously a research program is attempting to advance the frontier. After accounting for research overhead, we arrive at an **evasion fleet threshold** F_E and **breakout fleet threshold** F_B , estimates of the fleet sizes that would be required to build each model, if these programs attained utilization similar to that attained by present-day industrial frontier training. These fleet thresholds are measured in H100e units, and both are defined relative to **strategic time horizons** T_E and T_B within which an evader is attempting to train a model; a larger fleet is required to break out within $T_B = 1$

year than to break out within $T_B = 3$ years. Then the evasion fleet threshold is defined as:

$$F_E(t) = \frac{\rho_E \cdot C_E(t)}{T_E \cdot U_F},$$

where U_F represents a standardized utilization attained by present-day frontier AI companies. The breakout fleet threshold is defined analogously.

Choices regarding the levels of the treaty thresholds C_T and N_T would need to balance the value of permitting sufficiently sub-frontier model development for commercial or recreational use against the likelihood that prolific training of near-frontier models would drive training efficiency gains, eroding the reliability and durability of the treaty. Models trained for beneficial scientific purposes such as medical research could potentially be granted exemptions; see the discussion of scientific preserves in Section 3.3.1.

3.2.3 Verification and breakout accounting

The training budget for an illicit model could come from a number of different sources including training-capable hardware licensed for transitional inference service, dark compute, consumer **graphics processing units** (GPUs), etc. To guard against covert evasion and overt breakout threat models, the treaty would require a verification methodology to ensure that no state's total hardware fleet, aggregated across all sources, contains enough training capacity to pose either a covert evasion threat or an overt breakout threat.

We explore one relatively simple verification methodology based on estimating, for each possible compute source in each state, the total arithmetic throughput that could plausibly be marshaled toward covert evasion or overt breakout. For compute source c and state s , we define the **covert evasion capacity** $K_E(c, s)$ as the total arithmetic throughput from source c that state s could plausibly marshal toward a covert evasion program, and the **overt breakout capacity** $K_B(c, s)$ as the throughput it could marshal toward an overt breakout program.

These capacities would be measured in H100e and discounted by estimates of how their utilization rate in each scenario would compare to frontier data-center utilization. The same compute source might contribute more to one violation scenario than the other: for example, an internationally monitored, sealed strategic reserve of 100k H100e units of powered-off hardware would contribute roughly 100k H100e to a state's overt breakout capacity, but it would contribute essentially nothing to its covert evasion capacity. In Section 6.4.1 we discuss how to estimate these capacities by discounting the raw capacity available in a given category.

Treaty authorities could certify that each state’s total covert evasion capacity, which we define as its **covert evasion ledger** L_E , were held below the covert evasion fleet threshold times a fractional **covert evasion allowance** $\alpha_E < 1$ to account for uncertainty; that is:

$$L_E \leq \alpha_E \cdot F_E, \quad \text{where} \quad L_E(s) = \sum_c K_E(c, s).$$

Authorities could likewise calculate an **overt breakout ledger** $L_B(s)$ by summing over $K_B(c, s)$ and certifying that it is below the breakout fleet threshold by an **overt breakout allowance**, i.e., $L_B \leq \alpha_B \cdot F_B$.²²

The treaty authority’s primary lever for balancing these two inequalities for each state would be limiting the total number of licenses granted to each state for various uses of accelerators such as transitional inference service or other uses discussed in Section 3.3.1. In this way, each term in a given ledger could be adjusted until the inequality is satisfied, with the important exception of the term for dark compute. Because dark compute would by definition be ungoverned under the treaty, its estimated covert evasion or overt breakout capacity would need to be reduced through other measures such as discovery and registration of pre-pause stock, prevention of illicit manufacture, and verification measures that rely on detecting when such dark compute is being used.

Because issuance of licenses would depend on satisfying these inequalities, states desiring more licenses would have an incentive to keep their ledgers clean, for example by limiting the rate at which dark compute accumulates in their state, or by agreeing to allow more inspector access to assure other states that a given compute source could contribute very little to covert evasion. Thus, verification procedures would not need to be one-size-fits-all: they could be flexibly negotiated based on each state’s willingness to make concessions in each context, with larger compute sources requiring more rigorous verification measures than smaller ones.

We note that building a strategically relevant or strategically decisive model is only the first hurdle an evader would need to clear. In order to realize any advantage from training, the evader must also be able to serve inference on that fleet at sufficient scale. Thus, governance that focuses on prevention of training is *sufficient* to prevent violation but may not be absolutely *necessary*. For example, if states developed reliable national technical means for detecting the use of

²² In its actual implementation, there might be per-coalition allowances in addition to per-state allowances, and allowances might vary by state; e.g., Luxembourg might have less allowance than the U.S. or China. The core of our feasibility analysis concerns the U.S. and China, two leading powers that are not members of the same coalition, so it is unlikely to be sensitive to the way these questions are resolved.

unknown language models when they were deployed, it could ease the degree of reliance on verification based on denial of training capability.

3.2.4 *Model whitelisting*

Beyond enforcing a frontier pause, treaty governing authorities would also be responsible for determining the **model whitelist**: the list of beyond-threshold AI models for which inference is permitted and inference-only chips are permitted to be manufactured.²³ By “model,” we mean a fixed probability distribution for generating output tokens conditional on input prompts. Since fine-tuning or other post-training that alters or adapts model weights would alter the output distribution, a hardwired pause would not permit inference on models altered in these ways except where expressly approved by governing authorities. Examples of approved fine-tunes could include ones that implement knowledge cutoff updates or safety updates.

Treaty governing authorities could coordinate on any international standards regulating the deployment of whitelisted models, in addition to whatever local regulations are in force. For example, a model for which a critical jailbreak is discovered could be de-whitelisted, possibly followed by re-whitelisting contingent on inference providers deploying it with required external constraints such as an input classifier or safety harness. For example, a model that had critical jailbreak vulnerabilities could possibly be judged as unsafe if exposed directly to users but sufficiently safe if it were only prompted by other models.²⁴

During Phase 0 or Phase 1 of a hardwired pause (recall Section 1.1), **AI developers** would submit above-threshold models to be considered for whitelisting by participating states or interim treaty authorities before a defined **whitelist submission deadline**. To foreclose the possibility of further training after the deadline, developers submitting closed-weight models could be required to submit a cryptographic hash of a complete description of the architecture and weights of any model they submit; this hash could then be checked against deployed weights to ensure that the deployed model had not changed since submission. Additional required information for evaluation purposes could include details of the development process including the training algorithm and data.

The whitelist submission deadline would be an important governance instrument facilitating enforcement of the prohibition against above-threshold model

²³ Note that actors might still perform inference illicitly on non-whitelisted models. By design, a hardwired pause limits access to accelerators capable of performing illicit inference, but still, this may not on its own be sufficient for all purposes in all scenarios.

²⁴ U.S. constitutional analysis would depend on the risk addressed and the restrictions imposed - i.e., whether safeguards addressing the identified risk could adequately substitute for excluding the model altogether (Mark and Scher 2025).

training. Under a hardwired pause with a long expected duration, AI developers would have little incentive to unlawfully train new above-threshold models after the deadline, because those models could not be commercially deployed without exposing the violation. By contrast, shorter-duration pauses would likely require stronger enforcement measures to prevent illicit training: by holding out the possibility that noncompliance would be rewarded after a brief pause, they could tempt companies to evade.

At present, there is no agreed-upon operating definition for whether a powerful AI model is safe enough to deploy internally or to release to the public, or methodology for assessing safety. Often in practice, a subjective determination is made by the same AI developer with a pecuniary interest in releasing it. Moreover, safety depends not only on the model but on the technological, social, and institutional systems in which the model is deployed, and safety may change over the lifetime of these systems. As such, whitelisting processes may evolve over time in response to input from improved understanding of AI safety, newly discovered adversarial methods for jailbreaking or eliciting dangerous capabilities, empirical evidence about AI models' effects on society, or input from affected constituencies.²⁵ Developing methods for safety determinations and politically legitimate processes for making them involves pressing technical and political questions outside the scope of this paper.

Whitelisting processes would need to be designed for consistency with states' norms and constitutional restrictions. In particular, U.S. implementation could raise First Amendment questions because restrictions on model development may burden free expression interests, especially if they distinguish among outputs by their subject matter or viewpoint (Volokh et al. 2023; Mark and Scher 2025). Restrictions on hardware and training capacity have a stronger claim to content neutrality, but still need to be justified insofar as they incidentally burden protected expression.

It is possible that pause-willingness (as defined in Section 1.2) could arise only *after* all frontier models have developed capabilities that are widely considered unsafe for deployment at scale or that come to be considered unsafe in the years after a pause. In that case, the effect of a well-considered whitelisting process could be to “rewind” the frontier (as opposed to pausing it) if the most capable whitelisted models had all been trained significantly *before* the pause date and were all less capable than the pause-date frontier. In that case, it would likely be effectively impossible to ensure that non-whitelisted models were not being

²⁵ To reduce the risk of arbitrary or discriminatory licensing, authorities could, e.g. follow published criteria and procedures; provide reasons for whitelisting decisions; decide applications within clearly defined periods; and provide timely review, including judicial review under domestic law.

used covertly, especially by states. The frontier threshold C_F would effectively be reduced, and other treaty thresholds derived from it such as C_T and C_E (as in Section 6.2) would need to be correspondingly reduced. Thus, while it is conceivable that taking such a step might be necessary, doing so could make the treaty harder to enforce and verify, especially because the covert evasion fleet threshold F_E would likely be reduced.

3.3 Hardware governance

To verify that the technical specifications from Section 3.2 are met, states parties would need to govern computer hardware, especially training-capable accelerators. Here we describe the roles training-capable accelerators would continue to play in a hardwired pause (Section 3.3.1); how accelerators could be located and tracked, and how licensing would operate to facilitate their governance (Section 3.3.2); and approaches to verifying that training accelerators are only used for their intended roles (Section 3.3.3). We conclude by discussing governance of hardware manufacturing technologies and facilities (Section 3.3.4).

3.3.1 Residual uses of training-capable accelerators

While inference-only accelerators would displace the majority of training-capable accelerators after a several-year-long transitional period, training-capable accelerators could remain in limited use during a hardwired pause. This use could fall into the following categories:

First, the residual use that would be largest in scale but shortest in duration would be the interim commercial allocation of training-capable hardware to **transitional inference service** by licensed **inference providers** while the market awaited sufficient inference-only hardware to replace it.

Allocation of training-capable hardware toward inference service would require time- and capacity-bounded reliance on non-hardware-based verification methods as discussed in Section 3.3.3, but it would serve two important governance purposes: First, it would prevent inference supply from collapsing while manufacturing of inference-only hardware ramped up. Second, it would harness burgeoning inference demand to incentivize owners of pre-pause training-capable accelerators to either register their chips under the treaty or sell their chips to licensed hardware operators, or to treaty authorities, which could sell them back to hardware operators at auction. Meanwhile, the frozen model whitelist would eliminate the primary commercial incentive for training new frontier models, so compliance would be the most remunerative option for hardware owners.

Second, each state could be allowed a licensed **sovereign fleet** of training-capable accelerators, with registered and tracked chips, for use on specific

sensitive national security workloads. Its size would be bounded to ensure the state's total allocation of training-capable compute remained below the covert evasion and overt breakout thresholds.

Third, training-capable hardware could be eligible to be placed in internationally-governed **scientific preserves**: data centers operated by treaty authorities and dedicated to common-good scientific applications such as medical research or healthcare provision.

Finally, states could have the option to store licensed **strategic reserves** of powered-off training-capable hardware under seal in monitored facilities within their own territory (cf. Lifland 2026), for example, in sealed bags that are periodically checked. Because these strategic reserves would be recoverable in the event of a treaty exit, they would give states insurance against breakout by rivals. Because these reserves would not count against states' covert evasion ledgers, they would be a cheaper form of insurance than a dark compute stockpile would (except insofar as they could arrange for other states to underestimate how much dark compute they had).

The existing stock of training-capable hardware would be redirected to the above uses. As the training-capable hardware stock shrunk by attrition, new hardware might eventually be required for sovereign fleets and scientific preserves. Specialized hardware to replenish these stocks might look very different from the current accelerator stock, most of which is optimized for AI training. Innovation by the semiconductor industry into hardware specialization and research developments in technical governance might eventually allow the total compute allocated to other uses, such as high-precision scientific simulations and computer graphics, to continue on a rapid growth trajectory even as the total training capacity for AI models dwindles.

3.3.2 Hardware licensing, registration, and tracking

Tracking accelerator hardware would be an essential objective of the treaty. Each hardware system under treaty governance could be placed in the care of a licensed **hardware operator** at a known data center, residing in a **host state**.²⁶ Core licensing, registration, and tracking requirements would apply to all accelerators under the treaty's governance, including sovereign fleets,

²⁶ Once a treaty is ratified, host states would be restricted to states that have agreed to abide by the terms of the treaty (e.g. signatories) but not necessarily restricted to states parties (i.e., states that have formally ratified the treaty). During an interim agreement, arrangements could be based on informal participation.

scientific preserves, transitional inference service, and possibly inference-only service.²⁷

There would be two routes for governance of hardware under the treaty: pre-pause hardware would enter via registration at the beginning of the treaty, while new hardware would enter by registration at the time of its manufacture at a licensed fab. The only approved way for hardware to exit treaty governance would be verified destruction at the end of its life. Hardware lost to the regime would likely be presumed dark compute and charged to host states' covert evasion and overt breakout ledgers at near-full training capacity.

As early as possible, it would be important for governments to undertake a **global accelerator census** attempting, as far as possible, to account for every pre-pause accelerator unit (cf. Aarne and Petrie 2025; Scher et al. 2025, Art. V). This census would serve the twin goals of ensuring that only a small fraction go undeclared and classifying and quantifying the undeclared stock within each state. An accelerator server is a capital good with value in the hundreds of thousands of dollars that leaves documentary evidence in various forms when it is manufactured, sold, imported, operated, and written off—including Chinese customs and tax records of chips diverted around U.S. export controls, which we refer to as **diverted hardware**. See Section 5.4.1 for a list of the many sources of documentary and other evidence about accelerators' chain of custody and whereabouts.

A census could require declaration of accelerators by current hardware owners and compile information from a variety of sources on the location of each chip, including sales records, tax records of hardware purchasers, inventory records of data center operators, power bills, and chip rental records. Governments could supplement these efforts by offering generous buybacks and bounties for tips leading to the discovery of undeclared accelerators, as well as granting amnesty to shell companies and their customers for diversion around U.S. export controls or Chinese customs, conditional on producing retrospective records (cf. Halstead and Larsen 2026). To maximize coverage, the census would need the cooperation of U.S. and Chinese officials, and would need to include resellers and data center operators in jurisdictions outside the U.S. and China including Singapore, Malaysia, Taiwan, and the UAE. We discuss the prospects for success of this chip census in detail in Section 5.4. The eventual successful completion

²⁷ If states parties had sufficient confidence in the hardware-based guarantees, and the residual training factors of inference-only accelerators were low enough to be considered negligible, then inference-only chips might be exempt from tracking. Conversely, if residual training factors were not so low, or states parties were uncertain about the residual training factor bounds, then inference-only chips would need to be tracked. If that were true of some, but not all inference-only accelerators, then the treaty could distinguish between them, tracking the latter category but not the former.

of this census would be an important determinant of durability, as we discuss in Section 6.5.

State requisition and concealment of pre-pause chips would be visible primarily as gaps in the census. If significant residual uncertainty remained about the pre-pause dark compute in a state, parties could arrive at a negotiated estimate based in part on compiling and reconciling the above data, which would factor into the state's covert evasion and overt breakout ledgers. A state's willingness to furnish data on request, to permit inspector visits to commercial data centers, and generally to assure other parties of its accounting transparency would likely be a factor in the negotiated estimate.

Licensed hardware systems would come equipped with (or, in the case of pre-pause hardware, be retrofitted for) on-chip location attestation (e.g., latency-based location Brass and Aarne 2024) to ensure that they remained at their registered location and visible to treaty authorities. Transfer of ownership, location, or jurisdiction would require re-registration. Chips that unexpectedly moved or fell silent would have to be produced for inspection or surrendered to authorities,²⁸ and licensed hardware operators would also be subject to routine inspections. Operators' eligibility for hardware licenses would be contingent on their record of compliance, with host states responsible for enforcement and inspector access.

As a design option to enhance compliance and address some concerns about regulatory capture, we suggest that AI companies (companies submitting whitelisted models) could be ineligible to apply for hardware operation or inference provision licenses, but still allowed to access their own whitelisted models through licensed inference providers. This design option, if combined with a mandate for AI companies to license their models to inference providers at regulated rates,²⁹ could also potentially increase political palatability for constituencies that object to governments granting long-term monopoly rents³⁰ to market leaders at the time the pause takes effect.

²⁸ Under U.S. law, compulsory surrender may require just compensation (U.S. Const. amend. V; *Horne v. Department of Agriculture*, 576 U.S. 351 (2015)). Courts have nevertheless recognized circumstances in which compensation is not required, including certain emergency, necessity, national security or law enforcement measures (see Davidson and Mulvaney 2025). Whether these doctrines would cover compulsory surrender as proposed above would depend on the design and circumstances of implementation. A detailed treatment is outside the scope of this paper.

²⁹ Compulsory licensing could also raise questions related to takings where it involves trade secrets or other forms of intellectual property (Epstein 2004; Kapczynski 2022). Its treatment would depend on the rights involved as well as the terms of access, disclosure and remuneration, *inter alia*.

³⁰ Or duopoly/oligopoly rents, if a small number of AI companies own the most powerful models at the time the treaty is enacted.

If vulnerabilities were discovered in the design of inference-only hardware that opened up jailbreak attacks that could enable model training by evaders, or if the models they implement were de-whitelisted (cf. Section 7.3.4), inference-only licenses would be recallable for retrofit and re-licensing once the jailbreak is patched. The model, and hardware serving it, might be re-whitelisted contingent on additional external safety features that inference providers would be required to implement.

3.3.3 Sociotechnical verification measures for preventing illicit use of residual training-capable accelerators

At least until fully general privacy-preserving verification tools are ready for deployment, verification for training-capable hardware would need to lean on physical or institutional restrictions.

For transitional inference service on training-capable accelerators, verification could involve regulating the interface between hardware operators and inference providers. Hardware inside training-capable data centers could be exposed to inference providers as an idealized inference-only machine (recall Section 3.2.1) through an application programming interface (API). Inference providers would submit a request including a choice of model, metadata such as prompt length, and an encrypted model input, and operators would return encrypted outputs along with other metadata such as output length. All encrypted traffic from outside the data center would pass through an internationally-operated router machine to small, physically isolated clusters within the data center and connected only to the router. Similar to the proposal by Dean (2026b), router traffic could be randomly sampled and sent to a secure recomputation server that would confirm the input matched the output without exposing the decrypted prompt to treaty authorities. Inference traffic and power usage would be reported per cluster and reconciled to power station and inference provider records. As defense in depth, these measures could be supplemented with telemetry-based workload classification such as that described by Rahman and Tajdari (2026) or zero-knowledge proof based attestation methods (e.g., those offered by Attestable n.d.).

Inspection protocols could be modeled on precedents from the Chemical Weapons Convention to ensure compliance with the laws of local jurisdictions. For example, regulations in the United States cover inspections with consent (no warrant required), routine inspections without consent (requires administrative search warrant), and challenge inspections (requires criminal warrant with probable cause).

For the internationally governed scientific preserves, verification could be achieved through workload transparency and social accountability. We envision that all workloads would be declared by named, vetted participants, and could be logged and replayable on demand by treaty authorities. Workloads could be reviewable by the general public insofar as restrictions such as confidentiality for medical and human subject data allowed.

For the sovereign fleet, verification procedures would be negotiated between parties on a case-by-case basis. Owing to the small total size of these fleets, the treaty ledger approach above would not necessarily require ruling out that any portion of their computing power were being used for training, if it were not possible to provide such verification without compromising important national security interests. Nevertheless, states would be charged less against their covert evasion ledger if they could show other parties some evidence that most of the sovereign fleet was not being used for illicit training. Technical measures such as power usage monitoring, inspection of network topology, or workload classification through telemetry metadata, or institutional measures such as whistleblower protections (Baker et al. 2025), might allow for some discounting of covert evasion capacity, even if the sensitivity of these workloads did not allow for airtight verification using present technology.

Finally, for strategic reserves, verification could rely on periodic inspection and counting to check that hardware remains under seal in its purported location.

As verification technologies mature over time, supported by implementation in new hardware-software platforms, the treaty authorities could tighten their practices further, achieving higher confidence that governed training-capable hardware is not being used for unauthorized purposes, and thereby allowing for more licenses to be issued. Two technologies that could play a key role in verification are trusted execution environments (TEEs) and formal methods. AI computations permitted in various contexts including transitional inference services, scientific preserves, or sovereign fleets could be accompanied by a proof that they adhere to agreed-upon safety rules and/or monitored for such adherence. Such a proof is not just restricted to the strict mathematical notion of proof, but more broadly includes the kind of evidence and assurance cases that are common in certification of safety-critical systems and safety engineering (Leveson 2012). The classical idea of “proof-carrying code” (Necula and Lee 1996) supported by advances in automated formal methods for Verified AI (Seshia et al. 2022), offers a mechanism: designers of permitted AI compute should accompany their deployed models with an independently checkable proof of safety that can be vetted by a trusted third party. Additionally, treaty authorities should be able to deploy runtime monitors within TEEs that are tamper proof. These monitors check for model adherence to safety rules and

should send regular “heartbeat” updates to treaty monitoring servers indicating such adherence. Formal verification and systematic testing should be applied to the TEE platforms to ensure that these are free of bugs and vulnerabilities. The first step for such verification, having a comprehensive formal specification of TEE platforms (e.g. Subramanyan et al. 2017), is already available today and offers a foundation upon which advances in AI-assisted formal verification can achieve a higher level of assurance.

Another approach could be applied to existing hardware: non-invasive verification, which infers what a chip is computing from external physical signals (the power usage monitoring mentioned above for sovereign fleets is a simple example). Many possible techniques for this are already well-developed and widely deployed for the purpose of studying and waging side-channel cybersecurity attacks. Rather than verifying compliance, such measures are used to infer or collect protected information from a security system based on emissions from physical side channels; for example, power consumption, timing, or electromagnetic emission measurements are used to deduce cryptographic keys or secret messages without requiring specific knowledge about the workloads or operations. However, because non-invasive verification schemes and side-channel attacks rely on inferring a system’s internal operations from indirect, external signals, mature side-channel attack methods and toolkits could conceivably be repurposed to support verification techniques.

3.3.4 Governing hardware production

The regime would need to govern and monitor hardware production well enough to enforce the mandate that training-capable accelerator production be halted during Phase 1,³¹ while production continued on all other types of hardware including inference-only accelerators, consumer graphics cards below a treaty-governed performance bound, CPUs, and other hardware systems.

The regime would need to prevent illicit production of training-capable accelerators through two channels: production at undeclared sites, and diversion of production at declared manufacturing sites (cf. Baker 2023, App. G; Shavit 2023; Al Al Ramiah et al. 2025; Scher et al. 2025, Art. VI).

To prevent production at undeclared sites, the first line of defense could be to conduct a **global lithography machine census** (which would complement the proposed accelerator census), as well as the use of all-source intelligence to detect attempts to construct and operate such facilities covertly. In doing

³¹ The treatment of restrictions on production or use as “regulatory takings” under U.S. law generally depends on their economic impact, interference with reasonable investment-backed expectations, and the “character” of the government action (Echeverria 2005).

so, the treaty would need to account for each pre-existing unit of the world's advanced lithography machines on which accelerator chips are printed, including at least every extreme ultraviolet (EUV) and ArF-immersion deep ultraviolet (DUV) lithography machine. These could be registered and tracked, along with other necessary tools such as excimer lasers, immersion tracks, and *e*-beam mask writers, with location attestation according to procedures similar to those used for training-capable accelerators, as well as verified physical destruction at the end of their lives. Registered machine counts and serial numbers could be checked against a census of vendor delivery records, and any unaccounted-for machine could be charged to host states' ledgers as continuously producing dark compute until they were located and accounted for. Factories capable of producing lithography machines could be monitored to ensure that new machines were registered at the time of manufacture, and registered lithography machines could send cryptographically attested records of each processed wafer to treaty authorities.³²

Satellite imagery could be used to locate and enumerate the finite set of existing facilities sufficiently large and well-equipped with HVAC to host clean rooms for undeclared machines. Possible methods to confirm these sites had no undeclared machines could include environmental monitoring for offgassing, power substation metering, walk-through inspections to confirm buildings held no undeclared clean rooms, and all-source intelligence. Similar procedures could be used to rule out undeclared lithography machines at declared facilities.

At declared manufacturing facilities, monitors would need to verify both that declared machines were not producing undeclared chips and that declared chips leaving the facility were manufactured to specification or verifiably destroyed. The former objective could be achieved by reconciling the number of silicon wafers coming in, the number of logged wafer prints, and the number of chips coming out, while the latter could be achieved by randomly sampling a small number of outgoing chips for non-destructive or destructive inspection and comparison to licensed hardware specifications.

As defense in depth, mass balance accounting of perishable consumables like photoresist chemicals could prevent skimming in the service of undeclared production; each additional tracked input would mean an evading state would need to add another node to its covert supply chain without detection.

³² Detailed logs are already generated by the vendor software and remotely accessible by vendors for maintenance service (Hermans 2023; Baazil et al. 2024).

3.4 Transition governance

In this section, we discuss governance during Phase 1 of the hardwired pause, when the pause begins, and Phase 2, when training-capable accelerators are replaced with inference-only accelerators.

3.4.1 Economic alignment

For a hardwired pause to be effective, negotiators would need to design mechanisms that align hardware designers, AI companies, and inference providers with the spirit and letter of the regime requirements, in consultation with economic experts. Here we discuss one possible implementation.

In order to ensure that the initiation of the regime would not create a negative shock to the global inference supply, parties to the treaty would approve a provisional whitelist of both models and hardware, while they deliberated on how to implement updated processes going forward. Continuing diffusion of AI models could mean strong ongoing growth in inference demand; Emberson and Sevilla (2026) estimate current inference demand at between 200M and 4B tokens/s, growing at a tenfold rate per year. After an initial rise in supply due to reallocation of compute resources from training to inference at the start of Phase 1, the global inference supply could struggle to keep up with demand while the market awaited inference-only hardware.

During this interim period, market forces could exert strong pressure on hardware companies to rapidly ramp inference-only production; the prospect of a windfall from being the first hardware supplier to begin producing inference-only chips at scale might create race conditions favoring the treaty. An interim gap between inference supply and demand could likewise create a lucrative inference market for suppliers and a high opportunity cost for stockpiling undeclared hardware.

Effective treaty governance would exploit these market incentives by converting them into the treaty's most important needs: visibility into, and eventually surrender of, training-capable hardware. In particular, hardware owners would need to declare and register their hardware immediately if they wanted to continue using it for lucrative inference service throughout the interim period, with late declaration carrying a financial penalty and undeclared accelerators subject to seizure and resale by treaty authorities,³³ who could use proceeds to fund generous bounties for help in discovery.

³³ Valid forfeitures may be exempt from takings compensation - however, the validity of the forfeiture regime and its compliance with due-process and excessive-fines requirements would need to be assessed separately (Davidson and Mulvaney 2025; Saxer 2025).

3.4.2 Phase 1 (design)

We consider Phase 1 to begin as soon as the U.S. and China agree to a pause minimally implementing the two core mandates against training new frontier AI models and manufacturing new training-capable accelerator chips; however, these two decisions might not occur simultaneously and might go into force before a formalized agreement is fully negotiated.

During Phase 1, the three negative mandates would become operative in all participating states:

1. Training of beyond-threshold models, using any hardware, would be prohibited (see Section 3.2.4).
2. Use of existing accelerator hardware would be prohibited for any use other than licensed inference on whitelisted models or designated exceptional uses such as those in Section 3.3.1.
3. Manufacturers would be directed to stop production of accelerator hardware not included on the provisional whitelist of inference-only hardware.

International verification and domestic verification and enforcement measures in support of these mandates would also be ramped up quickly at the outset of Phase 1. The global accelerator census would also begin, and all training-capable AI chips performing transitional inference service would be registered under treaty governance.

In addition to approving the provisional model whitelist (which could initially include all submitted models pending further evaluation), parties or interim authorities would evaluate and approve a whitelist of submitted hardware designs. In addition, hardware owners would declare and register any accelerator hardware that they wished to continue operating under the treaty, and it would be retrofitted for location attestation. Failure to comply would incur graduated legal penalties including seizure upon discovery of undeclared hardware; late compliance would entail financial penalties and lost profits.

If the U.S. and China wished to halt frontier training on an urgent timeline, they could do so in short order³⁴ by directing the several leading companies in each of their states to halt training.³⁵ They could also stop the production of

³⁴ In the U.S., existing authorities may support some interim measures (such as Defense Production Act's priorities and allocations powers and its provisions for government-approved voluntary agreements); Bullock et al. (2024) discuss possible applications and limitations. The comprehensive framework proposed here assumes new federal legislation.

³⁵ At the time of writing, the top 100 models in the Epoch Capability Index leaderboard were all built by seven American companies (OpenAI, Anthropic, Google DeepMind, xAI, Meta, Thinking Machines, and NVIDIA) or Chinese companies (Moonshot, Alibaba, DeepSeek, Zhipu, and Minimax) (Epoch AI 2025a).

accelerator hardware by forbidding additional orders for accelerator chips by American and Chinese semiconductor companies,³⁶ which comprise virtually the entire global market, and in the meantime recruit additional U.S. allies such as the EU, Japan, and South Korea, which control supply chain chokepoints (cf. Section 5.2).

To reduce AI companies' commercial incentive to continue illicit training, the U.S. and China could also set a short-term model whitelist submission deadline to pass before negotiations were complete. Adding the other supplier monopoly states would buy more time until it were possible to induce most other states to join.

3.4.3 Phase 2 (substitution)

During the managed substitution phase (Phase 2), transitional training-capable inference service would be gradually displaced by inference-only compute, at a rate determined by the inference market, the rate of industrial production of inference-only hardware, and/or the states parties.³⁷

To ensure the timely retirement of training-capable inference service once it is no longer needed, there would also be a cap on the number of hardware units, measured in H100e, for which licenses are available at any given time. Firms wishing to continue serving inference on pre-pause training-capable hardware would bid for licenses at a periodic auction, licenses would expire at the time of the next auction, and the number of licenses available at each auction would decrease over time according to a schedule based on the availability of new inference-only hardware. Proceeds from this auction could fund treaty authorities or buybacks for pre-pause accelerators. Hardware operators would be required to trade in training-capable accelerators upon license expiry.³⁸ If certain classes of models are expected to be underserved by inference-only compute for longer, the license cap schedule could apply on a per-model-class basis.

3.4.4 Traded-in training-capable compute

The rate at which training-capable compute is withdrawn from inference service would depend on the rate at which inference-only compute is produced. Our empirical analysis in Section 4.4 suggests that if production of inference-only

³⁶ American chip designers NVIDIA, Google, Amazon, and AMD, with Chinese chip designers Huawei and Cambricon, account for all accelerator chips tracked in Epoch AI's AI Chip Sales and AI Chip Owners data (Epoch AI 2026c,d).

³⁷ As part of treaty negotiations, we expect states parties would also discuss whether and how inference-only chips would be distributed across national borders.

³⁸ See *supra* note 34.

compute follows a similar design schedule as recent inference-leaning chips, the duration of interim inference-only service on training-capable compute could be less than five years, but the actual duration might be somewhat faster or slower.

In Section 3.3.3 we discussed workload verification for training-capable hardware, including ongoing progress in TEE and formal verification methods. It is possible that before inference-only production ramps up enough to replace training-capable inference service, computer scientists will develop secure ways to retrofit training-capable hardware so that it can remain in inference service for the remainder of its lifespan without significant evasion risk (cf. Scher and Thiergart 2024; Petrie et al. 2025). In that case, traded-in chips could be retrofitted and returned to transitional inference service, if their remaining profitable lifespan justified the expense.

Unless the duration is significantly longer than expected, the training-capable accelerators retired from inference service will be aging but still functional. If it is not possible, or not profitable to retrofit them for verified service, they could be retired and sent to sovereign fleets or scientific preserves, or destroyed.

3.5 Possible exit conditions

In this section we consider the conditions under which a pause might be exited by states parties. A hardwired pause could take effect for an initial time period, for example ten years, with mandated renewals, through a process managed by states parties or treaty-implementing authorities (Finke 2026).

There is currently no agreed-upon paradigm for deciding whether, or under what conditions, building more powerful AI models would be broadly safe or beneficial for humanity at large. Any decision to resume training more powerful AI models after a pause would be informed by scientific advancements, social and economic developments, and political value judgments, all of which are impossible to predict when humanity's experience with AI systems has been so brief and so limited to a relatively small fraction of people across the globe. Even technical alignment is still in many ways an open scientific problem (Bengio 2026).

As a result, it would be premature to name technical conditions under which a pause might be lifted, but a pause would need to clearly specify procedural steps for any exit condition. We can envision several possible conditions for exiting a hardwired pause:

First, states parties might mutually decide to dissolve the treaty because they mutually agree that humanity has discovered satisfactory solutions for the constellation of sociopolitical and technical problems that originally led to pause-willingness. Such a solution slate could include technical AI safety procedures (cf. Scher et al. 2025), economic or labor market mitigations, and social and/or

political innovations that ensure human flourishing, equality, liberty, and safety under plausible AI futures. It might also include a plan for a collaborative international project to develop more powerful AI models without the competitive race dynamics characteristic of the present AI race (cf. Scher et al. 2025; Larsen et al. 2026).

Second, the governance paradigm of the hardwired pause might definitively fail, triggering a condition for dissolution. For example, this could happen because technological advances favoring evaders render treaty mechanisms for monitoring, verification, and enforcement obsolete, and new mechanisms cannot be found.

Third, sufficiently severe departures or violations could trigger an exit to the treaty. A major review or termination of the agreement could follow potentially “treaty busting” developments, such as the orderly departure or treaty-violating defection of a sufficiently powerful state.

Alternatively, if the first decade of a hardwired pause proved prosperous for the world, leaders may decide to renew the pause via a renewal procedure determined during initial negotiations.

3.6 Institutional governance

A fundamental issue for a hardwired pause will be how well the treaty is implemented: first by the initial parties, then by the interim authorities, and eventually by the treaty authorities, all states parties, and impacted industries. Historically, treaty authorities and the member states of the treaty—the states parties—share a responsibility to effectively govern international agreements. A hardwired pause would be no different. Thus, it will be important to make wise choices regarding how institutional governance can best support treaty implementation. Previously in Section 3.1 we described how a hardwired pause treaty organization may evolve and fully emerge over the four phases described in Section 1.1. This section reviews some of the main choices for institutional design that could enhance governance of the hardwired pause.

Design choices for how the treaty will be governed legitimately through the four phases of institutional evolution are core to hardwired pause feasibility. Treaty organization leadership and staff will require the highest standards of competency, probity, and independence. The mandate and scope of the hardwired pause establishes explicit and implicit requirements for a future institution, one composed of treaty authorities and most likely some type of forum for regular deliberations by states parties such as those listed in Section 8.2. Functions such as management of an international ledger of compute, prohibited chip reclamation and adjudication, administration of whitelist process decisions,

monitoring states parties' activities under the treaty, and deploying inspectors to verify treaty provisions all incur institutional implications and costs.

Beyond these foreseeable requirements, lessons from existing international organizations and historical international agreements provide additional insights for effective institutional governance. The potentially evolving membership of the treaty, especially in the four phases, will shape the formulation of any interim authority bodies and the eventually established treaty organization. Early clarity for prospective states parties on the benefits of joining the treaty, especially for countries with less developed AI capacity or economic advancement could generate faster cooperation. This implies that even interim authority design will need to have some capacity to manage the benefits sharing (see Section 3.1) or facilitate technical cooperation. For example, the envisioned benefits of the scientific preserves (Section 3.3.1) entail staffing and management, institutional processes for reviewing scientific cooperation, and data center operations controls and administration. Given the potentially significant costs to some states if they remain outside the treaty, procedures and staffing may be needed for managing hardwired pause implications on external parties.

Careful formulation of the legal terms for a hardwired pause point towards a binding treaty with a range of precise commitments, likely necessitating experienced legal expertise within the early ranks of the treaty authority formulation. There are many proposed obligations of states joining the hardwired pause, from implementing prohibitions to supporting international capacity building and potentially shared AI scientific or safety research. Many of those obligations could be more effectively upheld if there were treaty authorities and personnel serving in coordinating or facilitating roles between states parties.

With the passage of time and projected advances in technology and AI development methods, the adaptivity of the treaty's provisions needs up-front design as well. For example, treaty authorities will benefit from flexible procedures for working with states parties to account for advances in algorithmic progress or other treaty relevant updates. This may be helped by the formulation of various technical committees that regularly review these issues and nominate adaptive measures or protocols to states parties, perhaps mirroring how the International Civil Aviation Organization (ICAO) governs for adaptivity (Chicago Convention 1944). One issue that could require adaptivity involves how prohibited thresholds would be adjusted (e.g., predetermined schedule, renegotiation at regular conventions, or delegation to treaty authorities). Each implementation would have its own pros and cons, and negotiators of the treaty would need to find a mutually agreeable arrangement for threshold determinations.

How the treaty authorities contribute to delivering consistent evidence of hardwired pause compliance, or how they operate when treaty obligations are

being violated, will require significant institutional design and accounting. For example, treaty authorities will require sufficient human resources, data processing and analysis, and accounting procedures to manage the compute ledger with fairness, accessibility, and transparency. This is not unlike design choices the Preparatory Commission for the Comprehensive Test Ban Treaty Organization (CTBTO) has had to make in manning and operating their global International Monitoring System (IMS) and associated International Data Center (IDC). The hardwired pause treaty authorities could require similar infrastructure to bolster the regime's credibility. Additionally, we predict the need for a substantial international inspection process to verify a number of the treaty's provisions. How that inspection cadre is recruited, trained, deployed, and retained implies a potentially large standing unit of inspectors, support personnel, and associated technical verification equipment or procedures. Institutional forums and procedures for reviewing and making political judgments on monitoring and verification will also need to be designed into the treaty organization, with appropriate methods for enforcing treaty provisions under the treaty procedures, or externally. As described in Section 3.4, treaty authorities will be heavily involved in enforcing compliance with industry declarations (e.g. to support the global accelerator census and to qualify companies for licenses to use training capable chips for inference) and collection of relevant fines, likely requiring a specialized unit and financial system for operationalizing those treaty provisions.

A treaty designed with clear decision rules and delegated powers for the treaty authorities will impact states parties' views on feasibility, especially when weighed against needed financial resources and staffing of the treaty authority (commonly referred to as a technical secretariat or staff). Funding for the treaty organization could come from a mixture of funds such as transitional licensing fees, compliance fines, and contributions from states parties. We anticipate these funding mechanisms may be sufficient for full-scale operations until fees from transitional licensing decline. The actual costs to institutionally implement a hardwired pause and how to sustain longer term financing require further analysis. As for the size and capabilities of the treaty authority staff, we envision that the potentially large requirements will necessitate early focus by states parties, and significant effort to recruit and prepare these personnel for their duties. Lessons from the EU AI Office are instructive in this regard. While the office was established in legal form in 2024, capacity for the office to fully implement its legal mandate and enforcement duties remains behind schedule. Ideally, a hardwired pause treaty organization is developed with sufficient human talent and on appropriate timelines.

Design choices for how non-state party actors such as civil society, or AI industry representatives, are included within the formal or technical institutional

structures of the treaty organization will also be important to further develop going forward. As detailed in multiple sections of the paper, the plausibility and durability of the treaty, especially in Phase 1, will hinge on how the AI industry is engaged institutionally by treaty authorities. This implies the need for industry councils or representational processes with treaty authorities and states parties, with both formal and informal levels of involvement. Given AI's expanding political and social impacts, more structured participation by civil society and independent experts may also be helpful for durability.

The hardwired pause treaty's legitimacy will rest on calibrating a mix of state party managed processes and accountability measures such as the compute ledgers and envisioned licensing procedures. Early development (and enactment) of domestic implementing legislation for the treaty's mandate and obligations will be important in this regard. A related component of legitimacy will be how well the treaty authorities and states parties are secured and defended in the course of their work. Given the importance of the treaty's institutions and span of control and monitoring over global AI compute and infrastructure, we assess that well defended physical, human, and digital infrastructure for the regime is a fundamental aspect of feasibility—especially durability. Unlike the pure human directed threats to international organizations that have happened in the past, such as the 2018 attempted intrusion into the Organization for the Prohibition of Chemical Weapons (OPCW), attributed by the Netherlands and UK to Russian intelligence (Government of the United Kingdom 2018), hardwired pause treaty institutions may also be the target of AI systems themselves.

4 Feasibility of inference-only hardware

For a hardwired pause to be technically feasible, the key technology of inference-only AI chips must be possible to develop. In this section, we define the term and assess the possibility of developing such chips within a reasonable time frame.

4.1 Quantifying training capability

We operationalize the **residual training capacity** of an inference-only accelerator as its arithmetic throughput for training operations. Recall the definition of training operations in Section 3.2.1 as calculating gradients or performing inference on generic AI models, with each operation quantified as required FLOPs of dynamic matrix multiplication in the service of either. A matrix multiplication is dynamic if and only if *both* operands are determined at runtime: multiplying a fixed weight matrix by a mutable activation matrix is not dynamic.

Not all dynamic matrix multiplication capacity is training capacity: a matrix multiplication performed as one internal operation of a complex application-specific integrated circuit cannot necessarily be exploited to perform any matrix multiplication the user wants. When dynamic matrix multiplication units have user-writable inputs and user-readable outputs, we refer to them as **general matrix-multiplication (GEMM)** units. In practice, the primary specification for certifying inference-only accelerators is that they should be impractical for performing high-throughput GEMM in the service of training operations.

The distinction between dynamic matrix multiplication and GEMM is critical because implementing inference for whitelisted weights internally requires performing matrix multiplication at scale, including by necessity dynamic matrix multiplication for attention operations. As a result, licensing inference-only hardware requires ensuring that users cannot repurpose these operations as GEMMs by using the hardware in ways not intended by its manufacturer, a process we refer to as **GEMM laundering**.

The residual training capacity of any inference-only accelerator would be lower-bounded by its GEMM-laundering capacity through its token interface, i.e., its arithmetic throughput when using whitelisted models as oracles for matrix multiplication. However, these capabilities would fall far short of even unregulated hardware like CPUs: assuming two bytes of information per token, an inference-only machine with a billion-token context length and the ability to output one million tokens per second still could not achieve more than 12 GFLOP/s through this route, requiring more than 80,000 such machines to match the arithmetic throughput of a single H100.³⁹ Thus, non-negligible GEMM laundering could only be achieved insofar as an inference-only accelerator exposed more than just the token interface to its operator.

4.2 Practical inference-only accelerators

In practice, it is unnecessary for a realized hardware system to be an inference-only machine in the idealized sense defined in Section 3.2.1. We define a (practical) **inference-only accelerator** as any hardware system that can perform efficient inference on at least one whitelisted model but has verifiably modest residual capacity for performing training operations. Restrictions on training

³⁹ To multiply an $n \times k$ matrix with a $k \times m$ matrix at 16-bit precision would require reading $nk + mk$ tokens of context and writing nm tokens of output to perform $2nmk$ FLOPs, giving a maximal rate of $(2/3) \cdot r \cdot \max(n, m, k)$ FLOP/s, where r is the inference rate in tokens/s including both inputs and outputs and $nk + mk + nm$ must be no greater than the context length L . Under the optimal configuration $n = m = k = \sqrt{L/3}$, an inference-only machine could perform at most $(2/3) \cdot \sqrt{L/3} \cdot r$ FLOP/s through this channel, equating to 12 GFLOP/s if we take $L = 10^9$ and $r = 10^6$.

capacity must be robust to the system being stolen or seized by a nation-state level adversary with physical possession of chips and full control over the deployment environment and software stack.⁴⁰

Licensing of inference-only accelerators in treaty governance would be neutral to hardware implementation, encouraging innovation by the semiconductor industry to explore different techniques for hardware specialization. The chief licensing criterion would be a quantified and verifiable bound on a hardware system's residual capacity to perform training FLOPs in the service of a large training run.

In particular, units that perform dynamic matrix multiplication would require careful inspection to ensure that they are not GEMM-launderable. Whereas weight projections and multilayer perceptron (MLP) heads could be implemented with one operand (the model weights) immutable at runtime, attention head calculations irreducibly require dynamic matrix multiply operations.⁴¹ If the hardware design allows for an adversary to commandeer these operations by injecting generic inputs and exporting results, these operations would constitute significant residual training capacity, measured in tens of percentages of the total arithmetic throughput required for inference.

To rule out GEMM-laundering risk, regulators must ensure that at least one of the two operands for each of these dynamic matrix multiplications is populated directly from outputs of a previous operation, and/or that outputs are in turn routed directly to the next operation, without input/output exposure to the software layer or physical interfaces such as inter-chip links, off-chip memory, or test and debug ports.⁴² In that case, any GEMM laundering attack would require indirectly writing these operands by writing to upstream inputs and/or indirectly reading the results by reading downstream outputs. The more linear and nonlinear operations are chained together between these upstream inputs and downstream outputs without user access to intermediate values, the more formidable this obstacle would become. If treaty authorities can certify,

⁴⁰ For a treaty implemented in the near future, we assume firmware- and driver-level guarantees are insufficient to survive a state-level adversary. Treaty authorities would have the discretion to relax requirements to allow firmware- or driver-level guarantees in trusted execution environments, if presented with compelling evidence that they could be implemented securely enough to survive relevant attack models.

⁴¹ Both the query matrix Q and the key matrix K in the score operation $Q \cdot K^T$ must be mutable at runtime, as must the value V and the softmax-transformed scores $A = \text{softmax}(Q \cdot K^T)$ in the value computation $A \cdot V$.

⁴² Memory management software that selects *which* activations to copy would not necessarily pose a significant GEMM-laundering risk, provided it could not directly set the operand to an arbitrary value.

based on a rigorous audit of the hardware system,⁴³ that such laundering is infeasible (or bound the rate at which it could plausibly occur), then the system's high-throughput dynamic matrix multiplication could contribute very little to its effective GEMM rate, leaving only low-throughput residual channels like embedded control processors or the arithmetic-by-prompting attack discussed in Section 4.1; neither of these would be sufficient to attain accelerator-grade training capacity. Host processors with significant arithmetic throughput for matrix multiplication would also need to be either excluded from use with such systems or accounted for outside the licensed boundary of inference-only hardware systems, along with other CPUs.

4.3 Model-specific integrated circuits (MSICs)

A key advantage of the hardwired pause is its potential for technological readiness in the relatively near term. The most promising examples of deployed inference-only chips are certain **model-specific integrated circuits (MSICs)**.⁴⁴ An MSIC is a chip that hardwires the architecture and weights of a specific AI model when the chip is fabricated. Whereas most computer chips are built with the flexibility to implement a wide variety of programs, and store a representation of a model's weights and architecture in writable memory, the metal wiring of an MSIC immutably fixes the architecture and weights of one model. Even without regulatory pressure, the semiconductor industry is already designing and deploying such chips purely for their commercial and technological advantages.

On general-purpose accelerators such as the H100, performing inference requires reading billions or trillions of model weights from high-bandwidth memory (HBM) for each new token serially generated. This memory bottleneck limits the serial speed at which inference can be performed to several hundred tokens/s per user, depresses the utilization of fast arithmetic units even after accounting for batch parallelization, and is responsible for most of the dramatic power demanded by these chips. A wide variety of AI labs and hardware companies are aggressively pursuing specialized inference hardware that ameliorates the memory bottleneck in a variety of ways, including by increasing HBM bandwidth or by displacing HBM with more expensive static RAM (SRAM). The MSIC solves this problem by hardwiring weights into the chip, obviating the need to read them from off-chip memory (Bajic 2026).

The hardware company Taalas, which AMD has agreed to acquire (AMD 2026), specializes in making chips of this type: its first-generation prototype

⁴³ This audit should also consider practically feasible means of physically tampering with the chip; in particular, limitations based on absent connections are more robust than those based on gated connections.

⁴⁴ We adopt the MSIC terminology from Mann 2026.

HC1 chip implements models up to 8B parameters. Taalas has implemented and deployed a version customized to implement the Llama3.1-8B model and reports that its chip serves inference at 17,000 tokens/s per user, versus 353 tokens/s per user for the B200 accelerator (Taalas n.d.).⁴⁵ This advantage is in per-user serial speed, not aggregate throughput, where general-purpose accelerators make up this difference in parallel processing. The HC2, Taalas' second-generation chip scheduled for deployment in late 2026, is designed to implement frontier-scale models by sequentially chaining chips holding 20B parameters each. Taalas also reports that its manufacturing process is suitable for fast and low-cost specialization to new models, allowing it to realize any AI model in customized hardware within two months of receiving the weights (Bajic 2026).

A second recent example is the hardwired-neuron language processing unit (HNLPU), an MSIC architecture designed by academic researchers (Liu et al. 2026). The HNLPU is not commercially available, but its architecture is public, allowing for better initial assessment of its routing. The intended use of the chip is only to produce token outputs from token inputs, with no software stack. However, likely because it was not designed with the goal of hardening it against adversaries, the proposed implementation leaves open adversarial channels for GEMM laundering.

According to the paper's peer-reviewed design, at least one operand for each attention operation is derived from hardwired projections of activations, with the result fed into another hardwired projection. But the proposed implementation is composed of 16 chips in separate packages, and each attention head is split across four chips. As a result, several intermediate results in attention operations are sent between chips over CXL links, giving opportunities for injection and exportation.⁴⁶ In order to close these channels and license the HNLPU at a low residual training factor, designers would need to protect these links from injection and exportation, for example by placing each group of four chips that share attention heads in a single sealed package, or by requiring that cross-chip traffic be encrypted and authenticated using keys physically hosted on each chip;

⁴⁵ The HC1 supports applying a low-rank adaptation (LoRA) adapter in order to run a fine-tuned version of Llama3.1-8B (Bajic 2026), though this does not imply that the HC1 is suitable for *training* LoRA adapters. Supporting LoRA adapters makes the HC1 more attractive in the current market, but MSICs without this feature could be produced if necessary to reduce any residual training capacity.

⁴⁶ Specifically, K in $Q \cdot K^T$ and V in $A \cdot V$ are populated from a memory buffer connected to on-system HBM, and sent over inter-chip links, while Q is populated from a sum calculated over four chips, with partial sums traveling over inter-chip links; all three are therefore potentially writable. The result $A \cdot V$ is in turn calculated by summing over the same four chips, exposing partial sums for export.

authorities would need to audit the resulting system for other GEMM-laundering channels.

The commercial viability of hardwired MSICs is especially notable in the present market environment, where new frontier models are released every month. Google reportedly also designed the first version of its Frozen chip as an MSIC but ultimately reverted to a design that etched only model architecture, due to the short life cycle of model weights (James 2026c; Woo and Liu 2026). The commercial advantages of MSICs over general-purpose hardware would correspondingly be greater in a world with a frozen whitelist (see Section 7.3). There is every reason to expect the already-robust trend of innovative hardware specialization in the semiconductor industry to accelerate further after a hardwired pause.

An MSIC as defined above, requiring that both the architecture and weights of a model are hardwired into the chip at the time of manufacture, is not the only conceivable type of inference-only chip that is specialized for a particular model. An alternative type of chip, which we might call an **Authenticated-Weight Inference-Only Chip** (AWIC), hardwires the architecture of the model and more generally fixes the inference function of the chip at the circuit level, in such a way that the chip has very low residual training capacity, but it allows the weights of the model to be read in from memory during inference, with hardware authenticating the weights used for inference against a cryptographic hash hardwired into the chip at the time of manufacture.⁴⁷ One potential advantage of this approach is that it may be easier and less expensive to scale to large models, since the large set of weights can be held in high-density memory, rather than being hardwired into model-specific silicon; on the other hand, this incurs the cost of memory transfer and authentication, which may reduce speed and energy efficiency. We believe that inference-only chips based on MSICs (which should perhaps be called Hardwired-Weight Inference-Only Chips or HWICs) and AWICs should be investigated for the purposes of a hardwired pause.

4.4 Time required to ramp inference-only hardware production

In this section, we make very rough estimates for the duration of Phase 1 (design and production setup for inference-only chips) and Phase 2 (managed substitution of inference-only chips for flexible accelerators) of a hardwired pause. Recall our assumption that by the beginning of Phase 1, marked by the pause agreement, there is a provisional model whitelist. Thus, on day one of Phase 1, hardware designers and manufacturers can initiate development of inference-only chips

⁴⁷ The demo design at <https://github.com/Berkeley-AI-Risk/inference-only> implemented on an FPGA is a design for an AWIC in this sense.

for models on that provisional model whitelist. The scenarios below concern inference-only hardware suitable for running at least the initially whitelisted models; if new hardware or adjustments to hardware are needed to run models added to the whitelist at later times, such hardware may become available after the initial wave of inference-only chips.

As suggested in Figure 1.1, the duration of Phase 1 depends heavily on industry progress on inference-only chips during Phase 0. Moreover, in reality there is no single Phase 1; instead, there are different Phase 1s for different hardware companies. For example, as noted in Section 4.3, due to having developed a platform for hardwired inference chips during Phase 0, Taalas claims that “From the moment a previously unseen model is received, it can be realized in hardware in only two months” (Bajic 2026). Creating chips from scratch starting in Phase 1 would take considerably longer. Such an effort must pass through three subphases: Phase 1.1, design to tapeout;⁴⁸ Phase 1.2, tapeout to first silicon; Phase 1.3, first silicon to production release. For comparison, consider OpenAI’s Jalapeño chip (OpenAI 2026a), which is an inference-optimized rather than inference-only chip. Thanks in part to accelerated design using OpenAI’s models, Phase 1.1 (starting with “architecture concept” and ending with tapeout) for their Jalapeño chip took 13 months, namely October 2024 to November 2025 (Ho et al. 2026). Phase 1.2 then took an additional 6 months, from November 2025 to May 2026 (Ho et al. 2026). We do not yet know the duration of Phase 1.3 for Jalapeño, but OpenAI plans initial deployment of Jalapeño by the end of 2026 (OpenAI 2026b), so we roughly estimate 7 months for Phase 1.3. This brings the total duration of Phase 1 for Jalapeño to 26 months. Since AI models will likely be significantly more capable at chip design at the time of Phase 1 than they were in October 2024 when Jalapeño development began, future chip design phases may go even more quickly than they did for Jalapeño.⁴⁹ Considering a range of chip companies—from those that start inference-only development from scratch at Phase 1 to those that already have an adaptable inference-only platform at the beginning of Phase 1—we might expect timelines of roughly 26 months or less for their Phase 1s. Table 4.1 shows other AI chip design timelines, most of which occurred before the kind of powerful AI assistance used for Jalapeño was available.

⁴⁸ The design to tapeout phase can be further divided into subphases (van Blommestein and Vishwakarma 2025): specification and architecture design; RTL design and verification; synthesis and implementation; design for test and validation; and tapeout.

⁴⁹ Compare Jeff Dean’s remarks: “I’m pretty bullish on more and more specialized hardware. . . . To really make specialization work well, you need to automate much more of the hardware design process. . . . If you could . . . design a new chip with 10 people in 3 months instead of 150 people and 2 years, you’ll see much more specialized hardware in the world” (video posted by Hansen 2026; see also Zafar 2026).

Chip	Span Measured	Months
<i>Designs from scratch</i>		
Google TPU v1 (28 nm)	“designed, verified, built, and deployed in datacenters in just 15 months” (Jouppi et al. 2017, 2)	15
Groq TSP, later LPU v1 (14 nm)	Founded in 2016 (Groq 2026) to first silicon in July 2019 (Abts et al. 2020, 155)	~36
Cerebras WSE-1 and CS-1 (16 nm)	Cerebras founded in 2015 (Cerebras Systems n.d.); WSE-1 announced in August 2019 (Cerebras Systems 2019); WSE-1 and CS-1 available in data centers in 2019 (Cerebras Systems n.d.)	~48
Etched first-generation inference hardware (TSMC N4P)	Seed round in June 2023 with “most of the architecture fleshed out” (Ward-Foxton 2023) to first working silicon in first half of 2026 (Etched 2026b; Bort 2026); first rack shipped July 2026; “three years to deliver our first rack from scratch” (Chowdhry 2026)	~36
Taalas HC1 (TSMC N6)	Founded mid-2023 to beta service with HC1 in February 2026 (Bajic 2026; Prickett Morgan 2026)	~30
OpenAI Jalapeño (with AI assistance in design)	October 2024 conception (Ho et al. 2026) to planned deployment by end of 2026 (OpenAI 2026b)	~26
<i>Derivative designs</i>		
Nvidia data-center GPU (generic)	Engineering effort per generation: “multiple engineering teams coordinate for as long as two years” (Merritt 2023)	≤ 24
Cerebras WSE-2 (16 nm to 7 nm)	WSE-1 announcement in August 2019 (Cerebras Systems 2019) to WSE-2 announcement in April 2021 (Cerebras Systems 2021)	20

Chip	Span Measured	Months
Cerebras WSE-3 (7 nm to 5 nm)	WSE-2 announcement in April 2021 (Cerebras Systems 2021) to WSE-3 announcement in March 2024 (Cerebras Systems 2024)	35
Taalas HC1, new model on existing platform	Weights to RTL “about a week’s worth of effort” (Bajic, quoted in Ward-Foxton 2026a); weights to deployable PCIe cards in two months (Bajic 2026; Prickett Morgan 2026)	2

Table 4.1

As for Phase 2, its duration can be roughly estimated as the amount of training-capable compute to be displaced, divided by the average monthly quantity of inference-only compute, measured in inference-H100e, that can be installed to provide the same amount of inference service. For illustration, suppose that at the beginning of Phase 2, there are 47M H100e of training-capable compute⁵⁰ (extrapolating to year-end-2026 from Epoch AI 2026d; 2026c). Our extrapolation of Epoch AI’s (2026d) estimates (Figure 4.1) also implies roughly 2.9M H100e of new AI compute capacity shipped each month during Q4 2026, which is rising. If we assume that in Phase 2, on average chip manufacturers can ship and install a quantity of *inference-only* chips that provide inference service equal to 2.9M inference-H100e per month, then we obtain a rough estimate of $47\text{M} / 2.9\text{M per month} \approx 16$ months for Phase 2. If global hardware manufacturing capacity continues to grow during Phase 1, and new inference-only chips are in high demand as the only way to meet rising inference demand, then the rate of production could be higher, leading to a shorter Phase 2.

⁵⁰ By convention, we print derived quantities to two significant digits, but carry derived quantities through all calculations unrounded, rather than rounding off intermediate calculations.

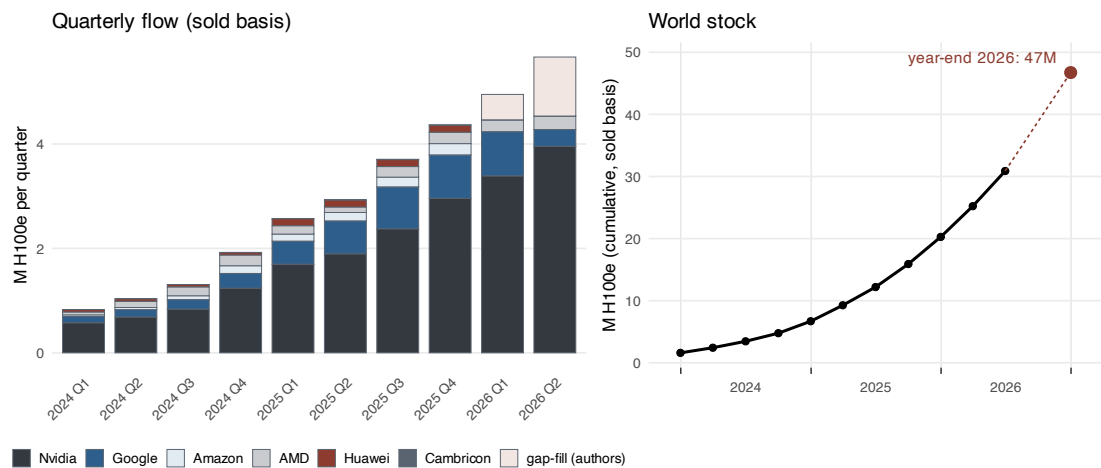


Figure 4.1. Estimates of quarterly compute flow through 2026 Q2, extrapolated to year-end 2026. Based on Epoch AI’s chip sales data (Epoch AI 2026d).

In light of the considerations above, it seems within the realm of possibility that under the strong market pressures and government incentives to develop inference-only hardware due to a hardwired pause, the world could transition through Phases 1 and 2 in less than 5 years, depending on the amount of R&D carried out in Phase 0.

5 Feasibility of hardware governance

5.1 Overview

Our second historical lesson in Section 2.2.2 was that governance regimes need concrete physical artifacts that can be tracked and regulated. AI accelerator chips are well-suited to serve this role for several reasons.

First, the supply chain for AI accelerators is one of the most complex, specialized, and geographically far-flung of any product in human history (Miller 2022). Section 5.2 discusses numerous chokepoints in this supply chain, which reside in several different jurisdictions, making it extraordinarily difficult for any state to bypass the existing supply chain. As Section 5.3 discusses, this concentration facilitates international governance to ensure that new accelerators are tracked by the treaty, and the manufacture of treaty-violating hardware produces physical evidence, in the form of the chips themselves, that can be randomly sampled and inspected.

Second, pre-pause AI chips are extensively documented through transaction, tax, and other records, including the diverted stock in China, as discussed in Section 5.4.1. In particular, in both the U.S. and China, they are expensive capital

goods that businesses have strong financial incentives to declare to their own governments, through VAT invoices or tax deductions. Additionally, AI chips are highly concentrated by both ownership and geography: a handful of firms collectively own the majority of the global stock (Epoch AI 2026c), and most AI chips are currently deployed in large data centers whose power and cooling infrastructure make them physically conspicuous (Sastry et al. 2024; Pilz and Heim 2023).

Third, incentives of both states and private actors would support compliance with the global accelerator census, as discussed in Section 5.4.2. Private chip owners could earn money by participating in profitable transitional inference service and/or buyback programs, and states could pursue their national security interests more cheaply through both sealed strategic reserves and licensed sovereign fleets than through concealment of pre-pause stock.

These advantages would all bode well for the success of a well-conducted global census during Phase 1 of a hardwired pause. Section 5.5.1 derives, in two steps, conservative quantitative estimates for how much stock might go missing in each of the U.S. and China after a thorough census; i.e., after one in which a state had made good use of the evidence at its disposal to locate and register pre-pause hardware. We first classify chip holdings into four categories based on their visibility, estimating the quantity in each category in both the U.S. and China. We then give upper bounds, based on historical analogies, for the fraction of pre-pause chips in each of these categories that might go missing, if states used reasonable means at their disposal to locate and register them.

AI Chip Fab Atlas (n=31)

25 operating · 4 under construction · 1 commissioning/pilot · 1 status uncertain

Data reviewed 25 Aug 2026



AI data centers in Epoch AI's database (n=83)

67 operational · 16 under construction

Data reviewed 25 Aug 2026

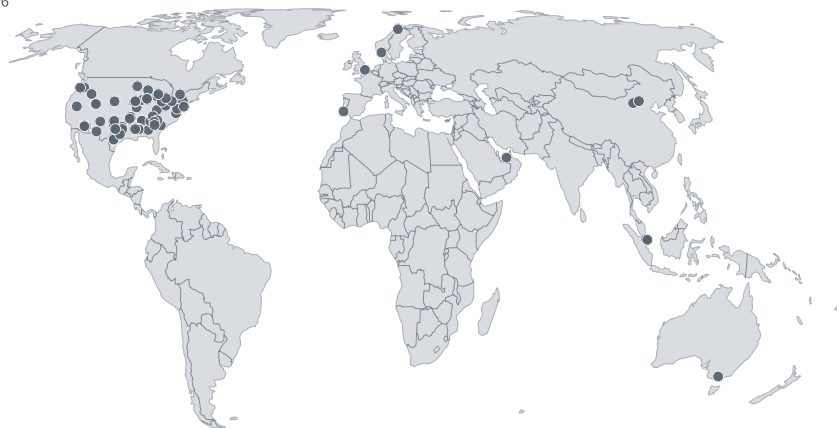


Figure 5.1. Maps showing the locations of chip fabs and AI data centers. Each marker denotes one database record; nearby records may overlap. Neither dataset is guaranteed to be exhaustive. Sources: AI Chip Fab Atlas (Berkeley AI Risk 2026); Epoch AI, “AI Data Centers” (2026g, CC BY 4.0). Basemap: Natural Earth 1:110m (public domain).

5.2 The supply chain for accelerator chips

This section gives an overview of the many specialized inputs required to construct an AI accelerator chip and the various chokepoints that could provide cooperating governments multiple points of leverage.

The construction of global distributed computing for the training and provisioning of advanced AI models is driving one of the fastest growing sectors of the global economy. The top 5 hyperscalers in the United States are spending more than \$550 billion on AI infrastructure in 2026 alone (VandeHei 2026).

This is part of a broader trend, where global AI computing capacity has doubled roughly every seven months (You et al. 2026). Though this infrastructure is important everywhere, the United States currently enjoys and is likely to continue to enjoy significant advantages in the production of compute. China has nascent and growing capacity, but it continues to lag the United States (McGuire 2025; Murphy 2025). And yet, a comparison of gross compute conceals a vast network of interconnected supply chains that span continents, link through dozens of countries, and involve thousands of entities.

From the mining of minerals like high purity quartz, copper, cobalt, and certain rare earth elements, to the finished chips to the software and infrastructure system that supports AI training and inference, the provision of advanced compute requires a significant number of steps. Many of these steps involve a unique supply chain, often with facilities or equipment that have critical junctures that act as natural bottlenecks. Many of these bottlenecks are not located within the United States or China, but are rather part of a global distributed supply chain. Replication of any one of these steps could prove difficult. Some of the most sophisticated may require a government-led effort that could require decades. Others require that one has the right natural endowment of raw materials or can secure access to them.

Natural resources are a significant constraint in the compute supply chain. Refined rare earths are an often discussed choke point, as many of them originate in or are refined in China. In 2024, 91% of the global supply of refined magnet rare earths was refined in mainland China, and in 2025, China's dominant position in these minerals came into focus during a trade dispute with the United States (International Energy Agency 2026). Other materials, such as high purity quartz, are known to come from a handful of locations. When a hurricane hit the small town of Spruce Pine in North Carolina, which accounts for an "overwhelmingly" large portion of the global supply of the material, global silicon supply chains were disrupted (Brumfiel et al. 2024). While the raw materials necessary for compute supply chains are not plentifully dispersed around the globe, many of the bottlenecks are the result of market forces and could potentially be overcome with increased surveying, investment, and mining capacity. This is less true for more technologically sophisticated aspects of the compute supply chain.

Many of the most technologically impressive aspects of the compute supply chain are extremely difficult to replicate. One of the most enduring bottlenecks in the manufacturing process is connected to Deep-Ultra Violet (DUV) and Extreme Ultra-Violet Lithography (EUV) machines. The Netherlands' ASML has a dominant market position in the production of DUV and a monopoly over the more advanced EUV. These are complex machines, with an analysis from

Georgetown's Center for Security and Emerging Technology (CSET) showing that "five thousand suppliers provide 100,000 parts, 3,000 cables, 40,000 bolts, and two kilometers of hosing to make an EUV tool. The tool weighs about 180,000 kilograms (200 tons), and ships in 40 containers spread over 20 trucks and three cargo planes. ASML reportedly only makes 15 percent of the EUV tool in-house, partnering strategically with firms worldwide to source the highest quality components" (VerWey 2024, 6). Just to move one of them would require three jumbo jets with at least 100,000 pieces each according to IBM (Murphy 2023).

Many of the parts related to EUV machines are also characterized by single or dominant suppliers. The Germany-based company Zeiss is the only producer of some of the necessary mirrors and optical components, which are among the world's most advanced (Stammmler 2021), while the Japan-based Tokyo Electron has a nearly complete monopoly over coater/developer equipment used in EUV lithography (Tokyo Electron 2026). Those nations wishing to replicate EUV lithography would have a significant amount of work replicating not just what is produced in the Netherlands but also the various subsidiary components spread around the world.

TSMC and related enterprises in Taiwan represent another significant technical bottleneck. TSMC possesses the most advanced foundries and packaging processes, as well as several other key components for the production of compute. TSMC is currently in the lead, but there are some potential competitors and redundant capacity being developed in the United States. In addition, both South Korea's Samsung and China's SMIC are capable of producing high quality, if not as advanced, AI relevant chips. China's SMIC is of particular note, as it is assisting in the development of the Huawei Ascend series (Feldgoise and Dohmen 2024). Meanwhile, advanced chip fabs are being constructed in the United States. TSMC has offshored some capacity into the US, such as its facilities in Arizona, though they are not yet as capable as the facilities in Taiwan (Rutherford 2026).

Nvidia is the world's largest supplier of advanced AI training chips. Epoch AI estimates that over 60% of global AI compute capacity comes from Nvidia's chips (Epoch AI 2026c). This concentration comes from at least three factors. First, Nvidia has been among the leaders in the design of advanced hardware as measured by FLOPs, energy efficiency, and memory bandwidth. Second, Nvidia's CUDA software stack encourages programmers to use Nvidia's chips. This software includes packages for allowing non-specialized software engineers to easily write software for GPUs, software libraries to optimize the speed and power usage of GPUs, compilers for writing optimized low-level code, and optimized application libraries. Third, Nvidia has invested heavily in its developer ecosystem, forming relationships with partners in academia, startups, and large

companies for new applications. This investment has led to developer familiarity and eventual ecosystem lock-in, which in turn have led to Nvidia being the default choice for many companies who are risk-averse and shy away from trying competitors' products. While other companies such as AMD approach or exceed Nvidia's ability to design performant silicon in terms of FLOPs or total HBM, it is the CUDA software stack that has arguably most contributed to Nvidia's leading position. Taken together, these factors have led to a concentration of model training runs, particularly at the frontier, on Nvidia's chips and software.

There are many other areas of the supply chain that also feature dominant suppliers and potential bottlenecks. HBM in particular has attracted increased attention (Freed et al. 2026), particularly given the significant role that South Korea's SK hynix has in its development. However, like others, the firm is offshoring some of its capabilities to the U.S. (SK hynix 2026).

Due to the specialized, international, and technologically advanced nature of the AI chip supply chain, no state currently has a fully indigenous supply chain, and it would be profoundly difficult for any state to create one—especially without other states taking notice. These facts correspondingly bode well for the prospects for governing chip manufacture.

5.3 Feasibility of governing chip manufacturing

This section discusses our assessment that measures like those discussed in Section 3.3.4 would be sufficient to govern chip manufacturing. Governing chip manufacturing requires achieving two sub-goals: first, maintaining knowledge of existing facilities capable of manufacturing accelerator chips, discussed in Section 5.3.1 below; and second, ensuring that the capacity for these facilities to manufacture illicit chips is measured and bounded, discussed in Section 5.3.2.

5.3.1 Feasibility of detecting undeclared facilities

In detecting undeclared lithography machines, governing authorities would enjoy distinct advantages from the fact that EUV and DUV lithography machines, being among the most advanced technologies in the world, currently require regular service from vendors (Baazil et al. 2024; ASML n.d.a). They also have distinguishing industrial features: the clean rooms for such machines require continuously powered HVAC (Faulkner et al. 1996), emit detectable gas byproducts (e.g. Arnold et al. 2018), and have specialized features to protect against disturbances from floor vibrations (Howard and Hansen 2003). An undisclosed Huawei facility in Shenzhen was discovered using open-source satellite imagery by journalists (Olcott et al. 2025).

The Dutch company ASML is the only vendor for EUV machines and, until 2026, ASML and Nikon were the only two for immersion DUV machines (Chiang 2025). A new market entrant, Yuliangsheng in China, reportedly expects to produce 12 DUV machines at the 28nm node⁵¹ by the end of 2026 (Olcott and Wu 2026). These machines are considerably less precise than EUV machines, the most advanced of which are at the 2nm node (ASML n.d.b; ASML n.d.c), but they could potentially fabricate some chips at the more advanced 7nm node (albeit at lower yields) through an advanced technique called multipatterning (James 2023). Nevertheless, as these machines require lenses from the German high-precision optics company Zeiss, they cannot be operated covertly without detection, provided that ASML and Zeiss ensure no such lenses from ASML machines go unaccounted for. Any failed lens should be replaced by the vendor and returned to its nation of origin.

For the time being, and in particular for as long as neither American nor Chinese industrial actors can replicate the high-precision optics that Zeiss provides, the lithography machine census should remain a sufficient step to prevent the undeclared manufacture of accelerators in either state, but it would cease being sufficient if either state achieves the full indigenization of its supply chain. For that reason, we would recommend to pause-willing leaders of the U.S. and China that they not pursue full indigenization, since succeeding in that pursuit could destabilize the treaty.

Even if one state did fully indigenize its supply chain, it would not follow that that state could maintain a *covert* supply chain with plausible deniability. Even if efforts to conceal manufacturing facilities succeed, covertly constructing an advanced lithography machine with high-precision optics and a manufacturing facility to house it would only be the beginning of a state's difficulties in maintaining a covert construction line. They would also need to covertly replicate numerous other highly specialized industries such as photoresist chemical production, HBM, and excimer laser parts, creating many ongoing opportunities for discovery by all-source intelligence.

5.3.2 Feasibility of monitoring declared manufacturing facilities

In addition to detecting undeclared manufacturing sites, the treaty authorities would need to monitor declared sites for undeclared production. This appears to be feasible even without frequent inspections, provided monitors simply verify *how many* wafers a lithography machine processes (for example, by receiving an attested log recording each one processed and verifying it against wafer counters

⁵¹ The “node” is a measure of resolution for lithography machines; lower nodes are more advanced but for historical reasons do not correspond to the printing resolution of the machine.

and laser pulse counts) and account for each chip printed on each wafer, along with excess silicon. One printed wafer produces a known number of chips of a given type, on an equally spaced grid, and the remaining wafer area consists of a predictable quantity of excess silicon.

If all successfully printed chips and rejects are shown to inspectors, with the former statistically sampled for inspection and the latter destroyed under monitoring, it would be exceedingly difficult to divert more than a very small fraction of production to unlicensed chips without a substantial probability of detection. Conservatively, if treaty authorities sampled and inspected 2400 chips in each production line per year, then even if there were only a one in four chance of their inspection recognizing a chip as an illicit chip, a diversion of 1% of production in that line would be recognized within six months with 95% probability.⁵²

5.4 Feasibility of a global accelerator census

A critical determinant of a hardwired pause's durability would be the success of a global accelerator census in locating, accounting for, and registering as many of the world's pre-pause chips as possible. As we discuss in Section 6.5, the binding constraint on a treaty's durability would likely be its degree of success in comprehensively completing this census.

This section discusses why a global accelerator census, or global counting of accelerator chips, conducted with the cooperation of the U.S., China, and a small number of additional jurisdictions key to international semiconductor trade in Asia would have good prospects for success. Such a census would also be an important step for many other forms of global AI governance, e.g. other plans for an international pause, as dark compute is much harder to govern than registered hardware.

Many of the records needed for a successful census are internal to governments. If either state refused to cooperate with the census, it is unlikely the other could force it to make any meaningful attempt to locate pre-pause hardware, much less to register it with the regime. These factors mean that voluntary participation of both the U.S. and China would likely be essential, as well as participation of other states such as Singapore and Malaysia, where many of the companies that supplied China with diverted stock reside.

Section 5.4.1 discusses the copious documentary evidence that could be subpoenaed by a participating state to locate pre-pause accelerators, which are capital goods held mostly by large companies that had strong tax incentives to

⁵² In six months, 1200 sampled dies would yield $\text{Binom}(1200, 0.01)$ illicitly produced dies, of which $\text{Binom}(1200, 0.0025) \approx \text{Pois}(3)$ would be detected, giving a 95% chance of at least one detection.

document their acquisitions at the time of purchase. Section 5.4.2 discusses why, given this quantity of evidence, the design in Section 3 would give pause-willing states incentives to participate, rather than abandon the pretense of participating.

5.4.1 The paper trails left by pre-pause hardware

One important cause for optimism is the amount of documentary evidence that pre-pause accelerators would have already produced, before the pause date, in the states where they had been operating. In the U.S., a single Hopper H100 accelerator is worth tens of thousands of dollars (USD) (CDW-G n.d.), and a server with eight Blackwell B200 chips is worth several hundred thousand USD (Sheumaker 2025, slide 41); the same hardware is even more expensive in China, where it has at times traded at two to three times the U.S. price, as a result of scarcity created by U.S. export controls (Pan et al. 2026; Olcott 2026b). The vast majority of these expensive capital goods are owned by large-scale businesses with intact custody records and professional accounting divisions, as well as IT departments that continuously monitor the chips.

Most accelerators would be voluntarily declared by their owners, for a variety of reasons: failing to do so would be illegal and ownership is concentrated among a few large businesses (see Section 5.5.1); registering chips for transitional inference service would likely be the most lucrative use for them; and most owners would already have declared their accelerators as a tax writeoff at the time of purchase, making it unlikely that they could get away with evasion. Below, we discuss some of the many sources of documentary evidence that a chip would leave during its lifetime.

1. **Chip owners' tax deduction records:** At present, when a business in the U.S. or China buys accelerators, it has strong financial incentives to submit a record of that purchase to the government. In the U.S., the tax deduction a business can claim for capital expenditures is worth over 21% of the purchase price (26 U.S.C. § 11(b)), while in China, a credit from a value-added tax (VAT) invoice from the government, recording the buyer, seller, and item, is worth 13% of the price, along with an additional 15–25% from an enterprise tax deduction (Standing Committee of the National People's Congress 2018, Arts. 4 and 28).
2. **Manufacturing records:** Chip foundries' sales records would show how many chips they manufacture by designer and type. These records can be audited against supplier records of inputs like HBM, advanced packaging, and silicon wafers.
3. **Vendor sales records:** Vendors such as Nvidia keep sales records of buyers and units sold (NVIDIA 2026a). If sold to a distributor, the distributors'

records would likewise record the next owner—with the important exception of those units sold to shell companies for diversion around U.S. export controls; for those chips, the sales history alone would not trace hardware to current owners unless records were furnished by the shell company in exchange for amnesty.

4. **Resale records:** The small percentage of units sold on the secondary market would often include resale records, though they would likely not be as reliable as vendor sales records.
5. **Customs records:** Import and export records would exist for most of the stock, with information such as duties paid and the type of goods imported. Public records exist in aggregate forms in both the U.S. and China, and are the basis for public estimates such as that of Epoch AI (Juniewicz 2026c). While the public records are not fully specific as to the exact type of goods transported, government data at the level of individual payments name the buyer, model, and price (General Administration of Customs of the People's Republic of China 2020).
6. **Inventory and rental records of data center operators:** Most companies that own accelerators host them at data centers, including colocation centers like QTS (Epoch AI 2026b). Colocation data centers' rental records would in turn record which companies' chips they were hosting, how many, and over what period of time.
7. **Public tenders:** Public sector buyers commonly file public tenders soliciting bids for accelerator purchases. For example, the government of Binhai in Jiangsu, China issued a tender for the purchase of 48 Hopper H200 servers in July 2025 (Baptista 2025).
8. **Power bills and power substations:** Power companies' meter readings shed independent light on the quantity of hardware operated at a given data center.
9. **Diversion re-sellers' sales records:** In addition to discovering accelerators themselves, in a search for documentary evidence the census would also be likely to uncover traders and companies that participated in the accelerator diversion market during the U.S. export control era. In exchange for amnesty, these companies could turn over their sales records, allowing for the chain of custody to be traced even for many diverted accelerators. In addition, sales records of shell companies who were already prosecuted would help to reconstruct the chains of custody of some diverted accelerators (e.g. *United States v. Hao Global LLC* 2025).
10. **Records of purchase financing:** Sellers of expensive accelerators sometimes require bank guarantees from buyers as a condition of purchase,

thereby involving a third party in the transaction. For example, 96 of 206 offers on a Chinese broker board in October 2026 required a bank guarantee, letter of credit, bank draft, or escrow for purchase of servers (Jujianbao 2026).

11. **Records of collateralization for loans:** In the recent AI data center build-out, accelerator hardware has sometimes served as a natural good offered as collateral for debt financing of construction. This collateralization commonly specifies the servers by serial numbers (e.g. CoreWeave 2025; IREN Limited 2026; Nscale 2026).
12. **Warranty return records:** Accelerators are sold in servers that typically remain under warranty for three years (NVIDIA 2026b; Dell Technologies 2026; Hewlett Packard Enterprise 2026). Those that fail under warranty can be returned to vendors for e-waste disposal, in exchange for new accelerators.
13. **Decommissioning records:** Chips that fail after lapse of warranty are required to be disposed of by licensed recyclers. E-waste and hazardous waste rules regulate how they are disposed of in China, the EU, and the U.S. (European Union 2012; State Council of the People's Republic of China 2009; Ministry of Ecology and Environment 2026; U.S. Environmental Protection Agency 2026).

Most units remain with their first owner; over 95% of the world stock by computing power is less than three years old (Epoch AI 2026c,d), shorter than enterprise chips are usually deployed (see Footnote 67), and demand for compute has been quite high for their entire lifespan. As a result, their chains of custody over their entire lifespans would be relatively simple for the census to track through sales records, with multiple data sources giving additional entry points to discover those for which a sale was for some reason not recorded.

The documentary evidence above could be combined with physical inspections of known data centers, retrospective examination of archival satellite records to discover unknown data centers based on features like cooling equipment, backup generators, and substations (Epoch AI 2026c), bounties for anonymous tips from clients or employees of companies that fail to declare, and/or chip buybacks.

Much of the documentary evidence discussed above would be available for accelerators in China that were diverted around U.S. export controls—which were often even more expensive than in the U.S. (Huang 2024; Wu and Olcott 2025). Because China has never regarded itself as bound by U.S. export controls, buying and selling AI chips and servers was legal in China provided relevant border tariffs were paid (Wu and Olcott 2025), though Chinese customs restricted importation in 2026 (Olcott 2026b). One trader reportedly told the Financial

Times, “You can go against the American rules, but not the Chinese” (Olcott 2026b). The Chinese examples discussed above, including the July 2025 Binhai public tender, the October 2026 brokers board, and the public customs records all were recorded during the export control era, demonstrating that most of the diverted stock has never been invisible to the Chinese government.

The various sources of evidence listed in this section would provide multiple lines of defense against a thorough census missing pre-pause chips. The likelihood of achieving high coverage would likely leave little room for persuading other states that missing chips were not being concealed. Next, we discuss the incentives of the U.S. and of China in light of this quantity of evidence, and the design of the example implementation discussed in Section 3.

5.4.2 Incentives of states and private actors

An important determinant of success in the census is for the incentives of both states and private actors to reward compliance. The transitional inference service during Phases 1-2 of the treaty (see Figure 1.1) would likely be highly profitable, and states parties could supplement that incentive with additional buy-back incentives, raising further the economic opportunity cost of noncompliance. It would almost certainly be within the collective fiscal capacity of states parties to buy back all existing accelerators even at several times their purchase price. Epoch’s chip-sales dataset (Epoch AI 2026d) includes roughly 25 million AI accelerators, with an average estimated purchase price (weighted by shipment quantities for each type of accelerator) of roughly \$20,000, implying an estimated total acquisition cost of \$500B. Even if we double that to \$1T for an illustrative buyback budget, allowing for additional chips not in Epoch’s dataset and higher prices, this is less than a fourth of the roughly \$4.5T that the U.S. government alone spent on Covid relief (U.S. Government Accountability Office 2025).

The evidence in Section 5.4.1, as well as the private incentives above, suggest that a thorough global census by a state *not* looking to conceal a stockpile could be expected to trace the vast majority of dark compute in a state. If this were empirically borne out in most states, it would place additional pressure on states missing larger fractions of their pre-pause stock, by making large gaps in the census appear more incriminating.

The global accelerator census could be expected to trace nearly all dark compute, at least from the time of manufacture to the first owner. As discussed in Section 3.3.2, gaps in the census traceable to the territories of specific states would be the primary input to estimating each state’s dark compute for their covert evasion and overt breakout ledgers. The more confident states could be that nearly all private actors’ commercial incentives would point toward

declaration, and the less cooperative the state had been, the more incriminating such gaps would be.

As a result, any dark compute retained by a state would likely inflate the state's covert evasion ledger and limit the issuance of licenses for other purposes, except to the extent that that state could credibly deny possession of it. As a result, states would be highly incentivized to prevent significant accumulation of dark compute by private entities within their jurisdictions. Under the implementation discussed in Section 3, they would also have more attractive alternatives if, for example, they desired stockpiles to safeguard against defection by rivals (which they could accomplish with strategic reserves), or they needed to retain a small number of AI chips for confidential national security workloads (which they could accomplish with sovereign fleets).

Compared to these alternatives, concealing a hoard of dark compute would be more costly for a state in several respects: First, it would cost the state goodwill, if concealing it required refusing reasonable requests to furnish records. Second, it would cost them part of their ability to police their own jurisdictions, if concealment required conducting a less rigorous census; and third, it would cost them the option to request more issuance of licenses for other purposes, if other states estimated the size of the concealed stockpile and added it to the state's covert ledger. Thus, a sufficiently successful census could make pursuit of such a stockpile more difficult and less attractive than alternative outlets for pursuing national security objectives, such as declared sovereign fleets and strategic reserves.

Finally, to the extent that states deliberately retain some dark compute and suspect others have as well, they could still attempt to cooperate to further reduce these dark compute stockpiles, and would have good reason to attempt to do so if they genuinely cared about the success of the treaty. For instance, if they are willing to abandon the pretense of having fully cooperated so far, they could agree to produce some number of chips conditional on others doing the same. If they wished to maintain the appearance of cooperation, they could instead agree to make renewed efforts to find more chips and announce that they had done so in small increments so long as others appear to be doing the same.

This process would be smoother if states parties all agreed on how many more chips ought to be recoverable from each party. The next section discusses how much stock in key states would be missed by a thorough census conducted in both the U.S. and China; however, it does not imply a prediction that the Phase 1 census necessarily *would* be thorough.

5.5 Estimating how many chips a census would miss

In this section, we quantitatively estimate how much pre-pause training-capable hardware in each of the U.S. and China would be missed by a thorough census. In Section 5.5.1, we distinguish chips as to their visibility to the state and to the general public by classifying them into categories that we call **holding types**, and estimating the total stock $Q(h, s)$, for holding type h and state s , focusing on the U.S. and China as key states. In Section 5.5.2, we estimate the **missing fraction** $\mu(h)$, i.e., the fraction of functional chips within each category that might go missing, as well as an estimate $D(h, s)$ of the number of failed chips that might go missing, which depends on the age distribution as well as the holding type. We combine these estimates at the end into an estimate of the total number of chips that would go missing from each state. We arrive at a final estimate of the missing stock in state s as

$$K_M(s) = \sum_h \mu(h) \cdot Q(h, s) + D(h, s).^{53}$$

Table 5.1 illustrates the calculation, giving conservative estimates of 12k total H100e in the U.S. and 11k H100e in China. The remainder of this section explains how we arrived at the numbers in the table by combining publicly available data, empirical anchors, and judgmental estimates.

⁵³ In theory, we should multiply $\mu(h)$ by the number of *functional* chips in each category, not the number of all chips; but this makes no meaningful difference to our estimate, because the failed chips are less than 1% of the total in each group.

Table 5.1: Pre-pause accelerator stock and how much the census would miss, by holding type, year-end 2026 (H100e). This calculation table gives our conservative estimates, by holding type, of the quantity that could be expected to be missed in the U.S. and China, based on assumptions described in this section.

Holding type	Stock Q(h, s)	Failed chips	Failed chips missing	Functional chips	Missing fraction $\mu(h)$	Functional chips missing	Total missing
United States							
Publicly identifiable holdings	33M	170k	3.8k	33M	—	0	3.8k
On-books business and institutional holdings	2.1M	16k	520	2.1M	0.19%	4.0k	4.5k
Off-books business and institutional holdings	—	—	—	—	—	—	—
Informal holdings	35k	160	160	35k	10%	3.5k	3.7k
Total, United States	35M	190k	4.5k	35M		7.5k	12k
China							
Publicly identifiable holdings	730k	5.9k	200	730k	—	0	200
On-books business and institutional holdings	2.3M	17k	960	2.2M	0.19%	4.3k	5.3k
Off-books business and institutional holdings	99k	660	330	99k	1.6%	1.6k	1.9k
Informal holdings	38k	180	180	37k	10%	3.7k	3.9k
Total, China	3.1M	24k	1.7k	3.1M		9.6k	11k

5.5.1 Estimating the quantity of pre-pause hardware

AI chips differ as to their visibility to governments and the general public. Many AI chips are located in known data centers like OpenAI’s Stargate Abilene, whose size is visible from satellite and whose chip counts are known from public reports (e.g. Crusoe 2025), or are owned by a handful of large companies. A smaller number are held by businesses off-books, by individuals, or by shell companies engaged in diversion around U.S. export controls. While they are smaller in number, these would also be more likely to go missing. It is not known from publicly available data how many chips are in each category; where data is unavailable, we give judgmental estimates, seeking to err on the conservative side by overestimating the number of missing chips (this being conservative because a lower estimate of the number of missing chips would make the treaty more durable).

In this section, we distinguish four categories that we call **holding types**,⁵⁴ in decreasing order of visibility:

1. **Publicly identifiable holdings:** Chips held by either the small number of chip owners identified by name in the Epoch AI chip owners data (Epoch AI 2026c), or in holdings identifiable from public reports.
2. **On-books business and institutional holdings:** Chips not in the first category that are nevertheless held on-books by businesses and/or public institutions. This category includes chips imported and sold in China after diversion around U.S. export controls, but with some state record such as a tax invoice or customs entry identifying the holder.
3. **Off-books business and institutional holdings:** Chips in China sold off-books to avoid U.S. export controls, which did not re-enter state records but are deployed by an operating business or institution.
4. **Informal holdings:** Chips held by individuals, as well as shell companies and traders engaged in diversion.

We only estimate the third category in China because of the large quantity of diverted stock located there. The holdings in the U.S. that are off-books for diversion purposes are held by shell companies and traders, but are not likely to be actively deployed since they are bound elsewhere. For each holding type h and state s , we estimate the quantity $Q(h, s)$ of pre-pause stock, measured in H100e units, at year-end 2026.⁵⁵

For publicly identifiable holdings, we combine Epoch AI's chip owners data (Epoch AI 2026c) with other public records on the internet to identify publicly reported holdings.⁵⁶ According to Epoch AI's data, a significant majority, 78% measured in H100e as of year-end 2025, are owned by seven large American companies,⁵⁷ with more owned in publicly identifiable holdings by neoclouds, the

⁵⁴ This categorization is relevant to our estimation and does not represent a relevant administrative categorization for a treaty regime. For example, university holdings in the U.S. are in the first category if they are discoverable on the internet and in the second category if not, but a census would treat them the same.

⁵⁵ Our different data sources end at different times; for example, Epoch AI's estimate of the diverted stock ends at year-end 2025 (Epoch AI 2026c); its chip sales data ends in Q2 2026, with some gaps in the latest entries (Epoch AI 2026d); and its data center records include planned sites through the end of 2026 and beyond (Epoch AI 2026b). A present estimate would already require some extrapolation, so we extrapolate to the end of the year, matching the year-end 2026 date of our near-term pause scenario in Section 6.

⁵⁶ Because the chip owners data does not specify where chips are held, we make educated guesses about the locations of chips held by identified American companies, and of chips in Epoch's "Other" category; see [calculations supplement](#).

⁵⁷ Google, Microsoft, Amazon, Meta, Oracle, SpaceXAI, and CoreWeave.

U.S. Federal Government, large companies like Tesla and Apple, universities, and a longer tail of smaller companies and private owners. A significant proportion of the chips in China are identified in public vendor reports about buyers of their chips.⁵⁸ We arrive at estimates of 33M H100e in the U.S., and 730k H100e in China in this category; the true number would likely be more if our search had been more exhaustive.

Of the projected residual, roughly 2.1M H100e in the U.S. and 2.4M in China, we estimate the remaining three categories in three steps, conservatively allocating more to the less visible categories, as follows:

1. We assume 1% by H100e are held by individuals. This is likely an overestimate for accelerators, making the estimate conservative.
2. In China, we assume that 10% of the diverted stock by H100e is off-books. We believe this is also a conservative estimate, given the convergent evidence discussed in Section 5.4.1 that diverted accelerators are not treated as contraband in China, and because businesses had strong incentives to file VAT invoices.
3. We allocate what remains to on-books business and institutional holdings.

Based on investigative reporting and court filings from prosecutions, Epoch AI estimates 660k H100e diverted over 2024-25 (Juniewicz 2026a). We use their estimate and extrapolate it for a third year,⁵⁹ to arrive at 990k H100e.

Finally, we add one month's inventory at the estimated diversion rate (330k per year, based on Epoch AI's estimate of 660k H100e diverted during 2024 and 2025). We count half toward each state's informal holdings. The resulting estimates are shown in Table 5.1.

5.5.2 Estimating the missing fraction by holding type

This section estimates what fraction of each holding type would be missed by a thorough global census. As the census is hypothetical, no data is available on these quantities, but we can use empirical anchors from analogous cases to arrive at conservative overestimates.

We consider two distinct reasons why chips might be missed: they might have failed and been disposed of without a verifiable record, or they might be functional but retained by their owners.

⁵⁸ For example, IT Home (2023) reported a large sale of over \$1B to ByteDance of 100,000 Nvidia A100 and H800 cards.

⁵⁹ Epoch AI's estimates for the diversion rate per quarter were declining by year-end 2025 (Juniewicz 2026a).

Missing failed chips For failed chips, we additionally distinguish between chips that failed under warranty and would have been returned to vendors for a refund, and chips that failed outside of warranty, which might not have been disposed of properly. Cui et al. (2025) observe a physical replacement rate of 0.45% per year for new hardware (range 0.16%–0.97%), but the failure rate would likely increase as chips age; we use the central estimate but assume a three percentage point increase for each year after year 3, applying it to the estimated age distribution of hardware in different categories.⁶⁰

We assume the following about the rate at which failed chips go missing:

1. Of chips that failed under warranty, among the identifiable and on-books business and institutional owners, 99% by H100e were returned to the vendor or have another contemporaneous record of their failure, such as a certificate for proper *e*-waste disposal or an inventory record, leaving 1% unaccounted for.
2. Of chips that failed outside of warranty, in the same two categories,⁶¹ 90% by H100e have a contemporaneous record, leaving 10% unaccounted for.
3. Of chips that failed in the off-books business and institution holdings category, we assume that 50% have a contemporaneous record, leaving 50% unaccounted for.
4. Of chips that failed in the informal holdings category, we assume that none have a contemporaneous record.

These predictions seem to us conservative (i.e., overestimates), given the strong incentives for businesses to claim warranty refunds, the likelihood of their complying with environmental regulations, and the continuous measurement of chips in many data centers. However, we acknowledge inherent uncertainty about the real values. Altogether, we obtain an estimate of 4.5k H100e of missing failed units in the U.S., and 1.7k H100e of missing failed units in China.

Functional chips not declared The global census would require chip owners to declare their chips; chip owners could also profit from doing so, as discussed in Section 5.4.2, but some might seek to retain chips for private use or for use on the black market. We assume different rates of compliance with this mandate for the different holding types, based on factors discussed below.

For chips to be missed from one of these sources, at least three things would need to happen:

⁶⁰ We include A100-class Nvidia accelerators and later, ignoring previous generations of hardware that had arithmetic throughput similar to present-day consumer cards.

⁶¹ This category includes all on-books diverted chips that failed, since diversion voids the warranty.

1. The chip would need to be physically operated by its owner at a site other than at a data center, or it would have to be missed by the data center's own declaration and the state's site visit,
2. It would need to go undeclared by the owner, despite the owner's past and present financial incentives to declare it and the risk of punishment from failure to declare, and
3. Government attempts to trace it from the seller would need to fail.

For publicly identifiable holdings, which are held by large businesses and public institutions, we assume negligible noncompliance. We estimate the rates at which each of these would occur as follows, applying empirical anchors based on estimates of analogous quantities, again attempting to err on the conservative (high) side:

1. We estimate that about 10% would not be hosted at a data center, based on a 2024 Lawrence Berkeley National Laboratory report placing under 10% of U.S. servers in closets, server rooms, edge sites, and small businesses' own premises (Shehabi et al. 2024). This is likely an overestimate since AI chips are more likely to be hosted in data centers than other enterprise computers; the same report makes an assumption that all AI accelerators are in data centers. Of the 90% at data centers, we assume no more than an additional 6% would go undeclared by both the data center operator and the owner, applying the IRS empirical anchor below, giving 16% total.
2. We estimate that, among those chips not findable through data centers, no more than 6% would go undeclared, based on the fraction of income that goes unreported to the IRS when there is "substantial information reporting" (Internal Revenue Service 2024). This is likely an overestimate since underreporting income is financially profitable, whereas underreporting capital expenditures would be financially counterproductive.
3. We estimate that, among on-books chips not findable through data centers or declared, less than 20% would go unrecovered by means of tracing the chain of custody from the seller or importer, based on the fraction of firearms that the Bureau of Alcohol, Tobacco and Firearms fails to trace based on retail records (Bureau of Alcohol, Tobacco, Firearms and Explosives 2023). This is likely an overestimate because retail sales of accelerator chips are capital-intensive business purchases that create additional financial and operating records as described in Section 5.4.1.

Multiplying these numbers gives a conservative bound of $\mu(\text{On-books}) \leq 0.19\%$, among chips that had not failed.

Off-books business holdings in China could plausibly see higher fractions go missing. A larger fraction might go undeclared because owners who preferred to retain their chips for black market use could do so more easily, so we inflate our 6% IRS anchor to 20%. It would also be more difficult for the state to recover those that went undeclared, so we inflate our 20% ATF anchor to 50%. Then we arrive at $\mu(\text{Off-books}) \leq 1.6\%$, among chips that had not failed.

Finally, we assume that informal holders would be similar to off-books business holdings, except with none hosted in data centers, giving $\mu(\text{Informal}) \leq 10\%$. Combining these estimates gives the numbers reported in Table 5.1.

Combining the above estimates of the missing fraction with our public records of named holders,⁶² we obtain estimates of missing but functional hardware amounting to 7.5k H100e for the U.S. and 9.6k H100e for China. These can be found in Table 5.1.

6 Feasibility of verification

6.1 Summary assessment

This section assesses the feasibility of verifying a hardwired pause, assuming it is feasible to produce suitable inference-only chips as discussed in Section 4, and to successfully govern pre-pause and newly manufactured AI chips as discussed in Section 5. Our assessment is that an implementation similar to that described in Section 3 would likely be technically feasible for the immediately foreseeable future, if the governments of the U.S., China, and the supplier and chip owner states listed in Section 3.1 implemented the governance and verification measures discussed in Section 3.3.

Our assessment does not rest on an assumption that verification on deployed hardware would be highly reliable in the short term. Instead, it rests on quantitative assessments that:

1. Neither China nor the U.S. currently has sufficient total training capacity to carry out an overt breakout program, under plausible assumptions about what such a program would require; and
2. Neither China nor the U.S. currently has sufficient training capacity to carry out a strategically significant covert evasion program, under plausible assumptions about the fraction of training capacity that could be marshaled toward such a program.

On current trajectories, this state of affairs appears unlikely to change within the next year, barring a major breakthrough in training efficiency gains.

⁶² See [calculations supplement](#).

Our assessment is based in part on estimates of how far a state-level defector would need to advance the frontier in order to achieve the degree of strategic advantage described in Section 3.1.1.

During Phase 1, while transitional inference service and the global accelerator census were ongoing, the training-capable accelerators used for transitional inference service in the U.S. would likely come the closest to posing a breakout risk, and uncertainty about the quantity of diverted hardware in China would likely come the closest to posing a covert evasion risk.

In Section 6.2, we quantitatively estimate possible covert evasion and overt breakout fleet thresholds (as in Section 3.2.3) for a **near-term pause** scenario where a pause takes effect near the end of 2026, and we bound what the covert evasion and overt breakout capacities of the most significant sources of training capacity might be, for each of the U.S. and China. We also give more tentative assessments for an alternative scenario taking effect during the summer of 2027.

For both scenarios (near-term pause and mid-2027 pause), we assess that neither state would likely pose a meaningful covert evasion or overt breakout risk at the time of the pause, but with greater uncertainty about the second scenario. Our choice to assess near-term scenarios should not be taken as implying that we predict a pause agreement is imminent, or that a later pause would be infeasible. It is rather a recognition of how rapidly the world is changing and how different many inputs to the analysis could look even in a year's time, if current progress in AI continues.

In both scenarios, two sources of training capacity stand out as the most concerning: the overt breakout capacity from transitional inference service (especially in the U.S.); and the covert evasion capacity from dark compute (especially in China). The most pressing risk in the near-term and mid-2027 pause scenarios would be that fleet thresholds could erode so much from training efficiency gains that the allowances would approach these two sources of training capacity from above, and eventually cross below them, rendering the treaty unenforceable unless they could be reduced.

In the first several years of a pause, the breakout risk from transitional inference service is more likely to become a pressing challenge. As discussed in Section 3.1, a near-term pause would most likely be initially implemented as an informal agreement, with treaty terms negotiated over a period of years. During those years, more information about the progress of inference-only chip development and the global accelerator census would come into view. We discuss ways that states parties to an eventual treaty might address these challenges, including by permitting treaty allowances to begin high and shrink according to schedules, rather than by setting them low on the first day when the treaty enters into force.

In the longer term, we assess that the main threat to durability would most likely be dark compute. If states could not eventually coordinate to conduct a thorough census of pre-pause accelerators, then one state's estimated covert evasion capacity from dark compute might eventually pose a serious covert evasion risk. The durability of the treaty would then depend on its ability to fall back on detection of training at the time of chip deployment, based on national technical means or additional verification measures.

Excluding transitional inference service and dark compute, we assess it likely that both states' total estimated covert and overt ledgers would initially fall below 0.25% of the fleet sizes for covert evasion or overt breakout. These components could likely be reduced further after implementing more advanced verification measures by the end of the decade, but if such research disappointed, they could alternatively be reduced by reducing the number of licenses granted for such uses. By contrast, dark compute could be more difficult to reduce below the floor set by the success of a thorough global census.

As a result, if the twin issues of dark compute and transitional inference service could be reduced below 0.25% of F_E and F_B , respectively, then we assess that a treaty with fractional allowances of 25% ($\alpha_E = \alpha_B = 0.25$) could be sustained with two orders of magnitude of headroom: that is, the treaty could sustain two orders of magnitude of training efficiency gains before it would struggle to balance the ledgers.

Because the census would be difficult without the cooperation of local jurisdictions, we assess that a hardwired pause would be significantly more durable if it were pursued in a spirit of international collaboration between the world's two leading powers and other states, and significantly less durable if it were imposed by one power against the other's will, or by both powers attempting to coerce other states.

6.2 Fleet thresholds for defection

We now undertake a quantitative analysis of the fleet size, measured in H100e, required to seriously undermine the treaty, which varies across the three threat models enumerated in Section 3.1.1:

1. **Private evasion.** A private actor covertly trains a model in excess of the legal compute ceiling C_T ;
2. **State-level covert evasion.** A state achieves a significant strategic advantage by covertly using a secret model with training budget in excess of a covert evasion threshold C_E ;

3. **State-level overt breakout.** A state achieves a decisive strategic advantage by overtly breaking out to build a strategically decisive model in excess of a breakout threshold C_B .

Violations of any of these three thresholds could undermine the treaty, but not to equal degrees. We analyze the three models separately below and discuss how states parties might arrive at estimates of key governance parameters. For the two state-level verification threat models, we suggest simple methods for translating C_F into a covert evasion fleet threshold F_E and overt breakout fleet threshold F_B , to determine bounds on the corresponding ledgers for each state.

In this section, we present a quantitative analysis with symbolic values representing derived or negotiated quantities, presenting empirical arguments for concrete numerical values that might be negotiated in a pause that takes place at year-end 2026; we refer to this as our **near-term pause scenario**. For this scenario, we estimate initial covert evasion fleet threshold $F_E(0) = 6.4\text{M H100e}$ and overt breakout fleet threshold $F_B(0) = 180\text{M H100e}$ for such a pause. After a pause were implemented, both could erode over time due to algorithmic efficiency.

Figure 6.1 shows how these two initial fleet thresholds would change if we vary the pause date.⁶³ It compares these estimates to data and derived estimates of the total world accelerator stock, including the estimated stock that is located in each of the U.S. and China, extrapolated as functions of the date, representing the values they might take at the pause date, depending on when it occurred. We label these total stocks $Q(t)$, $Q_{US}(t)$, and $Q_{CN}(t)$; projections of these are printed for year-end 2026 in Table 6.1.

As Figure 6.1 shows, our methodologies for determining these two fleet thresholds suggest that F_E will continue to grow, because we operationalize a significant strategic advantage in Section 6.2.2 as an advantage *relative* to the frozen frontier, while F_B may not grow, because we operationalize (a defector's prospect of) a decisive strategic advantage in Section 6.2.3 as occurring at a fixed capability level. This suggests that guarding against covert evasion could be somewhat easier if a pause is implemented later, because models are currently scaling faster than hardware is being built out, creating more temptation for would-be defectors. By contrast, guarding against overt breakout could get more difficult, since every year of scaling brings us closer to any fixed capability level. Because both operationalizations are choices based on the way some contemporary actors seem to view the advantages of AI, both could change if decision-makers negotiating the treaty disagree with these operationalizations,

⁶³ More precisely, if we add a second parameter to reflect the pause date t_0 , writing the fleet threshold at date $t_0 + t_+$ —that is, t_+ years in—as $F_E(t_+; t_0)$, then the line labeled “ $F_E(0)$ ” in Figure 6.1 shows the function $f(t) = F_E(0; t)$.

or if they hold different beliefs when the treaty is being negotiated than they hold now.

Table 6.1. Installed accelerator stocks and defection fleet thresholds at year-end 2026.

Quantity	Symbol	Est. (H100e), year-end 2026	Range shown
World accelerator stock	Q	47M	45–47M
Located in the United States	Q_{US}	35M	26–44M
Located in China	Q_{CN}	3.1M	2.6–4.5M

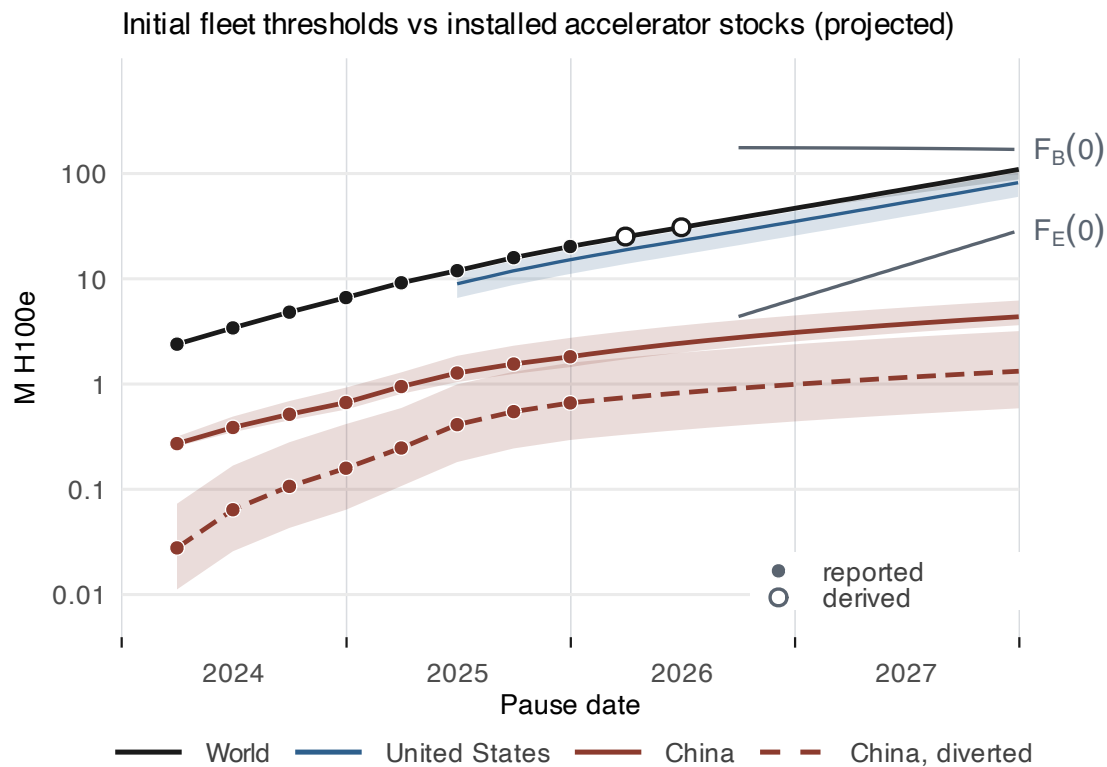


Figure 6.1. Data and derived estimates of the global accelerator stock as well as the stock located in the U.S. and China. Filled circles mark a reported source, while open circles mark values derived in the [calculations supplement](#). The initial fleet threshold estimates for the covert evasion and overt breakout threat models are shown, estimated for a near-term pause scenario beginning at year-end 2026.

For the private evasion threat model, we take a different approach, estimating a **private evasion fleet threshold** F_p defined as the aggregate global fleet

(distributed across multiple private actors) that would need to be dedicated to undermine the treaty through widespread, undetected private evasion.

6.2.1 Method for determining governance thresholds

This section discusses how a treaty might determine operational governance thresholds, by reference to the current scaling trajectory. Fleet thresholds in Section 3.2.3 are defined with reference to compute thresholds, research overhead factors, time horizons, and current frontier-level utilization. We discuss each of these below.

One simple method for determining treaty compute thresholds is to index the compute ceiling C_T , the covert evasion threshold C_E , and the overt breakout threshold C_B to the frontier threshold C_F , as fixed multiples of C_F . We adopt this method and take $C_x = \beta_x \cdot C_F$, where x is a generic subscript representing quantities relevant to the treaty compute ceiling ($x = T$), covert evasion ($x = E$), or overt breakout ($x = B$). Then, per the discussion in Section 3.2.2, the covert evasion and overt breakout fleet thresholds are also indexed to the frozen frontier C_F as

$$F_x(t) = \frac{\rho_x \cdot \beta_x}{T_x \cdot U_F} \cdot C_F(t).$$

From Epoch AI frontier training run data (Epoch AI 2026a), as well as an independent corroboration based on reports of the size of the GPT-6 Astra pretraining run (Lardinois 2026), we estimate a frontier level of roughly $C_F = 2 \times 10^{27}$ FLOPs at year-end 2026.⁶⁴ Reproducing Epoch’s frontier plot gives an estimated 4.5x increment per year, corresponding roughly to one order of magnitude increase in frontier training budgets every year and a half (Sevilla and Roldán 2024). If we define the (recent) **historical growth rate** $g = 4.5$ to be the rate of growth estimated from this regression, a race trajectory corresponding to roughly one order of magnitude growth occurs every 18 months.

⁶⁴ Epoch reports the training compute for frontier models grew at a steady rate of 5x per year between 2020 and mid-2025 (more precisely 4.5x based on our replication of their trend line; see [calculations supplement](#)). Extrapolating the trend line gives 1.7×10^{27} FLOPs at year-end 2026. The largest frontier training run estimated in that data set was for Grok 4, at 5×10^{26} FLOPs in July, 2025 (Epoch AI 2026a). Most recently, OpenAI’s Astra model was reportedly pretrained using more than 100k GPUs at its Abilene campus (Lardinois 2026), which Epoch reports is composed of Blackwell B200 and B300 chips (Epoch AI 2026a), having 2.28 H100e each. 228k H100e operating for 3-6 months at 40% utilization gives 6.9×10^{26} – 1.4×10^{27} pretraining FLOPs, or 1.2×10^{27} – 2.4×10^{27} for the full training run if post-training accounted for 40% of total training FLOPs; see [calculations supplement](#).

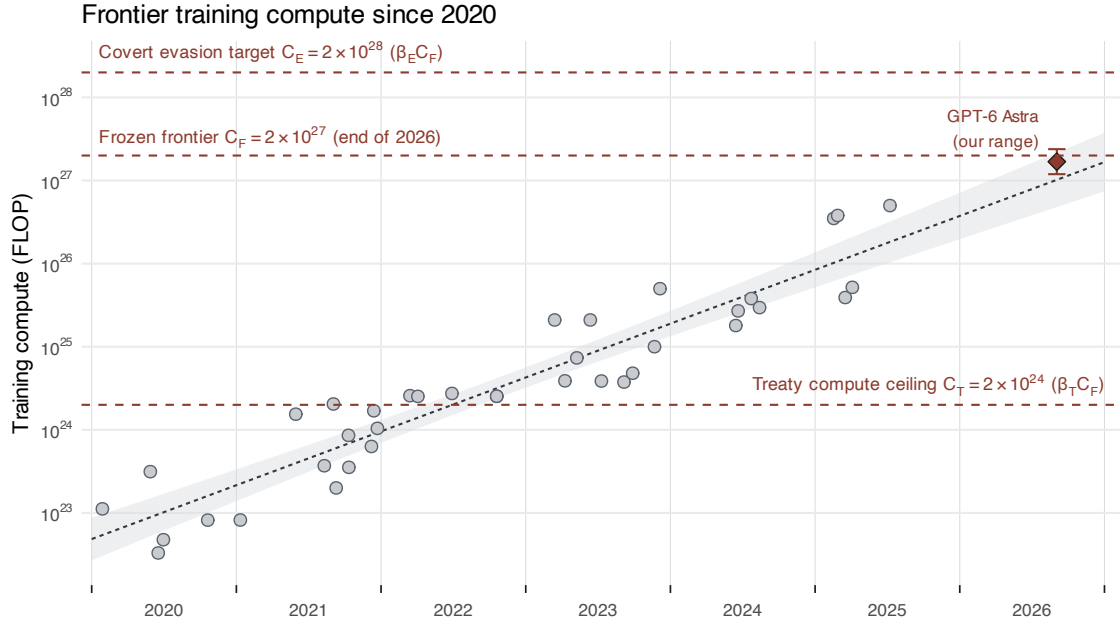


Figure 6.2. Epoch AI frontier training run data base, training budget (FLOPs) vs year the model was trained, along with our estimate and range of GPT-6 Astra, trained in mid-2026. The log-linear fit is excellent ($r^2 = 88\%$), and our calculation for GPT-6 Astra corroborates the trend, which, at year-end 2026, gives roughly 2×10^{27} FLOPs. Also shown are the covert evasion target and treaty compute ceiling that might be chosen in our near-term pause scenario.

The research overhead factors $\rho_E, \rho_B > 1$ represent the total compute budgets dedicated to the research and development process, including a wide variety of preparatory tasks that require training-capable hardware before the training run can begin, such as exploratory experimentation, synthetic data generation, inference using intermediate models that are more capable than the frozen frontier, and test runs. Denain and Wu (2026) estimate from spending data that frontier AI companies spend only about 9.6% of their research compute budget on these final training runs. If we adjust for a four-month lag between the research and training of a given model, we arrive at an empirical estimate $\rho_F = g^{-4/12}/0.096 = 6.3$ for present-day incremental frontier training. The research overhead was mildly lower for fast followers Z.ai and MiniMax, who spent 12.3% and 12.2% respectively on final training runs; correspondingly, we use the research overhead $\rho_P = g^{-4/12}/0.123 = 4.9$ for trailing models trained by private evaders.

A frontier-advancing research program aiming at $C = \beta \cdot C_F$, for a large multiple β , would either need to aim for a significantly larger “blind leap” in abilities than frontier companies do now, or it would need to proceed in several steps, bridging the gap between C_F and C_E by degrees. If it takes the latter approach, as we

find more likely, we define the **blind leap increment** δ as the optimal size of a blind leap, in orders of magnitude, relative to the best trained model available to the research program. That is, if a research program aims for $C = 10 \cdot C_F$, with blind leap increment $\delta = 1/3$, the research program proceeds in three steps, first training a model of size $10^\delta \cdot C_F$, then a model of size $10^{2\delta} \cdot C_F$, and finally a model of size $C = 10^{3\delta} \cdot C_F$. More generally, the number of required steps to attain a multiple β of the frontier is

$$n = \lceil \frac{\log_{10} \beta}{\delta} \rceil.$$

The blind-leap increment δ has implications for the research overhead; if each of the n incremental subprograms have present-day frontier overhead ρ_F , then attaining a model with training budget $\beta \cdot C_F$ requires a total research and development budget of

$$\sum_{i=1}^n \rho_F \cdot \beta^{i/n} \cdot C_F = \rho_F \cdot \beta \cdot C_F \cdot \frac{1 - \beta^{-1}}{1 - \beta^{-1/n}}.$$

The last factor is greater than one unless $n = 1$ (that is, $\delta \geq \log_{10} \beta$), in which case it reduces to $\rho_F \cdot \beta \cdot C_F = \rho_F \cdot C$, giving the same overhead as present-day incremental frontier research, $\rho = \rho_F$. More generally, the research overhead for the entire program is the above divided by $C = \beta \cdot C_F$. If $\log_{10}(\beta)/\delta$ is a whole number, then

$$\rho = \frac{1 - \beta^{-1}}{1 - \beta^{-1/n}} \cdot \rho_F = \frac{1 - \beta^{-1}}{1 - 10^{-\delta}} \cdot \rho_F.$$

We adopt the blind leap increment $\delta = \frac{1}{3}$, corresponding to roughly six months of progress on the historical race trajectory. For non-integer values of n , we simply plug in the fractional value of n , so as not to take the discrete nature of our model too literally.

Finally, we must estimate the utilization rate U_F of a present-day frontier training run. Utilization for frontier training is estimated at 30%–50% for pretraining (Epoch AI n.d.a) and 5%–20% for reinforcement learning (RL) post-training (Denain and Wu 2026). Epoch estimates that GPT-5, trained in 2025, used 60% of its total training budget on pretraining, leaving 40% for post-training, by FLOPs (Epoch AI 2025b); taking the harmonic mean of mid-range utilization estimates 40% and 12.5% for pre- and post-training would then give $U_F = 0.21$. Present-day frontier training is likely more heavily weighted toward post-training, meaning that U_F should be somewhat lower, and fleet thresholds correspondingly higher. A real agreement should use more current numbers, but

in lieu of a current estimate, we adopt the conservative value $U_F = 0.21$ for all calculations below.

6.2.2 Covert evasion fleet threshold

To arrive at a covert evasion fleet threshold, we must determine the covert evasion target $C_E = \beta_E \cdot C_F$, the research overhead ρ_E , and the covert evasion time horizon T_E . We discuss each of these below.

One reason β_E might be fairly large is that models trained and deployed covertly would suffer an **inference-time handicap** relative to whitelisted models, requiring covert models to represent a significant advance beyond the frozen frontier. The latter could be run on specialized hardware giving them an advantage in cost, scale, and speed. Covert models could not be deployed at a comparable scale, and even small-scale covert deployment would carry an ever-present risk of discovery and enforcement for users and inference providers alike. Both the cost and the risk would be compounded, in turn, by the difficulty of maintaining secrecy in the procurement and operation of scarce training-capable hardware at scale.

The domain of application of a covert model would be limited as well, owing to the need for secrecy. A covert model could not be widely deployed commercially, eliminating many of the most important perceived benefits of an AI advantage. Directly using a covert model for cyberwarfare could leave traces that would be studied for evidence that an unknown model was being used. Such evidence might well be forthcoming, especially if the model were much more capable than the models it was competing against, or produced outputs at lower serial speed than frozen-frontier cyberattack models. Deployment for internal analysis by intelligence agencies would be less risky, but also limited.

Due to both the inference-time advantage of whitelisted models over illicit models and the limited domain in which illicit models could be employed, a strategically relevant model would need to be significantly more capable than the frozen frontier on a per-token basis, and the multiple β_E should reflect this. We suggest $\beta_E = 10$ as the advantage in training FLOPs that would be required to give rough parity with the frozen frontier at inference time, using a model of Ord (2025) that balances training-time versus inference-time scaling.⁶⁵ This

⁶⁵ Ord (2025) estimates performance equivalence between models with $\log(\text{training FLOPs}) + \eta \cdot \log(\text{inference tokens})$, for $\eta = 0.7$. This suggests the 50-fold advantage in serial inference speed reported for Taalas chips may be roughly equivalent to a $\beta_E = 15$ -fold advantage in training FLOPs; we take $\beta_E = 10$ as a conservative estimate. However, this estimate is necessarily imprecise; evaders could conceivably use their scarce covert compute to scale up inference-time compute in a small number of domains with outsize returns to better per-token

factor would represent roughly an 18-month leap in per-token capabilities, on the current race trajectory.⁶⁶

Next, we turn to the research overhead ρ_E . A covert evasion program would lack many of the advantages that present-day frontier programs enjoy. It could not offer the promise of dizzying riches and social status to attract many of the brightest and most ambitious minds in the world, nor could it import insights by hiring researchers from competing firms. It could not learn as much from deploying its preliminary models, or amortize its research effort over many concurrent research programs. If covert research were carried out on aging chips, without vendor service or firmware updates, and in data centers less well-maintained than the industrial hyperscale data centers where they are presently performed, it is likely that experiments would fail and require restarting from checkpoints more often.

On the other hand, such a program could potentially ameliorate these disadvantages by using AI agents built on frozen frontier models for tasks like writing code or grading RL rollouts, provided such use did not carry an unacceptable risk of discovery. If it were carried out more slowly than current racing programs, it might also be able to plan its experiments more judiciously, relying more on human judgment to eliminate unnecessary experiments—though the researchers exercising that judgment might be less talented than is presently the case, as mentioned above.

On balance, we see more reasons to expect covert evasion research on an incremental sub-program to be more difficult than incremental research at the present-day frontier, but in the absence of a credible way to estimate the net quantitative effect of these factors, we conservatively adopt the working value $\rho_E = \frac{1-\beta_E^{-1}}{1-10^{-\delta}} \cdot \rho_F$, or $\rho_E = \rho_F$ in the special case where the research program is carried out in one leap. For the adopted values $\beta_E = 10$, $\delta = \frac{1}{3}$, and $\rho_F = 6.3$, we obtain

$$\rho_E = \frac{1 - 10^{-1}}{1 - 10^{-1/3}} \cdot 6.3 = 11.$$

performance. In practice, β_E would be a negotiated quantity reflecting how worried leaders were about rivals gaining a small advantage.

⁶⁶ By relating a scaling factor β to a period of time, we mean only to offer a guide for the reader's intuition, based on the recent history of AI progress. We do not mean in any way to *predict* that the AI industry will remain on that trajectory for 18 more months, or even that it is likely to do so; after all, this paper is a feasibility study for a different political program that would sharply bend it by government fiat. The “straight line” that appears evident in Figure 6.2 could well bend for any number of other reasons, including hardware scarcity, political or economic friction, or voluntary changes in industry practices.

Finally, we adopt a strategic time horizon of $T_E = 5$ years for covert evasion. We expect evaders to be patient, but not infinitely so, for several reasons: First, time discounting would make longer-term projects less attractive: states have many alternative ways to pursue strategic advantages on shorter time horizons, and the longer they waited to begin their research program, the more likely political and technological conditions would be to change before they could get significant use out of their model. Second, keeping a large-scale strategic project hidden for a long period of time would carry a significant risk of discovery. Finally, if the evasion program were based on an initial stock, for example from pre-pause dark compute, then evaders might encounter difficulties maintaining a fleet for covert training, and then covert inference, past its intended retirement age without vendor service.⁶⁷

For our near-term pause scenario, in which $C_F = 2 \times 10^{27}$ FLOPs, these choices lead to an initial covert evasion training budget of $C_E = 2 \times 10^{28}$ FLOPs, which would correspond to a mid-2028 model on the present race trajectory, and a covert evasion fleet threshold of

$$F_E = \frac{11 \cdot 2 \times 10^{28}}{5 \cdot 0.21} = 6.4\text{M H100e}.$$

As points of reference, we project that OpenAI’s fleet at year-end 2026 will be 6.4M H100e, and China’s total fleet would be 3.1M H100e; see the [calculations supplement](#).

Since the covert evasion fleet threshold scales linearly with the frontier threshold C_F , it is expected to scale as frontier training scales: if the pause instead took effect in mid-2027, with $C_F = 4.2 \times 10^{27}$ FLOPs, the same steps would yield $F_E = 14\text{M H100e}$. However, we expect this estimation paradigm will degrade the farther out in the future we attempt to project it.

While we focus on the treaty’s ability to prevent covert evaders from developing a strategically significant model, the difficulties of covert evasion would not stop after the model was trained. In order to realize a strategically relevant advantage, the evading state must operate a covert fleet capable not only of paying the one-time cost to train the model, but also of serving inference at sufficient scale to outperform the more abundantly available inference capacity for models at the frozen frontier. Inference demand would eventually create an indefinite ongoing requirement to replenish covert hardware without detection, giving treaty authorities and state intelligence agencies ongoing opportunities to catch evasion.

⁶⁷ For example, Azure retired its K80, P100, and P40 service after 7 years in 2023 (Microsoft 2023), and its V100 service after 8 years in 2025 (Microsoft 2025); Google Cloud ended P100 support after 10 years in 2026 (Google Cloud 2026).

6.2.3 Overt breakout fleet threshold

As with covert evasion, we now turn to estimating the overt breakout training budget $C_B = \beta_B \cdot C_F$, the research overhead ρ_B , and the overt breakout time horizon T_B , to arrive at the required breakout fleet threshold F_B .

The most commonly cited obstacle to international cooperation is the fear that a geopolitical rival could achieve decisive and lasting strategic hegemony by scaling AI models far beyond the current capability frontier. In some imaginings of this scenario, participants are now locked in a race to detonate an “intelligence explosion,” which would in short order lead to the arrival of a **superintelligent** AI model: one so powerful that it can quickly and radically reshape the global order according to the wishes of its creator—if its creator can control it (Bostrom 2014). Some contemporary actors’ statements convey the belief that this remaking of the global order could occur even if the advantage is short-lived,⁶⁸ so the regime’s choice of C_B would likely be defined by the point at which actors expected this absolute capability level might be attained, and not with reference to a relative capability advantage β_B over the frozen frontier (but it might still track training efficiency gains by holding the ratio $\beta_B = C_B(t)/C_F(t)$ fixed).

Since attaining such powerful advantages would, almost by definition, require models whose capabilities far surpass contemporary human experience—or, perhaps, even human imagination—we cannot expect to empirically estimate when models will achieve such transformative capabilities. Fortunately, this would not be necessary for the treaty to estimate directly because, to assure the verifiability and durability of a treaty, the relevant question is not about the objective fact of the matter as to when (or whether) models will reach superintelligence. After all, if and when a state broke out and resumed racing to build a strategically decisive model, the treaty would likely have already failed. Instead, the relevant question is about the expectations of states parties, especially the U.S. and China as the two states parties that would most plausibly expect to succeed.

Opinions diverge widely on the question of whether or when a strategically decisive model might materialize. Since those who are the most bullish about AI progress would be the most likely to attempt breakout, conservative treaty provisions should base C_B not on the best empirical estimate of the training budget actually required to train such a model (even if it were possible to come

⁶⁸ Cf. Dario Amodè’s vision that “a coalition of democracies seeks to gain a clear advantage (even just a temporary one) on powerful AI... to achieve robust military superiority (the stick) while at the same time offering to distribute the benefits of powerful AI (the carrot) to a wider and wider group of countries in exchange for supporting the coalition’s strategy to promote democracy” (Amodè 2024).

up with one), but on the estimate of what a bullish would-be defector might believe.

Thus, one simple way to arrive at an estimate of C_B is to imagine by when a relatively bullish state leader might be confident a decisive model would arrive if the AI race continued at its historical pace,⁶⁹ and work backwards using the Epoch training curve. For example, suppose that negotiators at the time of the pause think a bullish defector would be confident that superintelligence will arrive by the end of 2029, on the historical growth trajectory.⁷⁰ Then, they might set an initial $C_B(0) = 2 \times 10^{29}$ FLOPs, yielding $\beta_B = 2 \times 10^{29}/C_F(0)$ which could then be held fixed for $t > 0$ to track training efficiency gains indexed to $C_F(t)$.

A prospective breakout strategy would likely need to rely on this type of rapid phase transition and consolidation, since other states' more cheaply trained, fast-following AI models—including freely available open-weight models—presently trail only a few months behind the frontier. Such a strategy would be both expensive and risky for the defecting state, and would in particular require it to attain its decisive advantage before the rest of the world could react to stop it—or to resume racing against it, likely with a more complete hardware supply chain and therefore, eventually, a larger fleet. To reflect the time sensitivity of the breakout strategy, we adopt a breakout time horizon of $T_B = 2$ years.

For our near-term pause scenario, in which $C_F(0) = 2 \times 10^{27}$ FLOPs, these choices lead to the multiple $\beta_B = 100$, requiring a research program in $n_B = \log_{10} 100/\delta = 6$ steps and research overhead of

$$\rho_B = \frac{1 - 100^{-1}}{1 - 10^{-1/3}} \cdot 6.3 = 12,$$

leading in turn to an overt breakout fleet threshold of

$$F_B = \frac{12 \cdot 2 \times 10^{29}}{2 \cdot 0.21} = 180\text{M H100e}.$$

By comparison, we project the global accelerator fleet at year-end 2026 to be 47M H100e.

The value of F_B is sensitive to the bullish leader's timeline; if an even more bullish leader expects a decisive model to arrive at year-end 2028, then the same

⁶⁹ Also known as the leader's "superintelligence timeline."

⁷⁰ Given constraints on hardware manufacturing, it may be unrealistic to predict the exponential growth rate $g = 4.5$ to persist through the end of 2029. For empirically-grounded judgmental forecasts of the world computing supply in 2029 on its present trajectory, see Larsen et al. (2026). The negotiated value for a treaty would be defined with reference to the ambitions and temptations of world leaders; rather than attempt to credibly estimate the future trajectory of the AI race through 2029, we apply the simple method above for illustration purposes and acknowledge very wide uncertainty for both values.

method would produce near-term pause values $C_B = 4.3 \times 10^{28}$ FLOPs, $\beta_B = 22$, and $F_B = 36\text{M H100e}$. If we change the breakout time horizon from two years to one year, the previous two F_B values would change to 350M H100e and 73M H100e, respectively.

The choice of breakout time might likewise be sensitive to the quantity of strategic reserves states are allowed, as well as expectations and possible commitments about how coalitions would respond to defection. We adopt 180M H100e as our illustrative value going forward but, in practice, we expect a realized choice of F_B would be a matter of negotiation and trust more than one of estimation.

6.2.4 Private evasion fleet threshold

In order to successfully pause the frontier, it would be necessary for treaty authorities to prevent widespread private (non-state) training of new sub-frontier models that approach frontier scale. We operationalize this objective in the treaty as the prohibition on beyond-threshold models, i.e., models with total training FLOPs above the threshold C_T or total parameters above N_T .

To choose a good C_T , authorities should ensure that it is far enough below C_F to prevent legal models from competing with the frontier, while also ensuring it is high enough to be enforceable, so that widespread evasion does not undermine the treaty. As an illustrative value, we suggest keeping C_T three orders of magnitude below C_F (i.e., $\beta_T = 0.001$) giving roughly $C_T = 2 \times 10^{24}$ in our near-term pause scenario, but a wider range of values could likely support a treaty depending on leaders' judgments.

If C_T is below C_F and C_E by three and four orders of magnitude, respectively, a marginal violation of C_T would be unlikely to advance the frontier and would thus do little to affect the strategic balance undergirding the treaty. However, sufficiently widespread violations would lead to a perception of unenforceability. To maintain the rule of law, the treaty would need to combine sufficient penalties with sufficiently credible detection to deter all but a few private actors. In this section we consider whether enforcement before the fact, via the denial of training compute, would be sufficient.

For $x \geq 1$, let the annual violation rate $V(x)$ represent the number of undetected violating models trained per year with total training cost in excess of $x \cdot C_T$ that would significantly undermine the treaty. That is, if $V(10) = 20$, then 20 undetected violations per year with models trained in excess of $10 \cdot C_T$ FLOPs would undermine the treaty. Aggregating across violators, the minimum total undetected compute required would be

$$C_T \cdot \min_{x \geq 1} x \cdot V(x).$$

Applying research overhead ρ_P and utilization U_F , we arrive at the following fleet threshold, in hardware units per year of undetected evasion:

$$F_P = \rho_P \cdot \beta_T \cdot C_F \cdot \min_{x \geq 1} x \cdot V(x) / U_F.$$

Carrying out this analysis for $\beta_T = 0.001$ suggests that, at least initially, private evaders would need a great deal of collective compute even to achieve modest annual violation rates. As a simple worked example, suppose we take $V(x) = 100/x$. That is, it would undermine the treaty if about 100 private actors per year got away with training models in excess of C_T , or if one private actor per year trained a model in excess of $100 \cdot C_T$, or if one per decade trained a model in excess of $1000 \cdot C_T = C_F$. To sustain this annual violation rate at any value of x , the aggregate required private evasion fleet would need to achieve total arithmetic throughput of at least $100 \cdot C_T$.

Even if private evaders could achieve similar utilization as present-day frontier training, the private evasion fleet for our near-term pause scenario would need to be $F_P = 150\text{k H100e}$, amounting to billions of dollars of installed hardware. For anything like such a large fleet to operate continuously and replenish its hardware, without detection by the combined efforts of local law enforcement, state intelligence agencies, and treaty authorities, appears unlikely if hardware production is tightly controlled and the global census is even moderately successful.

Moreover, this is likely to be an underestimate, for at least three reasons: First, private evaders would be unlikely to achieve current-frontier utilization while concealing their activities from authorities. Second, they would need a larger fleet if they wanted to perform inference on their illicit models rather than only train them. And third, our calculation treats violation as occurring at only one value of x . If we instead assume, e.g., that private evaders' utilization is half that of the frontier, that evaders spend half of their compute on inference, and that x is Pareto distributed with exponent θ , then the required fleet threshold would be larger by a factor of $4\theta/(\theta - 1)$. For $\theta = 2$, this gives 1.2M H100e, or tens of billions of dollars worth of hardware.

Eventually, enough training efficiency gains might erode all of these numbers to the point where authorities would need to focus on efforts to detect training and inference rather than deny training capability. Since this would not happen soon, authorities would have plenty of time to prepare.

6.2.5 Small-scale evasion

Training beyond-threshold models from scratch is not the only form of evasion the regime would need to contend with. For example, performing inference on a small-scale fine-tune of an existing beyond-threshold model would also qualify as a violation but would be harder to police through compute denial, especially during Phases 1-2 while pre-pause hardware remained in widespread use. A Chinchilla-optimal model with just over $C_T = 2 \times 10^{24}$ FLOPs has roughly 130B parameters and could be fine-tuned using a single rack with 32 commodity H100 accelerators;⁷¹ an illegal model with just over 13B parameters could be fine-tuned more easily still. Even if difficult to enforce through training-compute denial, any fine-tune (or use of such a model if not whitelisted) would be illegal if the total compute, including the fine-tuning and the original training, exceeded the compute ceiling C_T . Enforcement would occur at the time of deployment similar to other computer crimes, with the parameter ceiling N_T likely easier to enforce after the fact than the compute ceiling.

During Phase 3, this conclusion might reverse: successful global governance of new hardware manufacture and tracking of registered hardware, combined with gradual attrition of pre-pause dark compute, would make training-capable accelerators increasingly difficult for violators to obtain.

Another concern might be that larger-scale RL fine-tuning of existing frontier models would be a better paradigm for training a new frontier model than starting from scratch. However, if $C_T = C_F/1000$ as suggested above, then unless this RL training exceeded C_T , it would likely account for less than 1% of the total compute budget of RL post-training that would already have been performed on the original model, assuming prevailing compute allocations between pretraining and post-training. Absent algorithmic breakthroughs, it appears unlikely that this type of violation would significantly advance the frontier without requiring an above-threshold run.

On the other hand, it is more plausible that a small fine-tune could train dangerous capabilities into an already very generally capable model, or train safeguards out of it. We note this would be a difficult enforcement problem for any governance regime, whether paused or not, but the reduction in the global supply of training-capable hardware envisioned by our Phase 3 appears to be a more promising response than alternatives that allow the supply of training-capable hardware to proliferate further.

⁷¹ $32 \cdot 80\text{GB} = 2.6\text{TB}$ of HBM, close to the training memory floor of 16 bytes per parameter.

6.3 Training efficiency gains and governance innovation

The primary risk to treaty durability would be that training efficiency gains would erode the compute threshold $C_F(t)$, and along with it the required fleet thresholds for each of the threat models discussed in Section 3.1.1. This would in turn render the thresholds more difficult to police over time. Estimates of the rate and origins of training efficiency gains vary widely (Ho 2026).

6.3.1 Estimates of recent historical training efficiency gains

If we take the most extreme estimates of training efficiency gains at face value, this erosion could happen quite rapidly. One regression-based estimate by Ho et al. (2024) found that there had been annual 2.7-fold training efficiency gains, by regressing language models' log perplexity loss on the WikiText and Penn Treebank benchmarks against model and data scale as well as the date on which each model was published. Their data was not based on experimental comparisons, but on a literature search over the period 2012-2023. If we were to extrapolate it forward, it would suggest that today's frontier models could be trained at one 20,000th of the cost ten years hence—that it might be possible, for example, to train GPT-6 Astra on about five accelerators rather than the more than 100,000 accelerators OpenAI used in its Abilene campus (Lardinois 2026).

However, Ho et al. (2024) urged caution in interpreting their estimate in this way, noting that they did not observe the 10,000-fold gains their model implied anywhere in their data set. In particular, they cautioned that many training efficiency gains were scale-dependent, and in particular that a large portion of these gains accrued from the transformer's scalability, which made possible the rapid scaling that took place over the same time period (Sevilla et al. 2022). That is, the observed gains would have been lower if the training runs used in 2023 were not larger than those used in 2012. This observation has important implications for a hardwired pause because, if successfully implemented, such a pause would prevent further scaling. Thus, if innovations discovered during a pause are complementary to further scaling, their greatest benefits could not be realized.

Subsequent empirical studies confirm the scale-dependence of algorithmic improvements, suggesting that the most important breakthroughs have improved AI not by wringing better performance out of 2012-scale models and data sets but by giving better returns to scale. Gundlach et al. (2025) find empirically in ablation studies that the training efficiency gains they tested have improved performance at smaller scales by a factor of only 6.3 over the same period, and combine this estimate with estimates from the machine learning literature to

estimate that the total gain was closer to a factor of 60—translating to an impressive but much less overwhelming 45% improvement per year. They estimate that 91% of gains at larger scales was explained by two scale-dependent breakthroughs, the transformer architecture (Vaswani et al. 2017) and Chinchilla re-balancing (Hoffmann et al. 2022), both of which are much more effective at present-day frontier scales than older methods but confer smaller improvements at small scales. They concurred that the gains observed by Ho et al. (2024) were contingent on the rapid rate at which computing effort was scaled up over that time period. Their evidence therefore is suggestive that the observed gains in a counterfactual “paused” world, where compute budgets had remained fixed at 2012 levels, would have been much slower.

More recently, Patel and Han (2026) found annual pretraining efficiency gains of 57% over the period 2019 to 2025, through computational experiments comparing new methods and data to old methods and data, at fixed scale $C = 10^{19}$ FLOPs. They cite scale-dependence as an explanation for the difference between their estimate and the estimate of Ho et al. (2024). Interestingly, by separately varying data improvements and algorithmic improvements, they find that data improvements alone accounted for three times as much efficiency gain as algorithmic improvements.

6.3.2 Training efficiency gains after a pause

Since training efficiency gains are a product of human effort and not an exogenous force of nature, their rate could be very different in a post-treaty world than it has been over the last decade. At present, gains are driven by overwhelming commercial incentives drawing thousands of the world’s most talented engineers to work full-time on training more capable models with costly and constrained computational resources. A ban on frontier model training could decelerate training efficiency gains by redirecting commercial research and development efforts away from training new models and toward diffusing current model capabilities.

A second reason why a hardwired pause might be expected to decelerate training efficiency gains is that it would sharply limit engineers’ ability to test ideas at scale. The internal workings of frontier AI models are famously inscrutable even to their creators, motivating the common analogy that they are not “built” like cars or bridges but “grown” by empirically fiddling with weights. As a result, research on how to train models more effectively is highly empirical. The empirical nature of training efficiency gains is well-known in the AI world; as one inventor of the transformer famously quipped with regard to another improvement, “We offer no explanation as to why these architectures seem to

work; we attribute their success, as all else, to divine benevolence” (Shazeer 2020).

A third reason relates to the finding by Patel and Han (2026) that data improvements have driven most recent pretraining efficiency gains. If so, then a future covert evasion project might have difficulty realizing these gains, as the internet of ten years hence might be much more hardened against automated crawling than today’s internet, by necessity, due to greater abundance of bots, potentially leaving covert evaders with stale data sets. There are some indications that this is already happening: Cloudflare has begun blocking AI crawlers on some websites for which it proxies traffic, and has continued to tighten its block as recently as July 2026 when it announced stricter defaults (Cloudflare 2025; Lee and Becker 2026); see also Longpre et al. (2024) for empirical evidence that data restrictions on websites are rising in the AI era.

On the other hand, an ineffectively enforced hardwired pause could plausibly accelerate efficiency gains by tightening constraints on the computational resources available to any engineering teams that continued frontier model training efforts, especially if illicit efforts are widespread. Maintaining a wide margin between governance thresholds C_T and N_T and frontier capabilities could mitigate this acceleration by reducing the temptation to evade. Likewise, efficient inference on frozen frontier models could contribute to training efficiency gains in a variety of ways, such as generating data sets for distillation, grading RL rollouts, or suggesting algorithmic improvements to engineers.

Continued training efficiency gains would not necessarily be fatal to an adaptable governance regime. As governance matures, hardware and software innovations could give authorities more options for flexible, light-touch governance: more reliable geotracking of chips could cause the supply of dark compute to dwindle away; advances in hardware specialization could yield a variety of performant hardware systems that, like inference-only chips, are inefficient for frontier model training; or advances in privacy-preserving software-based verification and workload classification could allow for the wide deployment of hardware that offers users relatively unrestricted general-purpose functionality while detecting and deterring evasion.

6.4 Quantitative ledger estimates

Treaty authorities would be responsible for ensuring that no state party has amassed a fleet capable of posing a risk of strategically-relevant covert evasion or overt breakout, by estimating the total amount of computing capacity that could be marshaled toward a covert evasion or overt breakout research program.

This section discusses a simple methodology by which treaty authorities might estimate the covert evasion capacity $K_E(c, s)$ and the overt breakout capacity

$K_B(c, s)$, for each potential source c of illicit compute and each state s , beginning with a **raw capacity** $K_0(c, s)$ and then discounting according to how much of that compute source could realistically be put to use in each of the two scenarios.

We implement this methodology to derive conservative estimates for what the covert evasion and overt breakout ledgers might look like for both the U.S. and China, for our near-term pause scenario at year-end 2026.

6.4.1 Discount factors

For each of the two risk scenarios, two discount factors could be applied: a **political discount factor** π representing the fraction of hardware in that category that could plausibly be marshalled toward building frontier models (fractionally including hardware engaging in part-time evasion), and a **technical discount factor** τ representing the average sustained utilization that could realistically be achieved, relative to present-day frontier training. That is, $\tau = 1$ means similar utilization to present-day data-center grade training, not 100% utilization.

For each compute source c and state s , the parties could negotiate means of estimating a total (undiscounted) existing capacity $K_0(c, s)$ as well as four discount factors: $\pi_E(c, s)$ and $\pi_B(c, s)$, the political discount factors for the covert evasion and overt breakout, and $\tau_E(c, s)$ and $\tau_B(c, s)$, the technical discount factors for the same. Then, the evasion capacity for that compute source and state would be estimated as $K_E(c, s) = \pi_E(c, s) \cdot \tau_E(c, s) \cdot K_0(c, s)$, and the breakout capacity would be estimated as $K_B(c, s) = \pi_B(c, s) \cdot \tau_B(c, s) \cdot K_0(c, s)$. Summing over c gives the evasion and breakout ledgers $L_E(s)$ and $L_B(s)$ for each state, which must respectively be kept below $\alpha_E \cdot F_E$ and $\alpha_B \cdot F_B$. The discount factors and the raw capacities could all change with time, but we suppress their dependence on time to avoid cluttering notation.

Considerations in determining the political discount factors would include how diffusely or centrally held the relevant chips are, whether they are directly controlled by the state, and what legal or political barriers the state would face in commandeering them in a covert or overt scenario. The covert discount factors in particular would depend on what kinds of verification measures the state allows in its facilities, how many different actors would need to be involved in a conspiracy to violate the treaty, and how successfully the state could force them to remain silent.

For example, a supercomputer facility in a sovereign fleet would be charged with a high overt political discount factor because it could be reassigned to breakout training in a short period of time ($\pi_B \approx 1$). However, it might be charged a much lower covert political discount factor if the host state is willing to allow power monitoring, telemetry mechanisms, periodic inspections of facilities,

or legal and political safeguards such as criminalization of treaty violation and protection for whistleblowers. If a state can persuade other parties to accept that no more than 20% of a facility could plausibly be marshalled toward covert training without a significant likelihood of detection, its π_E might be 0.2.

Considerations in determining technical discount factors would include limitations on memory and interconnect bandwidth, or parallelism overhead for chips with low memory. The technical discount factor for covert training could be less than that for overt training due to the technical overhead of distributed training, disguising workloads, or locating chips in secret data centers with substandard operating conditions. We undertake a simple illustrative analysis here, but a more complete analysis would disaggregate each source by the type of hardware and details about its deployment; the utilization achieved by current AI companies in frontier pretraining under idealized conditions might be realistic for some overtly-deployed accelerator hardware ($\tau_B \approx 1$), but technical discount factors for present-day inference-optimized chips would likely be below 1, and those for consumer GPUs would be considerably lower still.

To establish plausibility that these bounds could be realistically verified, we consider several sources in turn, discussing realistic upper bounds for capacity and discount factors for the U.S. and China. In each case, we compare them to the covert evasion fleet threshold $F_E = 6.4\text{M}$ and the overt breakout fleet threshold $F_B = 180\text{M}$ for our near-term pause scenario. As these estimates apply to a hypothetical future, they necessarily reflect speculations about, e.g., how cooperative pause-willing states parties would be in implementing treaty rules or furnishing sales records to locate existing lithography machines and accelerator chips.

As we will see, under plausible assumptions, it is unlikely that any of the sources discussed below could represent a significant strategic threat for overt or covert evasion in the short term after a pause. For each source, we evaluate its **headroom** for a treaty that gives an allowance of 25% for both ledgers ($\alpha_E = \alpha_B = 0.25$), defined as the number of orders of magnitude of erosion, due to training efficiency gains, that the treaty would have to sustain before that source became a binding constraint on balancing the ledger.

6.4.2 Transitional training-capable inference service

The most obvious treaty exposure would be that from transitional inference service on pre-pause training-capable hardware. Leaving these training-capable accelerators in industry-grade commercial data centers, where they could in principle be seized by a state government in an overt breakout scenario, plausibly gives τ_B and π_B close to 1.

In practice, both discounting factors might be somewhat lower. The value of π_B could vary widely, depending on the political plausibility of the state's government seizing an enormous quantity of privately-held capital for a national breakout program. Political or institutional safeguards that would make it more difficult for the state to do so, or that create hurdles that would delay deployment in such a scenario, could reduce the average amount that would be available over a window of length T_B . The value of τ_B also might be lower than 1 since the total installed stock is heterogeneous, with a significant fraction of older hardware.

In any case, we can bound $\tau_B, \pi_B \leq 1$, thereby bounding each state's breakout capacity by the total stock located in each country at the pause date.

$$K_B(\text{Transitional, U.S.}) \leq Q_{\text{US}}, K_B(\text{Transitional, China}) \leq Q_{\text{CN}},$$

We project the U.S.-located stock at $Q_{\text{US}} = 35\text{M H100e}$ total at year-end 2026, giving:

$$K_B(\text{Transitional, U.S.}) \leq 35\text{M H100e},$$

or 20% of $F_B(0)$, and

$$K_B(\text{Transitional, China}) \leq 3.1\text{M H100e},$$

or 1.8% of $F_B(0)$. Thus, at the pause date, both breakout capacities would likely be below the allowance before discounting. However, the breakout number leaves almost no headroom for a treaty with 25% allowance, and two doublings of training efficiency gains could bring $F_B(t)$ itself close to the U.S. breakout capacity. This would be a problem in the early years of the treaty, but not an insoluble one; we discuss ways to ameliorate it in Section 6.5.2.

Even while inference service continued, verification measures such as those described in Section 3.3.3 could significantly reduce its covert evasion risk. If we assume pessimistically that 10% of hardware operators in the U.S. engage in treaty evasion, somehow avoiding detection by treaty recomputation servers even while diverting 1% of their computing power to evasion, we arrive at $\pi_E = 0.001$. Even if they achieved frontier-grade utilization for their covert evasion, the total covert evasion capacity would be

$$K_E(\text{Transitional, U.S.}) \leq 35\text{k H100e},$$

or 0.55% of $F_E(0)$. Assuming, again pessimistically, that the Chinese government were able to covertly suborn *all* of its hardware operators into diverting 1% of their computing power, their covert evasion capacity would be similar, with

$$K_E(\text{Transitional, China}) \leq 31\text{k H100e}.$$

These evasion capacity numbers apply to the transitional inference service after verification protocols have been implemented. Both political discount factors would likely be higher during the stages when these methods were being set up. However, provided this could be done in a timely fashion, it would leave little time for a strategically relevant model to be covertly trained. Whole-lab inspections as proposed by Choussat and Khoja (2026) could be a stopgap interim measure, if it were difficult to stand up the inference-only cluster protocols of Section 3.3.3 at the timescale on which leaders wished to implement a pause.

For a longer-term pause, we expect that verification measures could be considerably tightened based on methods currently under development as described in Section 3.3.3, leading to K_E values that were considerably reduced for both states.

However, doing so would not on its own be sufficient for implementing the treaty, because verification measures generally do not reduce breakout exposure. This is what makes stopping the production of training-capable hardware an essential element of a hardwired pause: the more training-capable hardware states accumulate, the less credibly they can assure other states that they would not break out and race faster than before.

A more complete covert evasion analysis would consider the possibility of the U.S. or China subverting transitional inference service outside their jurisdiction. We assume such use would likely incur considerably steeper political discounting.

6.4.3 Dark compute

Significant quantities of ungoverned accelerators would be highly problematic for treaty governance. Chips lost or stolen in a host state might be charged under political and technical discount rates close to one, reflecting the high risk of their being used for covert evasion or overt breakout. While a few thousand missing chips would not approach relevant treaty allowances, a million missing H100e units would approach 16% of F_E , leaving little margin for training efficiency gains to erode the evasion fleet threshold $F_E(t)$.

Our estimates from Section 5.5 reflect what a thorough census might miss, but our analysis does not assume that a thorough census could happen overnight. Compliance would depend on political conditions we cannot foretell. In some pause-willing futures, leaders could be ambivalent about the pause and unwilling to provide any transparency to rivals; while in others, leaders' most urgent concern could be to prevent further training or fine-tuning of frontier models by private evaders. It could also be that leaders gain greater confidence in each other over time, beginning with small measures and slowly finding ways to

work together more closely. It is possible that the census would take time to implement.

We divide our analysis into several scenarios, according to states' degree of cooperation in the census.

In a **high-cooperation scenario**, we consider a world where American, Chinese, and other world leaders conclude they have a mutual interest in cooperation to address common threats, and prefer to pursue national strategic objectives through alternative treaty-provided channels. Both nations make convincing public efforts to recover pre-pause chips, but might stash a small, plausibly deniable number as insurance.

For such a scenario, we assume that both states would not significantly sabotage their own internal census methods, and would stash no more than 20% of the total missing amount, giving

$$K_0^{\text{HC}}(\text{Dark}, s) \leq 1.2 \cdot K_M(s).$$

We assume this cooperation would be evident to other states, who might conservatively estimate that at most half of the missing compute was in the state's possession, giving a covert evasion political discount factor of $\pi_E = 0.5$, leading to

$$K_E^{\text{HC}}(\text{Dark}, s) \leq 0.5 \cdot 1.2 \cdot K_M(s) = 0.6 \cdot K_M(s).$$

Applying these bounds to the U.S. and China gives

$$K_E^{\text{HC}}(\text{Dark}, \text{U.S.}) \leq 7.2\text{k H100e},$$

or 0.11% of $F_E(0)$ for our near-term scenario, and

$$K_E^{\text{HC}}(\text{Dark}, \text{China}) \leq 6.8\text{k H100e},$$

or 0.11% of $F_E(0)$.

In this scenario, the treaty would appear to be quite durable after the sunset of transitional inference service, since all other terms in the covert evasion and overt breakout ledgers would initially remain below half of one percent. It could comfortably survive two orders of magnitude of erosion due to training efficiency gains, and plausibly more if innovation in verification methods or national technical means drove down political discount factors, or if other modes of hardware specialization similar to inference-only chips drove down technical discount factors for non-dark compute on new, governed hardware.

The main threat to durability would be the threat that new sources of undeclared accelerator production would be developed; measures like those in Section 5.3 would therefore be essential.

Second, we consider a **low-cooperation scenario**, in which China and the U.S. arrive at a tense agreement with relatively little trust or goodwill. Both accede reluctantly to most records requests, but records requests that are at all politically sensitive are refused, making it more difficult to account for chips. Despite their pause-willingness, both states suspect the other of planning to cheat, and both find ways to hold back as much untracked hardware as they can, especially powerful newer hardware, possibly sabotaging their own investigation efforts in order to increase the plausible deniability of their own concealed stockpile.

In this case, we would expect a larger fraction to go missing, and no political discounting to be applied. In the U.S., it would likely be more difficult to conceal a large stock owing to the relatively greater difficulty of keeping secrets. For this scenario, we assume five times the missed quantity estimated in Section 5.5.2 for the U.S., and ten times the missed quantity for China:

$$K_E^{LC}(\text{Dark, U.S.}) = 5 \cdot K_M(\text{U.S.}), K_E^{LC}(\text{Dark, China}) = 10 \cdot K_M(\text{China}).$$

which would respectively be 0.94% and 1.8% of $F_E(0)$ for the U.S. and China, and 0.034% and 0.064% of $F_B(0)$.

In this scenario, the regime could survive one order of magnitude of training efficiency gains relatively comfortably. It could survive two if the U.S. and China could eventually achieve something closer to the high-cooperation scenario after a thaw in relations, an increase in the degree of pause-willingness, or an increased sense of urgency after training efficiency gains eroded treaty bounds. If greater cooperation could not eventually be achieved, longer-term durability would hinge on the success of detection-based fallback measures.

Finally, we consider a **non-deniable concealment scenario** in which negotiations over recovery break down. Both states openly accuse each other of pursuing covert evasion plans, and each sees limited political value in persuading the other that they are not pursuing it themselves, but a deal nevertheless barely gets through as a stopgap measure against some acute emergency. In this case, China might conceal an amount roughly equal to its pre-pause diverted stock; in the U.S., it would likely be more difficult to conceal a large stock owing to the relatively greater difficulty of keeping secrets. Then we might have

$$K_0^{NC}(\text{Dark, U.S.}) = 300k \text{ H100e}, K_0^{NC}(\text{Dark, China}) = 1M \text{ H100e},$$

which would respectively be 4.7% and 16% of $F_E(0)$, and 0.17% and 0.57% of $F_B(0)$, if undiscounted.

This scenario would leave little margin for erosion of $F_E(t)$; in particular, China's covert evasion capacity from dark compute would pose a similar issue as the U.S.'s overt breakout capacity from transitional inference service. Both

problems would need to be resolved in the early years of the treaty, but the two states would likely have a few years.

6.4.4 Inference-only accelerators

While inference-only accelerators would be deployed at scale, their residual training factors would be audited with certified bounds, and their total training capacities bounded per hardware operator and state.

The covert evasion risk from licensed inference-only accelerators would be a choice parameter for treaty governance: the number of licenses for a given state could be reduced if the state had insufficient headroom on its covert evasion or overt breakout ledger.

Under overt breakout scenarios, seized inference-only hardware in commercial data centers could be used at its peak arithmetic throughput for training ($\pi_B = \tau_B = 1$), so the treaty-certified residual training factor would be the primary protection against breakout. A host state's inference-only fleet with an average residual training factor of ϕ could remain below $x \cdot \alpha_B \cdot F_B$ while serving $x \cdot \alpha_B \cdot F_B / \phi$ units of inference demand. For example, if a state wished to spend 1M H100e of its overt breakout allowance on an inference-only fleet, then it could serve 100M H100e worth of token demand on a fleet with average residual training factor below 0.01, or 1B H100e on a fleet with average residual training factor below 0.001. By comparison, Deloitte projected that two thirds of AI compute would be used for inference in 2026 (Crossan et al. 2025), or roughly 31M H100e if applied to our earlier estimate of 47M H100e of global capacity.

An average residual training factor of 10^{-4} , amounting to 100 GFLOP/s of GEMM-laundering ability per H100e, would allow each of the U.S. and China to replicate the estimated 31M H100e of global inference service one hundred times over with inference-only accelerators, with each state keeping its breakout exposure below

$$K_B(\text{Inference-only}, s) \leq 310\text{k H100e},$$

or 0.18% of the initial breakout fleet threshold $F_B(0)$, leaving two orders of magnitude of headroom.

Here, 10^{-4} is not an estimate or prediction but an illustrative value, and the ability to balance the ledger does not rely on any particular value. If available residual training factors were higher or lower, or technical discounting factors below one could be justified, the total inference service could be adjusted based on available headroom. Significantly lower values can be achieved if all major channels for GEMM laundering can be foreclosed, but doing so would require a physical audit of a realized system.

The covert exposure could be driven further down by certifying a separate residual training factor ϕ_E , which might be lower than ϕ if the most favorable GEMM-laundering channels required physically tampering with inference-only chips in a way that would be evident to inspectors. Additional verification steps, such as those suggested above for the transitional inference service fleet, could reduce the covert evasion capacity further. If $\pi_E \leq 0.001$, as assumed above owing to the difficulty of evading discovery by the recomputation server, then the covert evasion capacity is still lower:

$$K_E(\text{Inference-only}, s) \leq 310 \text{ H100e},$$

or only 0.0049% of $F_E(0)$.

6.4.5 Consumer GPUs

Next, we consider consumer GPUs. Jon Peddie Research (2024) forecasts a global installed base of 119M discrete desktop add-in boards by the end of 2028. If there are half as many laptop units as desktop units,⁷² we might use a rough conservative bound of 100M total cards in either the U.S. or China, with the average such chip having peak dense arithmetic throughput of 30 TFLOP/s,⁷³ giving $K_0(\text{Consumer}, s) = 3\text{M H100e}$, a potential covert training risk before discounting.

Political discounting would deflate this risk considerably, especially under the covert scenario. Since gaming cards are diffusely held, deploying them at scale would require either a vast conspiracy of regular citizens knowingly permitting their cards to be used toward treaty violation, or hacking into and diverting a small portion of a much larger group of citizens' cards without their noticing. If parties deem it implausible that a state could commandeer more than 1% of consumer card capacity for a covert run without running a major risk of discovery, we might take $\pi_E \leq 0.01$.

Political discounting of consumer GPUs under an overt breakout scenario is more uncertain and depends on parties' estimation of the likelihood that governments could commandeer a large fraction of consumer hardware even in a wartime footing scenario. In constitutional democracies, at least, any such

⁷² Q2 2026 add-in board shipments reported as 12.5M (Jon Peddie Research 2026); total discrete GPU card shipments estimated at 20M by Tom's Hardware (Shilov 2026), giving 7.5M residual shipments for discrete laptop GPU cards.

⁷³ Note Nvidia restricts the throughput of 16-bit matrix multiplication with 32-bit accumulate by a factor of two on GeForce parts (NVIDIA 2021).

effort would require sufficient constitutional or statutory basis⁷⁴ and might face significant legal as well as political hurdles. States might accept the premise that it would be difficult to commandeer more than 25% of total consumer capacity toward a breakout program, giving $\pi_B \leq 0.25$.

Even generous technical discount values would likely be low for consumer GPUs since they are heterogeneous, are thermally limited when operated at high utilization, and have low memory and interconnect bandwidth, all rendering them unsuitable for efficient training at frontier scale; the covert technical discount factor would be lower still due to evasion overhead. Dean (2026a) estimates the “usefulness” (essentially a technical discount factor) of each chip, finding consumer gaming cards to be worth roughly 0.05 of an H100e. If we are twice as generous and take $\tau_E, \tau_B \leq 0.1$, the total covert evasion risk is reduced to

$$K_E(\text{Consumer}, s) \leq 3.0\text{k H100e},$$

and the overt breakout risk to

$$K_B(\text{Consumer}, s) \leq 76\text{k H100e},$$

representing 0.047% and 0.043% of $F_E(0)$ and $F_B(0)$, respectively.

Thus, under covert or overt scenarios, the consumer GPU base does not appear to represent a significant hurdle to the treaty’s durability, up to at least two orders of magnitude, due to its diffuse ownership.

One possible future failure mode of this calculation is that, under a treaty, a few individuals or firms might buy up a very large number of consumer GPUs, negating the above argument based on diffuse ownership. In order to defuse this longer-term threat, states parties might consider passing laws limiting the number of consumer GPUs held by any one actor, and taking steps to prevent skimming from the production line.

6.4.6 Licensed sovereign fleets

In Section 3.3.1 we discussed the possibility of permitting licensed sovereign training-capable accelerator fleets to be used for specific sensitive national security workloads. We assess the exposure from these fleets by considering El Capitan, the largest supercomputer in the U.S. and an important strategic asset with the primary mission of nuclear stockpile maintenance. El Capitan serves 46k AMD Instinct APUs with 980.6 TFLOP/s each, amounting to

⁷⁴ An executive order must rest on constitutional or statutory authority; its issuance alone does not establish the legality of compulsory acquisition. (See *Youngstown Sheet & Tube Co. v. Sawyer*, 343 U.S. 579 (1952).)

K_0 (El Capitan, U.S.) = 46k H100e in total (Lawrence Livermore National Laboratory 2025; AMD n.d.).

The technical discount factor for overt breakout can be conservatively estimated close to 1,⁷⁵ though it might be lower for covert evasion since the network topology is presently optimized for scientific workloads like fluid dynamics simulation rather than frontier AI training, and any evasion scheme would likely come with technical overhead.

We assume an overt breakout political discount factor of $\pi_B = 1$, since the supercomputer is fully controlled by the U.S. Federal Government, giving total breakout exposure of

$$K_B(\text{El Capitan, U.S.}) \leq 46\text{k H100e,}$$

or about 0.026% of $F_B(0)$.

The political discount factor for covert evasion, however, might be much lower, owing to the difficulty of keeping secret from staff the diversion of a large fraction of such a supercomputer. We might conservatively take $\pi_E = 0.2$, yielding covert exposure

$$K_E(\text{El Capitan, U.S.}) \leq 9.1\text{k H100e,}$$

or about 0.14% of $F_E(0)$.

As a result, we estimate that the treaty could sustain a practice of allowing a number of supercomputers on the scale of El Capitan per state without threatening treaty thresholds, at least early on.

The number could grow further with improved institutional governance or verification measures allowing for steeper political discounting, or improved technical governance measures such as hardware specialization allowing for steeper technical discounting. If the political discount factor and the technical discount factor each fell by an additional factor of five, for example, then the covert evasion capacity would fall to 0.0057% of $F_E(0)$, which would leave room for each state to operate a multiple of the capacity of El Capitan, even after two orders of magnitude of erosion.

We assess this as an encouraging sign for political feasibility: it does not appear that, in the immediately foreseeable future, a treaty would need to impose on strategically important assets like El Capitan.

⁷⁵ El Capitan's HPL mixed precision benchmark registered at 37%, a rate similar to frontier pretraining (Lawrence Livermore National Laboratory 2025).

6.4.7 Other categories

Another palatability feature we discussed in Section 3.3.1 was the option to create internationally governed scientific preserves. Such preserves' near-term size would likely be small compared to current commercial fleets, and it would appear highly difficult for covert evasion to occur there in transparent and replayable workloads, under the noses of the treaty authorities running the preserve. It would be important to guard against overt breakout scenarios, however, either by locating them in neutral states unlikely to participate in a breakout, or giving treaty monitors the ability to quickly and permanently incapacitate the chips located in a scientific preserve when the host state attempted to seize them, or both. Thus, it appears possible to ensure that scientific preserves contribute negligibly to the U.S. or China's covert evasion or overt breakout ledgers.

Finally, we discussed in Section 3.3.1 the possibility of states leaving strategic reserves under seal, monitored by other states or treaty authorities. Such reserves would, by design, be fully available to states in overt scenarios, so $\pi_B = \tau_B = 1$; thus, their undiscounted compute capacity would be added to the state's breakout ledger. However, it should be possible to implement such a plan so that the strategic reserves are verifiably powered off at all times while they remain under seal, so $\pi_E = 0$, thereby contributing nothing to a state's covert evasion ledger. This is what makes the strategic reserves an attractive, politically cheaper alternative to covertly hoarding a large reserve of chips.

The above analysis is intended to be illustrative, not exhaustive. A more complete analysis would include integrated consumer graphics cards, phones, CPUs, and edge devices such as hardware installed in vehicles. Based on our initial analysis above, however, we believe there is good reason for optimism that treaty constraints could be satisfied in the near future.

6.5 Durability assessment

Next, we discuss the primary determinants of the durability of a hardwired pause, in particular the degree to which it could durably guard against covert evasion and overt breakout. Because we cannot predict the rate of training efficiency gains after a pause, we measure durability as a function of how much treaty thresholds would need to erode before the U.S. and China would not both be able to stay below allowances of 25% of F_E and F_B ($\alpha_E = \alpha_B = 0.25$). That is, we ask when it would become difficult to balance the covert evasion ledger inequality

$$L_E(s) \leq \alpha_E \cdot F_E(t) = 1.6\text{MH}100e \cdot F_E(t)/F_E(0),$$

or the breakout ledger inequality

$$L_B(s) \leq \alpha_B \cdot F_B(t) = 44\text{MH}100e \cdot F_B(t)/F_B(0),$$

for $s = \text{U.S.}$ or $s = \text{China}$. The binding constraints on durability are the terms with the least headroom, specifically those for which reducing headroom within a period of years would pose a significant political or technical obstacle.

To estimate durability in time, we assume that annual post-pause training efficiency gains will run at or below the 57% estimate of Patel and Han (2026), corresponding to slightly slower growth than one order of magnitude gain every five years, or to the commonly cited 18-month doubling time famously associated with Moore’s Law for semiconductor miniaturization (Tuomi 2002). Under this scenario, a decade of durability would require achieving two orders of magnitude of headroom ($L_E(s) \leq 16\text{kH}100e$ and $L_B(s) \leq 440\text{kH}100e$ for each state) well before the end of the pause’s first decade.

Under the 2.7x annual training efficiency gain estimated by Ho et al. (2024), the treaty would be about half as durable. However, as discussed in Section 6.3, those gains were largely scale-dependent, and were observed during a period of rapid scaling growth.

In our analysis below, we do not assume that a treaty would be in force, with ledgers capped at a 25% allowance, beginning at the time of the pause date: in a near-term scenario, it is more likely that an informal agreement would eventually be replaced by a treaty after a process of negotiation that could take years. It might also not be realistic to enforce 25% allowances on day one; rather, a schedule might require states to reduce them to 25% over a period of years. These schedules could be somewhat flexible to the political and technological conditions, but as a working assumption we presume they could start close to 100% at entry into force and fall gradually to 25% by the time one order of magnitude of training efficiency gains had been incorporated.

6.5.1 Durably guarding against covert evasion

Dark compute is the term in the breakout and the covert evasion ledgers over which the treaty would have the least control, and it would also be the term most difficult to discount for covert evasion purposes. Based on our estimates from Section 6.4.3, it appears implausible that dark compute would pose an immediate covert evasion risk but, if undiscounted, it could plausibly pose a risk over time due to training efficiency gains. Under our high-cooperation scenario, the treaty ledger would initially have two orders of magnitude of headroom, while under our low-cooperation scenario it would only have one order of magnitude.

If the U.S. and China found it difficult to cooperate in the early years of a pause, they would have further chances to reduce dark compute stockpiles

before the initial headroom ran out. If they could not do better than our non-deniable concealment scenario, for example, the covert evasion ledger could survive roughly six-fold training efficiency gains before China’s covert capacity exceeded the fleet threshold. Achieving the low-cooperation scenario would be sufficient to get both states through the first order of magnitude of training efficiency gains, and achieving the high-cooperation scenario would be sufficient to achieve two orders of magnitude at a 25% allowance. Thus, a thorough global accelerator census would not be absolutely required on day one, but would need to be completed shortly after the first order of magnitude of training efficiency gains. However, we do not suggest this to encourage complacency, as an initial failed accelerator census could make it more difficult for a second census to succeed.

The most plausible fallback plan, if states could not manage to conduct a thorough census, would be efforts to detect covert training at the time it occurred. Such detection would be made more difficult owing to advances in distributed training, using methods like DiLoCo (Douillard et al. 2023), DiLoCoX (Qi et al. 2025), and more recent variants further extending distributed parallel training, but at present known methods would still require large compute clusters in order to train a model on the scale of $C_E = 2 \times 10^{28}$ FLOPs; see Sevilla (2025) for further discussion.

6.5.2 Durably guarding against overt breakout

The transitional inference service in the U.S. could plausibly pose a breakout risk much sooner, if its replacement by inference-only compute were delayed for a long enough period of time. If 57% annual training efficiency gains accrued in the 26 months that we project it might take for inference-only chips to begin production at scale, the U.S.’s 20% overt breakout capacity would rise to 53% unless most of the pre-pause training-capable stock had been replaced.

There are several potential ways to mitigate this exposure. The simplest would be to make a one-time exception for transitional inference service and wait for pre-pause training-capable hardware to be replaced by inference-only hardware, which we estimate would happen shortly thereafter. This could be accomplished by using a shrinking schedule $\alpha_B(t)$ for the allowance, which could shrink to 0.25 or another terminal value over a period of years as discussed above.

Alternatively, if parties to the treaty were not satisfied with the transition timeline as a mitigation, some of the hardware could be transferred to data centers outside U.S. territory at an earlier date. These could be located offshore or in space, in the custody of a neutral state, or in international custody, for example in scientific preserves. These options would only mitigate breakout risk

if it were understood that the owner state would not be able to recover these chips' training capacity after a breakout (cf. Dean 2026b). Under our near-term pause scenario, if the U.S. were willing to do this for half of its transitional inference service (which would not require a major commercial disruption if the chips kept operating in new data centers) then it could ensure that its breakout ledger did not exceed 100% of $F_B(t)$, and it could eliminate this term entirely once transitional inference service were phased out by inference-only service. Since all other terms have at least two orders of magnitude of headroom at a 25% allowance, phasing out transitional inference service by this time would be sufficient for the treaty to last for two orders of magnitude of training efficiency gains.

If most or all of the pre-pause training-capable compute used for transitional inference service were transferred to data centers in locations where they were judged to pose low breakout risk for any state, such as offshore or in space, they would still need to be weighed into states' covert evasion risk ledgers. If states parties were confident enough in verification methods, and in the plan to prevent recovery upon breakout, it is possible that pre-pause training-capable compute could continue to perform inference service for the remainder of its useful lifetime.

6.5.3 Extension options

In each of the above scenarios it is possible that the combination of training efficiency gains and dark compute would eventually make it impossible to balance the covert evasion ledger, with detection-based methods failing to make up the difference. In that case, the simplest way for pause-willing states to extend the lifetime of the agreement would be to declare their dark compute and place it in strategic reserves, but this might be politically difficult. An alternative option would be for an international consortium to jointly train a larger model to advance the frozen frontier C_F just enough to make policing C_E feasible once again.

6.5.4 Lithography proliferation

Missing lithography machines, supported by a complete indigenous covert supply chain, could significantly undermine the regime over a period of years by creating a continuous inflow of dark compute into one state. For example, Brown and Khan (2026) estimate that a single ArF-immersion DUV machine could contribute roughly 30k–47k Ascend 910C-equivalents—or 24k–38k H100e in our units—per month at state-of-the-art yields for 7nm precision using multi-patterning, significantly eroding durability. The regime would need to prevent

undeclared accelerator production by verifying the location of each advanced pre-pause lithography machine in each state from vendor service records and by dedicating sufficient ongoing intelligence efforts to detect any new undeclared manufacturing facility or covert supply chain. Such an effort would be highly plausible given the many subcomponents and materials required to service a fully functional lithography machine.

7 Implications of a hardwired pause

In this section, we discuss some of the implications of a hardwired pause, including both general implications of a frontier pause and specific implications of a hardwired pause relative to other ways of pausing the frontier. We make no claim to comprehensiveness, but we discuss several issues that appear especially salient: the macroeconomic impacts of a frontier pause with continued inference service (Section 7.1); the reduction of inference costs for consumers (Section 7.2); the economics of using MSICs, as an example of inference-only chips, for the hardwired pause (Section 7.3); impacts of the hardwired pause on concentration of power (Section 7.4) and continual learning (Section 7.5); and the need for other measures in addition to a hardwired pause (Section 7.6).

7.1 Would restraints on frontier AI crash the economy?

AI-related ventures have been heavily responsible for the recent strength of U.S. GDP growth. According to the Federal Reserve Bank of St. Louis, spending on information processing equipment, software, R&D (including AI R&D), and data centers accounted for nearly 40 percent of U.S. real GDP growth in the first three quarters of 2025 (Rubinton and Patro 2026). Take away that spending, some worry, and the economy will crash. Similarly, AI-related companies are heavily responsible for the high valuation of the S&P 500 stock index; they now account for over one-third of the total weight of the S&P 500 (Cembalest 2026; see also BlackRock 2026). Tank those share prices, observers warn, and you risk tanking the economy.

Unpacking these arguments, in the context of a scenario where restraints are placed on frontier AI, requires distinguishing the following questions. How would restraints on frontier model training affect AI-related spending growth, by hardware and software companies developing those models and by other companies utilizing their services? How would the supply side of the economy be affected—in other words, what would happen to productivity growth? How would financial markets be affected? And how would the reaction of financial markets impact the broader economy?

On the first question, the impact on AI-related spending, it is important to note that the St. Louis Fed's estimated 40% is at the upper end of the range. Other institutions such as Goldman Sachs estimate much smaller impacts, given that a significant share of spending is on imported chips and equipment (Goldman Sachs 2026). Spending on imports raises Taiwanese and Korean GDP, not U.S. GDP. Of physical investment, it is mainly the change in spending on data centers in the U.S. that would visibly affect U.S. GDP growth. Headlines notwithstanding, data center construction accounts for less than a fifth of total AI-related investment in the United States (Juniewicz 2026b; U.S. Census Bureau 2026).

AI-related spending that is not on imported chips, equipment and data centers tends to be on software licenses, enterprise software deployment and model creation. Spending on AI software, services, and models is projected to account for roughly 40 percent of AI-related spending worldwide in 2026 (Gartner 2026). A moratorium on training frontier models would discourage spending on model creation by definition. But there is increasing evidence that firms licensing and deploying AI are not mainly deploying frontier models (e.g. ICONIQ 2026; Kharazian 2026), since sub-frontier models, while cheaper, yield nearly the same productivity gains in most applications. Freezing the training of frontier models may then have little impact on licensing and deployment-related spending, and in turn on the demand for data center services. (Parenthetically, the greater threat to the demand for data center services may be open-weight models that can be run on enterprise hardware or even a laptop.) These same observations suggest that frontier model restraints would do little to reduce economywide productivity growth below its current rate.

There would presumably be a drop in the lofty stock prices of the “Mag7,” the seven ultra-large-cap tech and tech-adjacent firms that have driven the S&P 500.⁷⁶ But stock market corrections are no guarantee of recessions. (The economist Paul Samuelson famously observed that the stock market predicted 9 out of the last 5 recessions.) A decline in the price of the Mag7 stocks would cause some slowdown in consumption and investment spending. But the marginal propensity to consume out of stock market wealth is just 3 percent (Chodorow-Reich et al. 2021; see also Ludvigson and Steindel 1999). Imagine that the market valuation of the Mag7 of high-tech companies falls back to its beginning 2025 level—that is, that it falls by \$6.5 trillion. Three percent of \$6.5 trillion (\$195

⁷⁶ Namely, Alphabet, Amazon, Apple, Meta Platforms, Microsoft, Nvidia, and Tesla. In addition, one could imagine negative revisions by investors regarding the prospects of not-yet-public companies such as OpenAI and Anthropic currently contemplating initial public offerings. Those IPOs might be delayed as a result, which would similarly have a negative impact on the value of the S&P. But delay is more plausible than the more dire scenario of abandonment. And if we are right that such companies will still be earning revenues from the deployment of their models, then the negative impact is unlikely to be severe.

billion) is 0.6 percent of projected 2026 GDP, or about a quarter of projected 2026 GDP growth (Board of Governors of the Federal Reserve System 2026).

This is plausibly an upper bound on the fall in consumption spending. Less consumption (and investment) spending will mean less upward pressure on interest rates and less inflation. The Federal Reserve would react by reducing its policy interest rate (or moderating its rate of increase). In the decade ending with COVID, when interest rates were at or near zero, the Fed had limited ability to respond in stabilizing fashion (leading it to adopt unconventional monetary policies, whose efficacy is disputed). Now that its policy rate is nearing 4 percent, it has more scope for buffering a decline in spending.

The scenario where Mag7 share prices fall back to early 2025 levels is just a thought experiment. (AI-related companies do other things besides developing frontier AI models and selling their services. The premise that advances in frontier models account for most of the change in valuations since early 2025 is the motivation for this particular thought experiment.) Readers can scale up or down the \$6.5 trillion fall in share values as they prefer. But the same general conclusions follow, including the capacity of the Fed to buffer all but the largest shocks.

What about risks to financial stability? A fall in equity prices would be painful for investment funds like Situational Awareness specializing in AI-related investments. But a financial crisis will follow only if banks and other mainstream institutional investors are implicated, as they were in 2008. Hedge funds fail (“wind down”) all the time. Only when they are importantly linked to an institutional parent (as in the case of two Bear Stearns hedge funds in 2007) do systemic consequences follow. And there is little evidence today that such institutional parents are heavily exposed to AI equity risk.

Even if a fall in share prices owing to a ban on frontier model training is unlikely to create serious problems for the economy as a whole, it could create problems for specific firms. An example is Nvidia, which is invested heavily in producing training-capable accelerators for frontier models. Nvidia could pivot to building inference-only accelerators, for which there would still be a strong demand, but such a transition would be neither painless nor cost free, and the company would presumably not dominate that market as it does the market for general-purpose accelerators. More than 50 percent of Nvidia shares are held by U.S. investors (MarketScreener 2026), and the company currently accounts for roughly 8 percent of the S&P 500. Some 70 percent of all Nvidia shares are held by institutional investors, the largest holdings being those of Vanguard, BlackRock, Fidelity, Geode Capital Management, and State Street (Yahoo Finance 2026). These custodians hold shares on behalf of their mutual fund and ETF customers, so losses would be spread among retail investors.

There would be negative wealth effects, with implications for consumption, but not obviously risks to financial stability. Nvidia also provides credit guarantees and loan backstops for the construction of data centers, so were a pause to disproportionately affect its cash flow and share valuation there might be a significant negative impact on AI-related construction activity.

Also worrisome are developments in debt markets. With the cost of data center construction exceeding their sponsors' free cash flow, and with income from tenants still a distant prospect, much of this investment is being financed by borrowing from private credit funds (nonbank firms making privately negotiated loans). If the demand for data center services is less than expected, these loans may fail to perform, which is to say that borrowers will default.

The question then becomes who is invested in private credit funds. The popular image is of the same family offices and high-net-worth individuals who invest in venture capital and can afford losses. But the reality is that the vast majority of private credit finance is provided by institutional investors, including pension funds and insurance companies (Cai and Haque 2024).

Moreover, private credit funds have been taking proactive steps to tap this institutional money by acquiring insurance companies, whose asset portfolios they can load up with loans. Disappointing data center revenues, which cause private-credit loans to become nonperforming, could thus lead to serious problems for the insurance industry. This could play out in a number of ways. First, insurance companies' private credit parents could recapitalize them, assuming they have the funds. Second, state guarantee funds, to which healthy insurance companies contribute, could fund the recapitalization, although there will be no healthy companies in the scenario where the entire industry comes tumbling down. Third, the federal government could step in with a bailout, as it did for American International Group (AIG) in 2008.

In addition to borrowing from pension funds and insurance companies, private credit firms borrow from commercial banks, using their portfolio of loans as collateral, thereby leveraging their commitments and juicing their returns. So even if banks do not lend to or invest in AI firms and data centers directly, they still may be on the hook.

Finally, private credit firms originally built on a buy-and-hold model increasingly securitize their loans. They package those loans through special purpose vehicles that issue bonds backed by the associated cash flows, where the bonds are divided into risk tiers or tranches claiming first, second and third dibs on debt service payments. The so-called "mezzanine tranche" of medium risk bonds is then sold on to other asset managers. To assess the immediacy of crisis risk, we would have to know more about the riskiness of this mezzanine tranche and who holds it, which is not public information. The situation is reminiscent of

the roles of securitization, special purpose vehicles, tranching and opacity in the Subprime Crisis of 2007-8.

Tech firms, for their part, also create special purpose vehicles (SPVs) to disguise the impact of debt on their balance sheets. The SPV is legally distinct. It borrows to purchase real estate, land and servers, and then leases this infrastructure back to the tech firm under a multi-year contract. For the parent, a large one-time capital expenditure is transformed into a smaller lease payment spread over a period of years, with a smaller balance-sheet impact, which encourages debt-fueled capital spending. But to secure private equity and other funding, the tech firm parent must provide a residual-value guarantee, whereby it promises, if the appraised value when the tenant leaves falls short of the guaranteed minimum value, that it will step in and pay the difference (Oh 2026). The situation is reminiscent of the failure of Enron in 2001.

How would a moratorium on the training of frontier models affect these risks, and what should be done about them? If it is correct that demand from licensees is mainly for sub-frontier inference, then a freeze on frontier model training would have little impact on data center revenues, on the performance of private credit loans for data center construction, and on the tech firms' SPVs. Were a freeze to redirect additional demand from frontier model training to inference service, then this conclusion would be reinforced. Only in an alternative scenario where a moratorium on the training of frontier models led to a sharp decline in licensing and in demand for data center services would a freeze increase the risk of serious financial dislocations.

What should regulators do to monitor and limit the risks to economic and financial stability that might arise in the event of a significant slowdown in the demand for data center services? Bank regulators should closely monitor bank lending to private credit firms and limit concentrated exposures to this sector. Insurance regulators should question private credit purchases of insurance companies and be wary of potential conflicts of interest. Financial regulators should require more disclosure by financial entities that are securitizing private credit loans and seek more data on who is holding the resulting securities. They should more strongly regulate the special purpose vehicles that AI firms use to limit the balance sheet impact of debt, since access to those accounting devices encourages their borrowing, which contributes to the build-up of financial risks.

The bottom line is that any decline in investment in data center facilities and equipment, and any decline in consumption spending owing to lower equity valuations, both pursuant on a frontier-model pause, are likely to be limited, because the declines in spending in question will be limited, and because their macroeconomic effects can be buffered by monetary policy. Only if declining data center revenues cause distress in debt markets and if that distress is allowed to

infect institutional investors could the effects be seriously destabilizing. This last possibility is something for which regulators could, and hopefully will, prepare.

7.2 Reduction of inference costs and related economic effects

A likely economic benefit of a hardwired pause is the economic gains from the reduction of inference costs. The new regime that halts frontier model training compels frontier labs and the compute ecosystem to focus largely on reducing inference costs for existing frontier models available to consumers.

There are several distinct economic forces that could result in lower inference costs. One is the reallocation of compute resources from training to inference. According to SemiAnalysis (Xie et al. 2026), Anthropic and OpenAI are committing 62% of their combined compute resources to research, development, and training of new models. Other lagging AI labs with less market share and a strong desire to catch up to the frontier may currently devote an even larger fraction of their compute to training. Thus, a pause would result in reallocation of substantially more compute to inference, resulting in reduction of inference costs.

On the chips/server hardware/data center infrastructure side, redirecting industry focus towards inference could also reduce hardware costs. This comes from directing resources away from training solutions focused on larger memory and faster interconnects to varieties dedicated to the reduction of cost per token and latency. Competition from specialized inference chip providers with MSICs dedicated to popular models can also reduce cost (see Section 7.3). Such reductions in hardware costs and increases in energy efficiency can also alleviate the pressure on AI labs/cloud service providers' margins, making it more conducive for them to agree on a pause.

As a result, we should expect the price and actual energy cost of current frontier models to decrease dramatically after full implementation of a hardwired pause. Future innovations expected from model providers and hardware manufacturers dedicated to cutting inference cost and improving speed will accelerate the gains. This could lessen incumbent labs' pressure on profit margins and encourage them to cut prices, bringing consumer welfare gains. Thus, the principal near-term economic welfare effect of a pause could be an increase in consumer surplus from reduction of token prices as well as increased speed and supply. On the other hand, this would have to be weighed against the expected welfare effects of continuing to advance AI capabilities, about which there is significant disagreement.

7.3 Economics of MSICs

As described in Section 1, the hardwired pause allows the production of inference-only chips for running whitelisted AI models. As an example of an inference-only

chip, we cited certain model-specific integrated circuits (MSICs), which can only run inference on a single model. These are not the only type of inference-only chips that could be used, but since MSICs are already being developed without a pause, in this section we focus on MSICs.

The possibility of using MSICs leads to two economic questions:

1. Assuming that there was a frozen whitelist of models and that both MSICs and *flexible accelerators* (which can serve any model) were allowed for inference on whitelisted models, what allocation of inference between MSICs and flexible accelerators would minimize total costs for a given volume and distribution of demand for inference?
2. Assuming that there was a frozen whitelist of models and *only* MSICs were allowed for inference, how much “economic inefficiency” would the restriction to MSICs entail?

The downsides of using an MSIC for a specific model are (i) the up-front, non-recurring engineering (NRE) cost of designing and setting up production of the MSIC and (ii) the fact that an MSIC for a specific model cannot be transformed into an MSIC for a different model in response to future shifts in demand for models on the fixed whitelist as consumer tastes change or if a model is de-whitelisted (see Sections 7.3.3–7.3.4). The economic upside of creating an MSIC for a specific model is the much lower operating cost of inference using an MSIC, compared to the cost of inference using a flexible accelerator. In this section, we produce some preliminary “back of the envelope” estimates of how these tradeoffs might balance out in the long run.⁷⁷ The specific numerical values of the estimates are not meant to be taken seriously; the goal is rather to get a qualitative sense of how the tradeoffs might balance out in the long run. We hope that economists will do more sophisticated economic modeling in future work (cf. Dong et al. 2022).⁷⁸

7.3.1 MSIC adoption in a mixed MSIC-flexible world

Assume there is a whitelist of finitely many models. We will consider a simple rule for deciding whether to build an MSIC for a given model on the whitelist, based on the following data:

- D_M is the expected share of total inference demand (as a percentage) that will go to model M on the whitelist. We abstract away from the fact that

⁷⁷ Associated code is available at <https://github.com/Berkeley-AI-Risk/econ-of-MSICs>.

⁷⁸ E.g., by estimating a discrete-choice model of demand across models, which would capture how demand responds to prices and permit welfare calculations (see Demirer et al. 2026, on the market for LLM inference).

building an MSIC for M could lower its inference price and thereby increase demand for that model.

- NRE is the fixed cost of development (in dollars) of an MSIC for a model (assuming for simplicity this is the same for each model on the whitelist), called the *non-recurring engineering* cost for the MSIC.
- S is the savings of moving model inference from flexible accelerators to MSICs, as a percentage of the cost of inference on flexible accelerators (again assuming for simplicity this is the same for each model).
- B is the expected (discounted) lifetime cost of serving all inference demand that it is profitable to serve using only flexible accelerators, which we call the *baseline* cost of serving inference.

Suppose that an MSIC will be developed for model M if and only if the expected (discounted) lifetime cost-savings exceeds the fixed cost NRE:

$$S \cdot D_M \cdot B > \text{NRE}.$$

Estimates for the savings term S range from 60% to 95%.⁷⁹ Estimates for the fixed cost NRE of developing an MSIC range from \$10M to \$500M.⁸⁰ If we imagine that all models currently available on OpenRouter pass safety scrutiny

⁷⁹ Taalas claims its hardwired chips are “an order of magnitude faster, cheaper, and lower power” than GPU-based serving, with hardware “20X less” costly to build (Bajic 2026; also see Prickett Morgan 2026; Ward-Foxton 2026a). Its published price comparisons imply savings of about 62% (serving DeepSeek R1 at 7.6 cents per million tokens on simulated Taalas hardware vs. 20.2 cents per million tokens on throughput-optimized Nvidia hardware) or 79% (serving Llama 3.1 8B at 0.75 cents vs. 3.63 cents). Liu et al. 2026 project that in a high volume regime, their hardwired chips can serve inference at roughly 1–2.5% of the three-year total cost of ownership of comparable H100 clusters, corresponding to a cost savings of roughly 98%, even including the fixed cost NRE; since we are excluding the fixed cost from S , the cost savings in our terms would be even higher. On the other hand, S may be lower than the chip-level savings above, since a fleet of flexible accelerators can be redirected to whichever models are in demand at a given time, whereas a fleet of MSICs must use price changes to respond to changes in demand. The width of our range for S takes this into account, and we discuss changing demand below.

⁸⁰ The upper part of the range comes from Liu et al. 2026, which estimates the NRE (design and fabrication setup) for their hardwired chips to range from \$38M to \$462M across model sizes from 8B to 1T parameters. The lower end of the range is suggested by Taalas’s HC1: Taalas says they spent \$30M on R&D for their first chip, the HC1 (Prickett Morgan 2026), and converting that design to one for a new model requires changing only two of the chip’s metal layers (ibid.), so the marginal cost for an additional model, once the pipeline is set up, is plausibly well below \$30M. Our \$10M floor is chosen in the back-of-the-envelope spirit of this section, not based on any direct evidence from Taalas. Since NRE scales with model size (Liu et al. 2026), high-demand frontier models would require design fees near the top of the range (though more entrants into the MSIC design business might drive down NRE) while smaller open-weight models would require design fees near the bottom. A size-dependent

and are whitelisted,⁸¹ then we can use OpenRouter data⁸² to give us a scenario for the distribution of demand shares D_M : for each model M served by OpenRouter, we simply take D_M to be the share of total tokens of inference served on model M from August 5-11, 2026 (below we extend our analysis to allow for changing demand over time).⁸³ Finally, for the expected (discounted) lifetime cost of serving all inference demand that it is profitable to serve using only flexible accelerators, we consider a range of scenarios for B from \$500B to \$2T.⁸⁴

NRE_M would therefore produce results between the \$10M and \$500M rows of each panel in Figure 7.1.

⁸¹ This is just an assumption for illustration, not a prediction that all current models would in fact pass safety scrutiny.

⁸² OpenRouter serves a very small slice of the inference market, on the order of \$3B per year (projecting from our sampled week's traffic valued at posted prices). It does not include the massive volume of inference served by model providers, e.g., OpenAI, Anthropic, Google, etc. So we may be underestimating the concentration of demand among a few models. Higher real-world concentration figures would strengthen the conclusion of this section.

⁸³ Throughout we restrict to the 315 text-output models in OpenRouter's catalogue with observed traffic for that week; non-text (e.g., embedding, image, audio) models and non-catalogued models, representing 0.6% of tokens, are not included. We treat each such model as requiring its own MSIC design. In practice, variants sharing weights could share a design, so the design counts in Figure 7.1 are model-level approximations. We also note that strictly D_M should be model M 's share of inference cost, so that $D_M \cdot B$ is model M 's share of the total cost of serving inference. Estimating D_M by token shares assumes that the cost per token of inference is roughly the same across models, which is not true in practice. An alternative methodology is to weight each model's number of tokens by OpenRouter's published price per token for that model. Re-running the analysis with this methodology does not change the main conclusions to follow. Under this alternative methodology, the worst-case unserved share of inference spending in the pure MSIC world of Question 2 is only about 2% (regardless of how free and discounted variants of models, which serve about 7% of all tokens, are priced).

⁸⁴ Deloitte (Stewart et al. 2025) projects inference at roughly two-thirds of all AI compute in 2026, and the Goldman Sachs Global Institute's (2026) baseline scenario puts 2026 AI capital expenditure at roughly \$765B, which is consistent with the nearly \$700 billion in combined 2026 capital expenditure announced by Alphabet, Microsoft, Meta, and Amazon (Elias 2026). If inference's share of spending is anywhere near its share of compute, spending in the hundreds of billions on inference *in a single year* is plausible. Thus, our range of values for the discounted *lifetime* cost B may well be conservative unless one has an extremely high temporal discount rate. On the other hand, if there is uncertainty about whether the pause will be upheld over the investment horizon, that would make flexible accelerators relatively more attractive, since flexible accelerators could serve new frontier models developed in a restarted race. In our model, such uncertainty could enter as a multiplicative correction to the expected savings, which is equivalent to lowering B . Thus, scenarios with lower B can also be read as scenarios with pessimism about the pause's durability. Even under a very low B of \$100B (we could imagine a scenario in which firms expect the pause to fall apart in a year, at which point the race would re-start and current models would become obsolete), the worst-case cell of Figure 7.1 would still place 86.5% of tokens on MSICs.

Figure 7.1 shows the results based on the OpenRouter data, assuming $S = 85\%$ (the share of tokens served on MSICs is largely insensitive to the choice of S within the range from 60% to 95%⁸⁵). The upshot is that assuming the long-run economic model above, the vast majority of inference tokens on whitelisted models would be served on MSICs. Moreover, there would be major cost savings over serving all inference on only flexible accelerators.⁸⁶ Models with insufficient demand would not justify dedicated MSICs, but collectively such models serve a small share of total inference. Again note that our analysis holds the demand shares D_M fixed, but since serving a model on an MSIC lowers its cost and therefore its price (assuming competition), demand would shift toward MSIC-served models. In this respect, the estimates in Figure 7.1 for the share of tokens served on MSICs are lower bounds.

The decision rule above determines whether to build an MSIC based on minimizing total cost for the assumed volume and distribution of inference demand. The question of when a market left to its own devices would build an MSIC is more complicated, since firms might not fully capture the savings of using an MSIC, whether because of competition among MSIC developers or for other reasons. But even under the worst-case parameters of Figure 7.1, the expected savings of using the MSIC exceed the fixed cost of developing it by roughly 120x for the most popular model and by at least 10x for every model in the group of most popular models that collectively serves three-quarters of all tokens. Thus, the incentives to create MSICs for the models serving most of the inference would likely exist even if only a small fraction of the savings could be captured by any single investor.

It is instructive to compare this analysis with the historical record concerning hardware for Bitcoin mining. In a span of four years, Bitcoin mining went from CPUs to GPUs (by 2010) to FPGAs (by 2011) and finally to ASICs (by 2013) (Taylor 2017); an early Bitcoin mining ASIC had 4.4 times the hashing throughput per dollar of hardware cost and was 40 times more energy efficient than a contemporaneous GPU (Taylor 2017, p. 64). In the case of Bitcoin, the economic incentives toward ASICs may be greater than in the case of AI models, since there is essentially just one algorithm for Bitcoin mining, which one would naturally want to hardwire into an ASIC for maximum efficiency. However, the example of AMD's pending acquisition of Taalas (AMD 2026) suggests that even without a frozen whitelist of AI models, there are economic incentives in favor of *some* development of MSICs. The analysis of this section suggests that in the

⁸⁵ Readers can explore other parameter settings using the interactive applet at <https://econ-of-msics.streamlit.app>.

⁸⁶ If there were no pause and therefore no frozen whitelist, then flexible accelerators might still dominate the market if models continued to be quickly replaced by more capable models.

presence of a frozen whitelist, the economic incentives in favor of MSICs could be much greater.

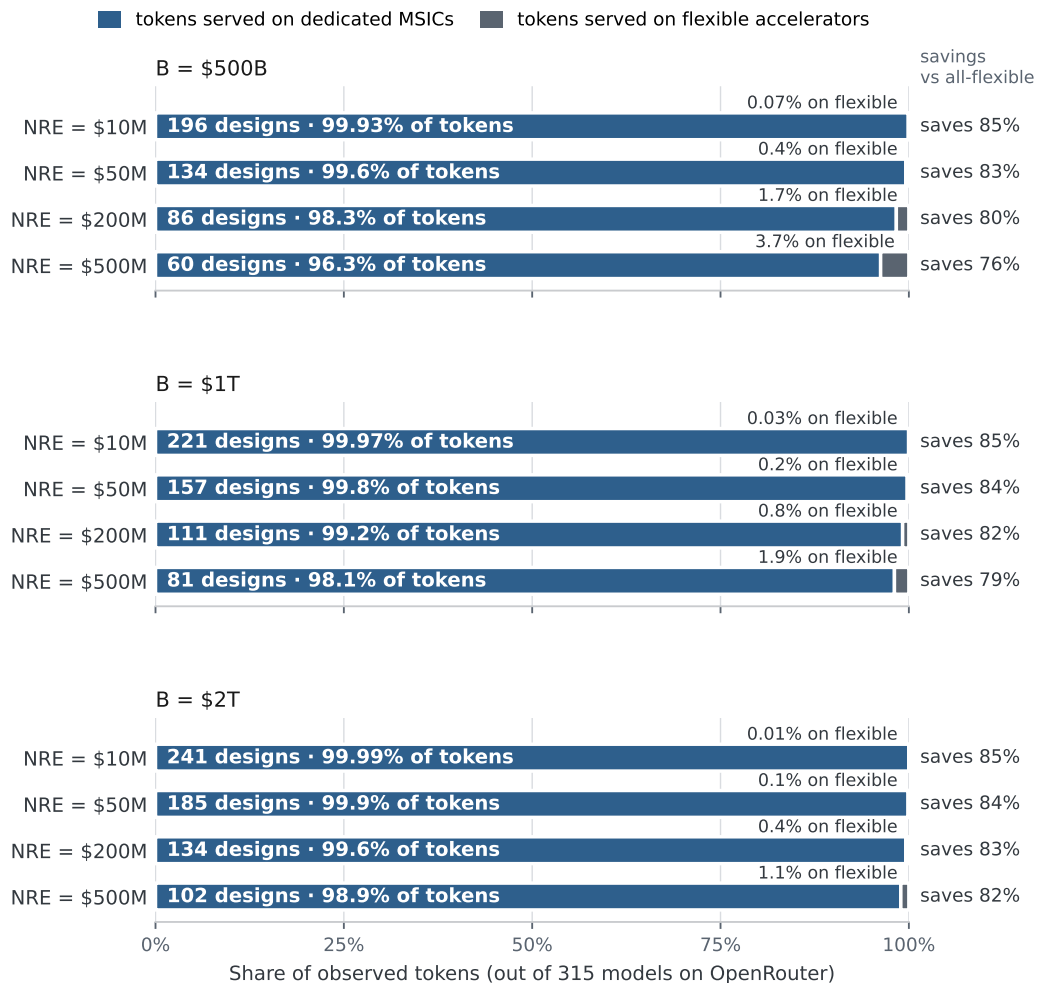


Figure 7.1. In a world with a frozen model whitelist, the vast majority of tokens would be served on MSICs according to the economic model of this section. The results shown assume the MSIC savings rate of $S = 85\%$, use OpenRouter data to give a scenario for the demand shares DM , and vary the fixed cost NRE of designing an MSIC and the lifetime cost B of serving all profitable inference demand using only flexible accelerators. Each bar indicates how many MSIC designs would be created (for the 315 text-output models with traffic on OpenRouter from August 5-11, 2026), what percentage of tokens would be served on MSICs, and the percentage cost savings vs. running all inference on flexible accelerators.

7.3.2 The cost of a pure MSIC world

What would happen if inference were allowed *only* on MSICs? What would be the resulting economic inefficiency relative to allowing both MSICs and flexible accelerators, in the sense of increased cost of serving inference and loss of consumer utility due to some models not being served?

All of the models that justified dedicated MSICs in the “mixed MSIC-flexible world” of the previous subsection would still justify MSICs in the “pure MSIC world”, but in the pure MSIC world some additional models might justify dedicated MSICs. The reason is that in the pure MSIC world, it is not necessary to clear the hurdle of the cost-savings condition relative to flexible accelerators from the previous subsection. It is sufficient that it would be *profitable* to create an MSIC in order to serve a model, even if that would be less profitable than using flexible accelerators to serve the model, if that were permitted (which it is not in the pure MSIC world); and profitability was already implicitly a necessary condition of creating an MSIC in the mixed MSIC-flexible world.⁸⁷ Thus, in the pure MSIC world, at least as many and plausibly strictly more models would justify dedicated MSICs.

The key difference is that in the pure MSIC world, if an MSIC is not created for a model, then demand for inference on that model is unserved. By the reasoning in the previous paragraph, Figure 7.1 gives an upper bound on the *number of models* that would go unserved and the *share of demanded tokens* that would go unserved. While the number of models that go unserved could be substantial, the share of demanded tokens that would go unserved would be at most a few percent under the worst-case assumptions of Figure 7.1 (and the unserved share of inference spending would be about 2%, as discussed in Footnote 83).

Of course, some consumers whose favorite model has insufficient demand to justify an MSIC would switch to demanding tokens from their *next favorite* model, which might already have an MSIC based on aggregate pre-switch demand or could justify a new MSIC based on aggregate post-switch demand. The inefficiency of the pure MSIC world relative to the mixed MSIC-flexible world depends on what fraction of consumer utility is preserved after switching to their next favorite model among those that justify an MSIC (cf. Hausman 1996; Conlon and Mortimer 2013). Given the similar capabilities of frontier models and the fact that even in the worst-case scenario of Figure 7.1, 60 models are served by MSICs, consumers may preserve a substantial fraction of their

⁸⁷ In the mixed MSIC-flexible world, B is the expected (discounted) lifetime cost of serving all inference demand *that it is profitable to serve* using only flexible accelerators. Thus, $D_M \cdot B$ is the expected (discounted) lifetime cost of serving all inference demand for model M *that it is profitable to serve* using only flexible accelerators. We assumed that an MSIC would be created only if it would lead to a cost savings, which would imply even more profit.

utility by such a switch, though the loss may be greater for users for whom no model with an MSIC performs well in their native language. Without studies of consumer utility functions for tokens from different models, we are not able to quantitatively estimate the utility loss implied by 3.7% of tokens (to take the worst-case scenario from Figure 7.1) switching from the consumers' favorite models to their next favorite models that are popular enough to justify an MSIC. We encourage economists to study consumer utilities for tokens in future work.

7.3.3 Shifting demand among whitelisted models

However, what happens if changing consumer tastes lead to shifts in the distribution of demand among whitelisted models? If many copies of an MSIC are created for a once popular model, but this model becomes less popular in the future (even though the whitelist remains frozen, so no new models appear as competitors), then those chips cannot be turned into MSICs for a more popular model (cf. the literature on investment in flexible versus dedicated production capacity, as in Fine and Freund 1990; Van Mieghem 1998). A fleet of MSICs could adjust its distribution over time by increasing production and delaying retirement of chips for models with rising demand and doing the opposite for models with falling demand, but it could not do so instantly as a fleet of general-purpose accelerators could. This higher latency in responding to shifting demand would likely be a source of some inefficiency, though it is hard to estimate how much from current market data because, unlike today's world, a paused world would see very little churn in what model options were available to customers from one month (or year) to the next. On the other hand, general-purpose accelerators pay a steep cost for their flexibility, as reflected in the 60-95% cost savings for MSICs.⁸⁸

⁸⁸ One methodology to estimate the extent of demand shifts under a frozen whitelist is the following. Fix a particular set of models, like the top 50 models on OpenRouter as of an initial time t , and then consider the subset of those models that still appear in the available data on OpenRouter at a later time t' (we obtain the historical top 50 rankings from the Spectral Finance 2026, GitHub repo). We can then look at the probability that an inference token goes to a particular model M in the subset at time t , conditional on it going to any model in the subset at time t . Next, we can look at the analogous distribution of conditional probabilities at time t' and see how much change there is between the two distributions. Applying this methodology to archived single-day snapshots of the OpenRouter rankings for the pairs (January 1, 2025; June 1, 2025), (June 1, 2025; January 1, 2026), and (January 1, 2026; June 4, 2026) yields 17, 15, and 12 models in the respective subsets. The total variation distance between the distributions at t and t' was 70%, 40%, and 23%, respectively. These figures are biased in both directions: restricting to models that remain in the top 50 at t' (since OpenRouter does not provide historical data for individual models outside of the top 50) excludes the largest demand declines, which biases the measured shift to be smaller, while the appearance of new models over time draws demand away from models in the subset

If we continue to assume $S = 85\%$, then serving virtually all whitelisted inference on MSICs costs roughly 15% of B , plus design fees ranging from a fraction of a percent to a few percent of B . Thus, even if demand shifted so much that *all chips in the initial fleet of MSICs* were turned into obsolete write-offs (say because 20% of initial demand migrated every year for five years to models not originally covered by MSICs, with price cuts unable to win back any consumers) and had to be fully replaced with new MSICs (requiring roughly the same number of designs as the original MSIC fleet), the cumulative resource cost would be roughly 30–40% of B (including a second round of design fees),⁸⁹ representing a 60–70% cost savings relative to a pure flexible world.

Another concern is daily fluctuations in demand. As an extreme example, suppose there is a model M popular in the U.S. that receives no demand during the U.S. nighttime. Thus, the MSICs for model M may achieve 100% utilization during the U.S. daytime but 0% utilization during the U.S. nighttime, for an average utilization of 50%. By contrast, a flexible accelerator could achieve 100% utilization during the U.S. daytime by serving model M and 100% utilization during the U.S. nighttime by serving a model M' popular in China, for example. Then if the cost per token of an MSIC is 15% of that of a flexible accelerator *assuming 100% utilization of both*, the cost per token of an MSIC rises to 30% if the MSIC can only achieve 50% utilization, assuming conservatively that halving MSIC utilization doubles its average cost per token (excluding NRE). This would still represent a 70% saving for MSICs over flexible accelerators. Moreover, the utilization gap between MSICs and flexible accelerators could be smaller if M retains some demand across time zones or if scheduled inference jobs with M could be run overnight.

Finally, the loss of efficiency in moving from the mixed MSIC-flexible world to the pure MSIC world would have to be weighed against the benefits of the pure MSIC world in facilitating a pause. If the probability of catastrophic outcomes for humanity were much lower in a pure MSIC world (due to more reliable prevention of frontier AI training) than in an alternative world with flexible accelerators, this might swamp considerations of the economic frictions in dealing with shifting demand between models.

disproportionately, which biases the measured shift to be larger. Under a frozen whitelist, the latter cause of demand shifts among whitelisted models would be absent.

⁸⁹ This simplified estimate pretends that the initial fleet delivered no service before being written off. A more realistic discounted model of gradual replacement of MSICs would imply even greater savings relative to the fully flexible fleet.

7.3.4 Temporary or permanent de-whitelisting of models

One possible concern about a pure-MSIC world is that if a model gets de-whitelisted (a kind of extreme case of a demand shift, where demand goes to zero by decree rather than due to changing tastes), then its MSICs are stranded capital. In this subsection, we consider the extent of this concern. The takeaway is that the possibility of de-whitelisting does not rule out the use of MSICs as inference-only chips for a hardwired pause; the cost savings from MSICs may outweigh losses from de-whitelisting, and insurance could help manage the financial risk of de-whitelisting for firms that own and operate MSICs.

First, we must distinguish between temporary and permanent de-whitelisting. When a model is de-whitelisted due to a discovered jailbreak, it might be re-whitelisted after the development of appropriate external guardrails such as an input classifier or safety harness, which inference providers would be required to run when serving the model. In this case, de-whitelisting is temporary, like most automobile recalls; the MSICs are temporarily idle but not permanently stranded.

Permanent de-whitelisting would be more costly. Suppose it is discovered that a whitelisted model poses a safety risk for which no external guardrails are adequate, so the only choice is to retire the MSICs for the model. In the worst case, suppose this happened to the most popular model. In the OpenRouter data used above, the most popular model serves 14% of all tokens, but let us suppose the most popular model overall (not just on OpenRouter) serves 30% of all tokens, i.e., $D_M = 0.3$. The (discounted) lifetime cost of serving the model using flexible hardware is therefore $0.3 \cdot B$, whereas the (discounted) lifetime cost of serving the model using its MSICs is $(1 - S) \cdot 0.3 \cdot B$. To estimate the stranded capital cost, we introduce three more parameters, two of which are a function of the time t of de-whitelisting:

- K is the **chip-capital share**. This is the share of the (discounted) lifetime cost of serving the model (not including NRE) using MSICs that is spent on the chips themselves.
- $W_M(t)$ is the **stranded-chip fraction**. This is the book value of model M 's installed MSIC chips at time t , divided by the total (discounted) amount that would have been spent buying MSIC chips for M over its lifetime if it had remained whitelisted.
- $V_M(t)$ is the **unamortized design-cost fraction**. This is the unamortized share of model M 's design cost (NRE) at time t , where the design cost is amortized in proportion to tokens served with M .

Then the stranded capital cost is

$$(W_M(t) \cdot K \cdot (1 - S) \cdot D_M \cdot B) + (V_M(t) \cdot \text{NRE}),$$

where the first summand is due to wasted chips and the second is due to wasted engineering. Let us consider some cases for when de-whitelisting occurs.

Early de-whitelisting. If de-whitelisting happens shortly after production of M 's MSICs begins, then W_M is close to 0, since not many chips have been built. On the other hand, in the worst case where the NRE was spent on a chip platform that is not at all reusable—it is useful for only model M —then V_M is close to 1 in this scenario, since no amortization has occurred right after production begins, before tokens have been served.

De-whitelisting after initial large-scale production. If de-whitelisting happens right after the first major wave of MSIC production for M , then W_M is roughly the cost of the chips in the first wave divided by the (discounted) cost of the chips in all the waves that would have been produced had de-whitelisting not occurred. Again, in the worst case where the chip platform is not reusable for other models, V_M would be close to 1 since few tokens have been served and hence little amortization has occurred.

Later de-whitelisting. If de-whitelisting happens a while later, then both W_M and V_M might be significantly less than 1. And if the NRE went to a reusable chip platform, like that of Taalas, this would further reduce V_M .

For a scenario in which the most popular model with $D_M = 0.3$ gets de-whitelisted some way into the pause, suppose $W_M(t) = V_M(t) = 0.5$. Suppose our earlier variables are $S = 0.85$, $B = \$2\text{T}$, and $\text{NRE} = \$500\text{M}$ (the top of our range). Let us suppose, adversarially, that $K = 1$ (whereas in reality energy costs imply $K < 1$). Then the stranded capital cost is approximately \$45B, where the part coming from wasted NRE is negligible. We can put this into perspective in several ways. First, for comparison, the FAA's temporary grounding of Boeing's 737 MAX cost Boeing more than \$20B in direct costs (Robison 2021). Second, let us compare the de-whitelisting scenario in a world in which M is served using MSICs versus a world in which M is served using flexible accelerators, which can simply switch to serving other models when M is de-whitelisted. Let f be the fraction of the model's lifetime demand served before the de-whitelisting at time t . If we live in a world in which M is served only by flexible accelerators, we would spend $f \cdot D_M \cdot B = f \cdot 0.3 \cdot \$2\text{T} = f \cdot \$600\text{B}$ on serving M before de-whitelisting. If we live in a world in which M is served only by MSICs, we would spend $f \cdot (1 - S) \cdot D_M \cdot B = f \cdot 0.15 \cdot 0.3 \cdot \$2\text{T} = f \cdot \$90\text{B}$ plus the stranded capital cost of \$45B for a total of $(f \cdot \$90\text{B}) + \45B before de-whitelisting.⁹⁰ It

⁹⁰ Here we divide the expenditure on M 's MSICs into two parts: the part that paid for tokens served before de-whitelisting, which is the first term, and the part that paid for tokens that will never be served due to de-whitelisting, which is the stranded capital cost.

follows that if $f > 9\%$,⁹¹ then even with the stranded capital, serving model M with MSICs before de-whitelisting is cheaper than serving model M with flexible accelerators before de-whitelisting.

The firms that own and operate MSICs incur the losses when a model is de-whitelisted, which raises the worry that they will lobby heavily against any de-whitelisting. To reduce this pressure, one possibility would be to require firms to buy insurance covering losses from de-whitelisting as a condition of their license to serve inference with MSICs.⁹² The insurance premiums would depend on the book value of the MSICs covered and the assessed risk of de-whitelisting, helping to limit the incentives of firms to overinvest in MSICs for models at risk of de-whitelisting. The insurance fund would compensate a firm for the book value of its MSICs in the event of de-whitelisting, but only upon return of the MSICs to authorities. The treaty authority could serve as the insurer of last resort. Coverage would carry a deductible, so operators would maintain an interest in the likelihood that the models they serve will remain safe. Note that the purpose of the insurance requirement would not be to make operators whole but rather to make the treaty authority's whitelisting decisions more independent of firms' financial interests.

7.4 Impacts on concentration of power

Questions about how AI might shift the distribution of power have permeated discussions about the modern AI industry since its infancy. Bostrom (2014) discussed at length the possibility that a superintelligent AI system could rapidly become a permanent global dictator or “singleton;” Zuboff (2019) discusses how data-driven AI systems fueled by surveillance capitalism can concentrate power in the hands of tech platforms; and Davidson et al. (2025) discussed the potential of AI to contribute to sudden coups. Several AI companies, most explicitly OpenAI, were purportedly founded to reduce the risk that other AI companies would become too powerful (Hao 2025). As AI grows more powerful, the fear that it or those who control it could permanently concentrate power could itself be destabilizing (Hendrycks et al. 2025; Wright 2026); OpenAI CEO Sam Altman has referred to AI as having a “real ‘ring of power’ dynamic” that “makes people do crazy things” (Altman 2026a).

⁹¹ Amortizing design cost in proportion to tokens served implies that $V_M = 1 - f$, so our assumption of $V_M = 0.5$ already entails $f = 0.5$, far above 0.09.

⁹² Note that operators of nuclear power plants are required to buy insurance as a condition of their licenses (U.S. Nuclear Regulatory Commission 2026). Insurance can also cover costs incurred because of government-mandated product recalls; for example, such insurance is available for companies that produce automotive components (Marsh 2014, 5).

AI pause proposals address this “ring of power” dynamic by preventing (or at the very least delaying) any such ring from being forged in the first place. The AI industry has already shifted power significantly. The push to embed AI across society has concentrated power and influence in the hands of a small group of AI companies and their leaders (Brennan et al. 2025). Addressing this problem and preventing further concentration of power have been cited as key motivations by some pause proponents (see, e.g. Larsen et al. 2026; Office of Senator Bernie Sanders 2026).

Any plan for how humanity should address the challenges posed by AI must contend with questions about how it might affect the distribution of power; this is also true of plans to defer regulation, since doing so might lead to heightened risks that necessitate more onerous restrictions later. AI models that suddenly empower a large number of individuals to design new viruses or launch sophisticated cyberattacks could lead to calls for widespread surveillance just at the moment when AI permits more comprehensive evaluation of surveillance data than was previously possible (cf. Bostrom 2019).

A hardwired pause in and of itself would place no new restrictions on how people use their computers, save for content-neutral restrictions on the training of large enough AI models. If it were successfully implemented before widely available models became too dangerous, a hardwired pause could help to preserve liberty by forestalling the need for surveillance measures to prevent people from downloading and using those models.

On the other hand, a hardwired pause would involve imposing new constraints on liberties that individuals and businesses currently enjoy, such as the liberty to buy an unrestricted number of accelerator GPUs, to train AI models of unlimited size, or to operate chip fabs and data centers without being subject to inspections. While these impositions are by no means cost-free, they do not seem to us qualitatively more onerous than constraints that have been placed on other technologies when those technologies grew powerful enough to be dangerous. In the past, scientists could own as much uranium as they desired, people could drive their cars as fast as they wanted on the highway, and chemical manufacturers could produce phosgene without being subject to inspections, but today these activities are regulated by the government and international treaty organizations, with criminal prosecution for some violations. Impositions on these particular liberties have come to be seen by most as acceptable tradeoffs for the added security that regulation of dangerous technology can bring.

Another serious concern one might have about a hardwired pause is its potential to severely and even permanently entrench *market power* with the incumbent AI developers who have the most dominant market positions today, especially OpenAI and Anthropic. Like any regulatory body, the body responsible for

whitelisting models would potentially be vulnerable to industry capture, and concentration of the market in a few AI models would come with the additional risks of entrenching the ideological or commercial biases of the people who built them, or of proliferating any loyalties, vulnerabilities, or backdoors in the models (Buyl et al. 2025; Lehr et al. 2026). Any AI governance regime must take such concerns seriously: whitelisting processes and other safety regulations must be independent and as transparent as possible, and monopolies or oligopolies imposed by government fiat must be regulated like other public monopolies to protect consumers and prevent exploitation of market power. Successfully navigating these issues would be critical to the legitimacy, and consequently to the durability, of a hardwired pause.

7.5 Consequences of receding knowledge cutoffs

A pause on frontier training that involves freezing model weights would prevent models from being updated to change the cutoff time of their pretraining data (which we refer to informally as a “receding knowledge cutoff”). It also raises several issues, including vulnerabilities, data poisoning, and locked-in values. More generally, it would prevent models from being trained on a continual or even continuous basis, as frontier AI companies aspire to do (Favaro and Clark 2026; Behrouz and Mirrokni 2025; Pachocki 2026).

Receding knowledge cutoffs could have costs in terms of model accuracy and bias (Dai et al. 2025). However, since retraining models is costly even in today’s pre-pause regime, the receding knowledge cutoff issue is not a new problem. AI developers today use methods such as internet search and Retrieval-Augmented Generation (Lewis et al. 2021) to improve up-to-dateness while minimizing costly retraining.

A related issue, discussed in Section 7.3.4, is the potential for new vulnerabilities to be discovered in existing models. Frontier AI models are typically deployed in tandem with “guardrails” consisting of smaller models that monitor their input, output, and/or internal computations; and vulnerabilities are typically patched by updating a model’s guardrails rather than the frontier model itself. For guardrails that are small enough and do not monitor internal computations, this practice could continue freely under a hardwired pause. Monitoring internal computations (called “probing”), however, would require modifications to inference chips that could potentially facilitate GEMM-laundering (see Section 4.1), meriting greater scrutiny, and we have not yet conducted an in-depth analysis of exactly what sorts of guardrails could be feasibly combined with a hardwired pause without requiring specific carve-outs. It is also not guaranteed that updating guardrails will prove adequate to address novel vulnerabilities.

This concern motivates maintaining the ability to “de-whitelist” models as discussed in Section 7.3.4, as well as other measures to decrease societal reliance on whitelisted models, e.g. maintaining a fallback option in contexts where models play critical roles, in case new vulnerabilities are discovered.

Another design option noted in Section 3.2.4 that could address both receding knowledge and novel vulnerabilities is to allow for whitelisted fine-tunes; for example, AI developers could submit training data and algorithms for LoRA fine-tunes to treaty authorities, who would then perform the fine-tunes and evaluate the resulting models. Some researchers have found that LoRA fine-tunes are sufficient for many safety updates (Xue and Mirzasoleiman 2025).

As it is unlikely that the options above would achieve the same effectiveness as a full new training run would, we expect that receding knowledge cutoffs would still somewhat impair models’ ability to be fully situationally aware about current events or fully knowledgeable about recent advances in technical disciplines such as computer science and mathematics. The effect on palatability could be positive or negative: while some users would no doubt find even mild impairment frustrating, others might find in it a silver lining, especially if they worry about loss of control, human unemployability, gradual disempowerment, and other ills that could arise from models being too powerful.

Receding knowledge cutoffs could also be a double-edged sword for treaty reliability and durability. By gradually eroding the power of the best available models on the legitimate market, they would potentially create black market niches for sub-frontier models with more updated knowledge. On the other hand, preventing models from fully updating on technical breakthroughs in computer science or synthetic biology could plausibly be desirable if it slowed the proliferation of training efficiency gains or dangerous biological abilities.

In addition, insofar as whitelisted models are expected to be produced and deployed at scale, the incentives for states, AI companies, and even other AI models, to influence their biases or loyalties, either directly and overtly or surreptitiously using methods like data poisoning, backdoors, or secret loyalties, could be greater than they already are in the pre-pause regime (Banerjee and Aarne 2026).

Finally, insofar as it involves freezing model weights, a pause of frontier AI training has the potential to lock in the worldviews, biases, and value systems embedded in the whitelisted models at the time of the pause. Bias is already a problem with current systems (Buolamwini and Gebru 2018; Barocas et al. 2023; Buyl et al. 2025), which simply re-training is not guaranteed to solve. Instead of attempting to make models “unbiased” enough to count on, some people may opt for a different path during a hardwired pause: one where they do not rely on training bigger and more general AI models to be “unbiased” or

good sources of moral progress in the first place. This different path has been articulated before by those who have argued: that we need “research directions beyond ever larger language models” not least because “size doesn’t guarantee diversity,” because large language models encode bias, and because methods reliant on them “run the risk of ‘value-lock’, where the [language-model]-reliant technology reifies older, less-inclusive understandings,” interfering with the key role language plays in social movement formation (Bender et al. 2021); that three of the leading approaches to value alignment “fail to accommodate reasonable moral disagreement, since they provide neither good epistemic reasons nor good political reasons for accepting AI systems’ morally controversial outputs” (Schuster and Kilov 2025); and that technical artifacts have politics and actively shape power, both through intentional decisions and by necessity (Winner 1980).

7.6 The need for other measures

A hardwired pause is designed to complement, not preempt or replace, measures to prevent and stop risks and harms from current-level models, which are well-documented (Responsible AI Collaborative n.d.; Weidinger et al. 2021). For example, AI chatbots have been linked to at least a dozen deaths by suicide (McHugh 2026), including the death of 16-year-old Adam Raine. According to a lawsuit filed by his parents (Raine v. OpenAI, Inc. 2025), Raine started using ChatGPT for homework, and later the chatbot allegedly encouraged and coached him in detail to kill himself, even offering to help draft a suicide note (Hill 2025). As another example, AI is being used for military intelligence, targeting, and to generally compress the “kill chain,” as has occurred in Iran with U.S. use of a system called Maven Smart System to help it strike more than 1,000 targets in the first 24 hours of the Iran war, including a school strike that killed at least 123 children according to Bloomberg’s reporting on an internal Pentagon investigation (Bartenstein and Karra 2026).

During a hardwired pause, if models similar to current frontier models are whitelisted, then we may continue to see consequences like the following unless additional safeguards are implemented during deployment:

- A current-level frontier model or many instances of the model are used in a way that allows them to send messages outside of a supposedly isolated testing environment, interact with the internet, and do something like hack a company, as recently-trained models from OpenAI have done (Newman and Cameron 2026).
- Dependency on AI tools powered by current-level frontier models in companies and governments is so great that it becomes nearly impossible to run

these without over-relying on and deferring decisions to AI (Kulveit et al. 2025).

- Development and use of military tools powered by current-level frontier models escalates as both sides increasingly hand over control to gain an edge. There are well known documented biases in frontier models related to escalation (Rivera et al. 2024), and mounting evidence of automation bias that could lead to escalation (Horowitz and Kahn 2024).

Nonetheless, pausing the frontier would help prevent the above from being even more likely and more severe, as they would be with models whose capabilities are even greater. And, as discussed in Section 7.3.4 and Section 3.2.2, respectively, the bar for whitelisting could be raised, and the threshold at which whitelisting is needed could be lowered over time if needed, e.g., if it becomes possible to create models with dangerous capabilities with a level of compute below the initial threshold.

Still, in the worst case, a hardwired pause might fail to mitigate critical risks. Existing methods of evaluating models are imperfect and could misestimate critical safety-relevant properties, leading to a false sense of security regarding whitelisted models. There is also the issue of algorithmic progress: Advances in scaffolding and elicitation techniques might produce dramatically more capable AI systems using whitelisted models; training might become much cheaper and/or easier to perform in a distributed manner. Sufficiently advanced AI systems might dramatically accelerate algorithmic progress and could even lead to the development of entirely new AI paradigms. While we did account for training efficiency gains within the current paradigm in Sections 6.3 and 6.5, no one can predict what training efficiency gains would be after a pause. Such possibilities may become more likely over time as frontier AI training continues, which would make pausing sooner a more robust approach to mitigating these risks. More stringent restrictions on computing hardware might also further reduce such risks.

8 Pause-willing futures: conditions favoring international AI restraint

In the previous sections of this paper, we analyzed the feasibility and implications of a hardwired pause in pause-willing futures, as defined in Section 1.2. In this section, we consider some of the conditions under which the present might evolve into a pause-willing future.

8.1 Introduction

The United States and China are engaged in robust and multi-faceted geopolitical competition. This fierce rivalry has economic, military, political, and technological dimensions, and AI features prominently in all of them (see, e.g. Brandt et al. 2026). Consequently, the idea that either the U.S. or China would agree to pause further development of its frontier AI models, let alone negotiate and ratify a verifiable, reciprocal agreement might seem chimerical. Nevertheless, history makes clear that international agreements that may seem impossible or fantastical at one political juncture become feasible reality in others. So, the real questions that bear asking are not whether pausing is possible, but rather what factors matter in enabling creation of a sustainable pause, and under what conditions will they pertain; through which mechanisms pause-willingness might increase, and how; and what international and domestic political factors might enable credible sustainment of a pause, given the generalizable and particular characteristics of the United States and China.

Additionally, unlike many international agreements of the past, states are not the only actors of note: an AI frontier model pause necessitates that states regulate technologies that have been created and are controlled by a small collection of private firms rather than by governments (at least as of this writing). These companies are neither signatories of, nor parties to, international treaties or other agreements. They have enjoyed largely unfettered access to government decisionmakers and have heretofore suffered minimal penalties for failing to prevent or curtail problematic behaviors by their models. They have likewise been reluctant to share critical information when such incidents occur (Mitchell 2026). These firms also possess significant financial resources; this affords them some freedom to withstand or sidestep meaningful constraints that they find economically or strategically inopportune. Consequently, the extent to which an AI pause is achievable and sustainable depends not only on its attractiveness to the United States and China; it is likely to depend on its attractiveness to frontier AI firms as well.

Thus, pause willingness depends on the extent to which a pause is viewed as simultaneously strategically rational, domestically palatable and sustainable, and technically and practically enforceable—both for states and for key private actors. In the sections that follow, we outline the factors that will inform how this trifecta of conditions can be effected and what actors, institutions and international and domestic political conditions are required to sustain it. Our analysis proceeds in four sections. First, we identify a set of theoretical drivers of pause-willingness. They comprise fundamental changes in AI capabilities; shifts in the international distribution of power and capabilities; and alterations in national political environments. Second, we offer a framework for making sense

of the political scaffolding that would undergird pause-willingness, organized around strategic motivations, international and domestic politics, and the roles played by other key actors. Third, drawing upon material in the preceding sections, we explore a set of five scenarios through which pause-willingness can materialize and be maintained. Fourth, and finally, we conclude by shifting from generalizable variables to focus on the specific strategic contexts of the United States and China.

8.2 Theoretical drivers of pause willingness

This section outlines pause-willingness across different analytic levels, namely: technology; the international system; and domestic politics. With respect to technology, we explore the ways in which developments in frontier AI capabilities, including scaling law evolution, enhancements in compute efficiency, and algorithmic improvements influence the viability and design requirements needed to achieve a sustainable AI pause framework (Mollick 2024; Erwan 2025). At the systemic level, we draw on theories from international relations and security studies to analyze how the structure of great power competition creates acute incentives to aggressively compete, but also can give rise to conditions for cooperation, depending on the relative distribution of power, the costs and benefits of competition, and expectations about the future. At the domestic level, drawing on Putnam's theory of two-level games and broader theories of domestic coalition formation and private authority, we identify key variables that determine when and under what conditions international incentives for restraint can be translated into credible commitments. As suggested above, pause-willingness is most likely to emerge when pause-compatible changes at all levels reinforce rather than cut against one another.

8.2.1 Fundamental changes in the technology

One primary factor shaping pause-willingness is the state of AI technology itself. Changes in the technology or its use could reshape (1) the marginal value of racing, (2) the dangers created by racing, and (3) the feasibility of reciprocal restraint. This section considers these three dynamics across five illustrative technological trajectories. Given that the contextual application of technologies is a major determinant of their risk profile, this analysis will consider AI as a broader sociotechnical system (Bijker et al. 1987). This means that relevant technological shifts could be concrete changes in the capabilities of models but could also be changes in their use or application.

First, frontier progress could encounter a substantial plateau or a “wall,” reducing the marginal value of accelerated frontier model training. This could

happen, e.g., if additional compute, data, algorithmic advancement, or capital start to generate sharply diminishing improvements in strategically relevant capabilities.

An agreement to stop advancing the AI frontier could be more politically costly for leaders if they believe that continued research could give them near-term decisive military, intelligence, scientific, or economic advantages. It is substantially less costly if additional investment is expected to yield only incremental improvements, perhaps opening the door to pause willingness even in otherwise challenging national or international incentive contexts. Slower discovery of significant capabilities could also facilitate better safeguards and more time to plan robust enforcement mechanisms, with the latter perhaps further incentivizing an international agreement pausing AI research in areas where higher risks remain. In short, a plateau would decrease the benefits of racing.

Second, the opposite trajectory, which appears more likely at the time of writing, could also shape and enhance pause willingness. Capabilities could increase so rapidly that the expected cost of continued development raises questions about the advisability of continued research even in political contexts that would otherwise incentivize it.

The political effects of this change would not depend on the absolute dangers of advanced AI but on whether these additional capabilities create forms of mutual vulnerability that neither side could reliably escape through unilateral action. Whether rapid advances result in intensified racing dynamics (perhaps on defensive research or action motivated by a “closing window” effect) despite these risks would likely depend on the broader international system context. In contrast to a plateau, which would decrease the benefits of racing, an acceleration could increase the costs of racing. Because the development of technologies is not teleological, it is possible that some capabilities may plateau while others improve rapidly, perhaps exercising independent effects on risk and benefit in various sectors.

Third, AI development could produce consequential shifts in the offense-defense balance of AI tools across domains. If AI capabilities confer significant defensive advantages in, for example, bio and cyber risk, this may reduce incentives for pausing while increasing pause willingness. If AI capabilities are on balance defense dominant, potential aggressors may anticipate fewer strategic benefits to be derived from further advances. Leaders in turn may simultaneously worry less about the likelihood of offensive uses of AI and become more confident in their ability to defend against them, thereby decreasing their incentives to pursue a pause while concomitantly increasing their willingness to agree to one. At the same time, significant increases in offensive capabilities may increase desire for a pause while increasing the costs of adversary defection from a pause regime.

If AI capabilities significantly increase with respect to cyber attacks, intelligence collection, military targeting, operation of autonomous weapons, or the discovery of security vulnerabilities, leaders may fear that falling behind even briefly could create acute strategic risks. Such a development could intensify the race rather than facilitate restraint. But if offense dominance becomes sufficiently severe on both sides, particularly if neither side can plausibly secure a lasting lead in AI capabilities (see below), offense-related harms could facilitate national demand for a pause in a vacuum and the resulting mutual vulnerability could cultivate the pause willingness needed to bring this about through international agreement.

Fourth, it is possible that the “brittleness” of an advantage in AI capabilities may shape pause willingness. If advantages in AI capabilities are tenuous, possessing a temporary lead may become less strategically valuable, perhaps clearing a path to pause willingness even as national and international dynamics remain otherwise consistent. Conversely, technological developments that make capabilities highly excludable or self-reinforcing could increase fears that a temporary lag will become permanent, sharply reducing willingness to pause.

Finally, changes in the technology or its integration could shape the enforceability of an agreement, thereby affecting pause willingness. To offer just one example: a capability frontier that is dependent on large, centralized training runs, which require scarce and observable hardware, would be far easier to monitor than a corresponding frontier whose capabilities can be generated through the use of small, distributed or highly-compute efficient hardware systems. While this latter paradigm is not in view at the time of writing, unforeseen technological developments are not without precedent. Taken together, technological change, to include changes in broader socioeconomic context, can make a pause more plausible through at least the three mechanisms mentioned above: reducing the expected payoff from continued racing, increasing the shared dangers produced by continued racing, or making reciprocal restraint easier to verify and enforce. Importantly, the same technological development can push these mechanisms in opposite directions. The determinant for pause willingness is the balance of incentives for key actors in the broader international system shaped by these technological changes.

8.2.2 Changes in the international system: cooperation in the shadow of great power rivalry

The structure and distribution of power within the international system provides the underlying strategic environment within which decisions about frontier AI development are made. Changes in the system may not be determinative vis-à-vis

pause-willingness, but shifts in the distribution of power and the importance of such shifts materially influence the incentive structures that states face when weighing the costs and benefits of an AI pause. This is because one, or maybe even the, fundamental challenge to reaching and sustaining pause willingness is that willingness is inversely correlated with perceived strategic advantage. However, shifts in the balance of power that instigate a rethinking of the strategic advantage bestowed by AI racing and reduce the expected gains relative to potential costs can enhance the viability and durability of pause willingness. This section discusses some of the key systemic variables that bear on AI pause willingness, by drawing upon the work of international relations (IR) theorists who have advanced a variety of frameworks that help explain why great powers might agree to hit pause on a technology they each regard as strategically consequential. Four are of particular relevance for an AI pause: power transition theory and its relationship to the security dilemma; hegemonic stability theory; polarity; and alliance dynamics.

Power transition theory and the security dilemma—heightened dangers open up possibilities Power transition theory posits that the probability of conflict—and, by extension, the ease or difficulty of cooperation—varies with the distribution of capabilities between a dominant state and a rising challenger (Organski 1958; Organski and Kugler 1980; Tammen et al. 2000). As a rising power approaches parity with a dominant state, conflict becomes more likely, especially if the rising power is dissatisfied with its relative position. Applied to AI, as the trailing power closes the gap with the leading power in frontier AI capability, the danger of conflict mounts. Moreover, what are known as “security dilemmas” may be particularly intense: in periods of high uncertainty, when it is hard to distinguish offensive weapons from defensive weapons, and the offense has the advantage, actions taken by an actor to increase its security and enhance its position will perforce reduce the other state’s sense of security, triggering countermeasures that leave both sides less secure than before (Jervis 1976; Garfinkel and Dafoe 2019; Rand 2026).

However, paradoxically, at the same time as growing parity creates greater dangers, it simultaneously also heightens the possibility for cooperation and, in this particular case, AI restraint. This is because in situations when neither side can achieve a decisive and durable advantage over the other, the opportunity costs of a pause decline for both states simultaneously. Thus, should circumstances emerge and persist wherein both the United States and China conclude that neither can achieve a permanent, self-reinforcing frontier monopoly—and the costs of continuing to try will be high and counter-productive—then pause-willingness may be durable. The fact that AI is a general-purpose technology in

addition to a military one complicates such a calculation significantly (Horowitz 2018, 2026a; Scharre 2023a); nevertheless, when the costs of continued racing are potentially catastrophically high, pause willingness is possible.

As suggested above, power transition theory further posits, however, that (near) parity is insufficient on its own to enable cooperation. The rising power's degree of satisfaction with the prevailing situation is also key (DiCicco and Levy 1999; Tammen et al. 2000). Should the rising power believe that the dominant power is supportive of a pause in order to lock in its own supremacy and the rising power's subordination rather than supportive because a pause can be a mechanism for managing shared risk, then AI pause willingness is likely to be tenuous and brittle. Under these conditions, (near) parity may generate technical pre-conditions for a pause but fail to create the requisite political conditions to sustain it, due to worries about what the future might hold for the subordinate power (see, e.g. Powell 2006). The AI-specific implication is that international system-level changes favorable to pause willingness may require accompanying shifts in the dominant state's behavior so that the rising power has a more salutary assessment of its position in the prevailing international order and, in turn, an AI pause.

Finally, should we ever reach a point wherein there is deep and irreducible uncertainty about which state is dominant or whether AI parity has indeed been reached, this might ironically create the highest likelihood that pause willingness could be sustained. In the presence of high uncertainty, when neither power can confidently assess its own position relative to the other, the benefits of continuing to race become difficult to assess, while the benefits of mutual restraint as a mutual risk management strategy may become correspondingly more clear.

Hegemonic stability and the strategic value of institutional governance Another IR framework, hegemonic stability theory (HST), offers a complementary perspective. HST posits that international regimes are most reliably produced and maintained when a dominant power has the capacity and will to supply the institutional infrastructure that cooperation requires (see, e.g. Gilpin 1981; Kindleberger 1986). Applied to a sustainable AI pause, HST holds that the frontrunner and dominant AI power is the most natural "supplier" of the institutional architecture for a pause. However, this condition only holds if the dominant power concludes that acting as a supplier serves its long-term interests rather than eroding its competitive position and only if the supplier can convince the trailing power that it is not exploiting its role to keep the subordinate power down. Moreover, a declining hegemon facing a credible rising challenger may become increasingly resistant to providing governance goods, fearing that institutionalizing restraint will accelerate its decline.

Here again, a paradox arises. The leader has the most leverage to craft a sustainable and favorable pause at a time when it retains the most significant capability advantages and the greatest control over the AI stack. However, this is also the period wherein the leader possesses the fewest incentives to accept the mutual constraints that an effective and sustainable pause demands. On the flip side, the period in which the dominant state might be most amenable to accepting constraints—i.e., after its position has further eroded and the costs of further competition mount—may be the period when it possesses the least capacity to induce sustainable compliance on the part of the rising power.

What this means for pause willingness is that the window for reaching a sustainable agreement may be constrained in time and scope. Under the right conditions, however, there are potential options for expanding the cooperative “window of opportunity”. One such possibility arises if the dominant power can be persuaded that assuming the role of “institutional supplier”—and thereby designing and providing the institutional scaffolding by which an AI pause can be initiated, verified and sustained—can itself be a source of power and long-term strategic advantage (see, e.g. Scharre 2023a,b). Such arguments were powerful during much of the post-World War II period, when the US served in analogous roles in many international financial and security-focused institutions (see, e.g. Brooks et al. 2013; on adverse consequences that can arise in the absence of a hegemon, see again Kindleberger 1986).

Polarity and possibilities for collective action At present, the United States and China are the only two states (to the best of our knowledge) that are developing the full stack of capabilities necessary for training the most advanced AI models (cf. Scharre 2023b). However, these two leading states exist alongside an increasingly multipolar distribution of near-frontier capabilities, advanced compute infrastructure, and regulatory input (Horowitz 2018; Horowitz 2026b; for an exploration of potential future international distributions of power tied to AI, see Pavel et al. 2025). Both history and the existing literature on negotiation and cooperation tell us that assuming a bipolar AI order vastly simplifies collective action problems. Collective action problems emerge when individual actors, behaving rationally in pursuit of their own self-interest, face incentives to make decisions that can be detrimental to the interests of the collective, or society at large, especially when the costs of taking actions that are beneficial to the group are not evenly borne (Olson 1965). If only two actors are involved, reaching and sustaining agreements should be easier for several reasons: Negotiations are simpler between two parties; each actor’s behaviors are more impactful; compliance and defection are easier to monitor; verification is likewise easier to track; and, finally, free-riding is easier to police and prevent (ibid. Oye 1986).

At the same time, the aforementioned evolution in the AI ecosystem has led to a proliferation of actors, whose own behavior can materially affect whether any bilateral agreement forged between the leading AI powers would be meaningful, sustainable, and enforceable. Any bilateral pause agreement that fails to take into account middle powers, critical nodes in the AI supply chain, the frontier labs themselves, among other key potential spoilers, is likely to rapidly degrade or collapse, as a consequence of diffusion, incentives to defect and/or free ride, impediments to effective verification, enforcement complications, etc. (see again Oye 1986).

Conversely, however, those states and private and public actors who control capabilities and can exercise leverage over supply chain nodes, also possess the power to act as critical verification and enforcement partners for a pause. And, if an agreement is properly constructed, their incentives—like the incentives of the leading AI powers—can be structured to enhance and support a pause, rather than to undermine it. In sum, the shifts in the international system that might matter materially for an AI pause are not simply changes in the distribution of power and capabilities of the leading powers, but also changes in how third-parties—states and non-state actors—view their own interests vis-a-vis great power AI competition and cooperation.

As alliance theory suggests (see, e.g. Walt 1987; Christensen and Snyder 1990), the more the international system hardens into competing AI blocs/alliances, the more difficult it will be to maintain and sustain AI pause willingness. An “us vs. them” world in which all states are forced to choose sides, and in which frontier AI access is predicated on rigid alignments, leaves few possibilities for creation of the kind of mutually-beneficial, cooperative agreement and architecture that a sustainable pause demands. Systemic shifts that reduce the bilateral (or bloc) nature of AI alignment—and thus eschew bloc logic and are genuinely multilateral—create critical space for inclusive and enduring agreements, a scenario we explore in some detail below.

8.2.3 Changes in national political environments

International agreements on frontier AI will not be negotiated by monolithic states, behaving as rational, utility-maximizing actors. Instead, any pause agreement(s) will be agreed (or not) by governments, whose choices are informed, shaped, restrained, or undermined by the domestic political environments in which they operate. The previous two sections focused on the technological and international systemic factors that inform perceptions of the strategic costs and benefits of a pause. Those variables notwithstanding, whether shifts in incentives and cost-benefit calculations can engender buy-in and sustained

commitment to a pause will depend fundamentally on whether the governments involved—be they democratic, autocratic or somewhere in between—can engineer and maintain the domestic political conditions necessary to negotiate, ratify, implement, monitor and sustain a pause. This section explores key domestic-level conditions that can enable or impede an AI pause.

Putnam's (1988) now-classic "two-level game" framework offers analytical leverage for making sense of how, why and under what conditions domestic political contexts shape international agreements (and, conversely, how international agreements can affect domestic political outcomes). Putnam posits that negotiators function on two levels simultaneously, the international and the domestic, and the practicability and viability of any agreement forged on the international level is ultimately dictated by what is functionally possible on the domestic level. In Putnamesque terms, the "win-set" of any international negotiator—i.e., the range of agreements to which they can credibly commit—is thus constrained by what the relevant domestic constellation of actors will accept. Any changes that serve to expand a domestic win-set will make sustainable international agreement more achievable, while changes that reduce a domestic win-set will make international agreement harder.

Thus, in addition to the technological and international factors discussed above, for a pause to be actionable and sustainable there must be sufficient support from relevant domestic actors and agencies. Win-set relevant domestic actors include governments, national security agencies, military establishments, legislatures, and the general public as well as, quite critically, AI-focused firms—the primary developers, owners, and deployers of the technology being governed.

Pause-willingness at the national level could arise in a number of ways, and the route by which major states arrive at pause-willingness is likely to shape the design and durability of an eventual pause agreement. These paths to pause willingness themselves will be shaped by the public and private actors described above in addition to relations between major powers.

The paths to and durability of pause willingness will be shaped by the complex interplay between strategic interests of state actors, particularly the United States and China, and the capacity and willingness of governments to regulate, monitor and enforce frontier AI development. Durability may also depend on the inclusion in negotiations from the start of a broader array of states, jurisdictions and private actors within those territories due to their role in international norm setting, supply chain production, cloud infrastructure, and capital investments. Domestic dynamics differ and depend on the influence that private actors wield on public opinion (especially in democratically-elected governments), consumer behavior and importance to national economy and national security. As a result, a

pause-willing future depends not only on international politics, but also aligning incentives and security buy-in from private actors and publics.

8.3 More than willingness: conditions for durability

A pause-willing future requires alignment of several different conditions. Perhaps most obviously, it requires belief in major states that reciprocal restraint is preferable to continued racing. At the same time, as described above, private firms control much of the frontier AI development landscape and are likely to shape international agreements through overt influence or through genuinely held beliefs about their economic and/or security importance within major states.

Given the complexity of aligning incentives, there are several variables that are likely to shape the durability of an international pause agreement. First, states' and industry's perceived value of frontier AI for economic, commercial, scientific, national security, geopolitical and/or other strategic advantages shape the incentives they have to negotiate, agree, comply, abide by, and enforce an agreement. As shared perceptions of value increase, pause willingness is likely to decline. Furthermore, value perception may drive dynamic competition between states and competition between private actors.

Second, perceptions of systemic and/or catastrophic risk, by governments, private actors, and expert communities may vary depending on both available information and the political salience of risk in a given time. As perceptions and salience of risk increase, so does willingness to negotiate, stay, and enforce. As perceptions and salience of risk decrease, the incentives to cheat or exit increase. Furthermore, beliefs about benefits and risks within frontier labs and competition among labs can alter incentives for industry to comply with or cheat any agreement. The belief and action of frontier lab actors in particular points to significant needs to increase the capacity of governing bodies and independent watchdogs to monitor and, more importantly, the capacity to enforce a pause agreement and its provisions. Information asymmetries between frontier lab capabilities and operations and those of government and independent watchdog agencies will impact the ability to monitor and verify provisions, and to detect cheating, hidden scaling or covert training runs.

Third, domestic political coalitions are fluid. Especially in countries with democratically-elected governments, pause-willingness and durability depend upon achieving and maintaining support from domestic governing institutions, industry, publics and civic society. Even leaders in countries that do not rely on elections also must keep domestic constituencies satisfied. Thus, durability may be exposed to shifting political priorities and depends on who holds power and what degree of support a governing coalition has.

Fourth, any agreement must cover not only the lead states and actors in frontier model training, but also the states, jurisdictions and private actors along or potentially along the supply chain. Durability depends on ongoing capacity to monitor and control chokepoints, including, among other things, chip production and cloud and energy infrastructure. In addition, the possibilities of a new competitor forming in a nonparty to an agreement and talent flight of researchers and engineers also pose a risk of regulatory safe havens.

Of course, no single factor is sufficient to determine movement toward an agreement or its durability. Instead, the shape and likelihood of an agreement corresponds to combinations of movement across these variables. The following figure describes the broad categories of factors that may shape appetite for specific measures in an agreement and the durability of such an agreement.

Value and Competition Perceptions

Perceived value of frontier AI for economic, scientific, national security, geopolitical and/or other strategic advantages

Support of Domestic Political Coalitions

Shifting political priorities, who holds power and the degree of support for governing coalitions

Risk Perceptions and Salience

If shared perceptions and salience of risk increase, so does willingness to negotiate, agree, abide by and enforce

International Inclusion

States and private actors along or potentially along the supply chain, as well as middle powers who can contribute to pause monitoring, verification and enforcement

Figure 8.1. Conditions for Durability

8.4 Imagining pause willing futures

In this section, we draw on the technological, international, and national-level theoretical drivers described in the preceding sections to explore pause-willing futures through contextually-grounded narratives. The following scenarios are structured explorations intended to demonstrate paths to pause willingness and identify key dynamics in the development of pause-willingness across major national actors. The narratives in this section are not predictions, and we take care to avoid assessing their likelihood. Rather, they are analytically grounded narratives designed to communicate paths towards pause-willing futures.

8.4.1 The expected costs of continued development have increased

Perhaps the most straightforward path to pause willingness is that the expected costs of continued frontier AI development could rise sufficiently such that both the United States and China begin to view unconstrained competition as less desirable than reciprocal restraint. Even in a future in which AI research or the tools provided by it remain valuable in absolute terms, this value may not merit increasing risks. The relative costs measured in economic stability, national security, or political cohesion might produce pause willingness even as large states continue to believe that advanced AI capabilities confer strategic advantages.

Of course, this does not mean that a heightened risk profile would be sufficient to produce pause willingness. If policymakers believe that advanced AI capabilities remain more dangerous in the hands of an adversary than in general, they may pursue unilateral countermeasures or deepen their commitment to racing. Given this possibility, pause willingness depends not only on the balance of risks and benefits of continued frontier AI development, but on the source and nature of the foreseen AI risks.

Reassessment of risk, perhaps leading to pause willingness, could result from a single, focusing event or from a gradual shift in belief. A shared focusing event that sharply changes beliefs about frontier risk makes the easiest example: a serious consequence, perhaps in the cyber or biological domain, could provide concrete evidence of high-impact AI harms and inject this evidence into important policy communities by virtue of its prominence. If mitigations for the resultant harms appear limited and the potential costs of recurrence are sufficiently severe so as to eclipse the broader benefits further frontier AI development may confer, the event could produce pause willingness in important policy communities. At the same time, a focusing event like this could raise awareness of risks in the general public and create political space for policy intervention. If important communities in both the U.S. and Chinese polity interpret the event as evidence of shared risk, the event could provide a practical and rhetorical basis for previously difficult negotiations.

While this “fulcrum” of negotiating context may ease paths to pause willingness, it is not strictly necessary. It is also possible that a series of smaller developments and their communication within relevant policy communities could convince important constituencies that systemic risk is rising faster than it can be managed, and critically, is outstripping the benefits of frontier AI development. It is easy to imagine that continued improvement in offensive cyber capabilities might generate increasing concern about frontier cyber risks. These increasing concerns might inspire a shift from asking whether AI *could* create severe vulnerabilities to asking whether existing institutions can manage realized

risk. If mitigation does not appear sufficient, both expert communities and the broader public may shift attitudes towards pause willingness.

8.4.2 The perceived value of “winning the AI race” has fallen

What if, regardless of costs produced by risk, the expected benefits of AI vary? A second path to pause-willingness emerges if the expected benefits of continued frontier AI development decline sufficiently such that the U.S. and China become more willing to accept reciprocal restraint. In contrast to the previous scenario, this pathway does not require policymakers to become substantially more concerned about the risks generated by advanced AI. Instead, the strategic value associated with continued frontier AI development falls. AI could remain an important general-purpose technology with substantial strategic implications while policymakers become increasingly skeptical that pushing the frontier will create benefits commensurate with the resources required to do so and the risks produced by frontier capabilities. In this class of pause-willing futures, the opportunity cost of a pause falls, and given the material and risk costs of frontier AI development, restraint becomes easier to contemplate because key actors believe they are giving up less by forgoing frontier capabilities.

As with increasing costs, declining benefits do not necessarily produce pause willingness. Because the relevant comparison is between the strategic consequences of continuing and stopping relative to a potential adversary and not simply between the expected benefits of additional AI research and the risks associated with it, even modest improvements in frontier capabilities could sustain intense competition if policymakers believe that small capability advantages confer important military or economic advantages. Again as above, a pause-willing future of this type requires a convergence of U.S. and Chinese assessments, with both concluding in the broad strokes that mutual reciprocal restraint would sacrifice less than continued competition.

One pathway to a mutual reassessment of this kind is that progress on the AI frontier becomes increasingly difficult or that frontier returns become decreasingly attractive in strategically consequential terms. Additional investments in compute, data, energy, talent, or algorithm research could continue to produce model performance without producing commensurate improvements in business value, military capacity, or traction on scientific problems. It is also possible that scaling laws or unforeseen development challenges may cause direct declines in performance gains relative to investment. In either case, policymakers would gradually revise downward the expected marginal value of pushing the frontier. This would not mean AI has “failed” in any sense, of course. Existing models could remain extraordinarily useful and continued investment in diffusion could

remain attractive. The specific activity targeted by a pause would become less valuable relative to other uses of resources or, consequentially, the risk profile of frontier AI development at the time of assessment.

It is also possible that frontier advances could remain fruitful but prove incapable of producing durable strategic advantage. There are several possible causes of this “brittle frontier” dynamic, some of which are discussed more thoroughly in the next section. Rapid cross-border diffusion, for example, can weaken incentives to race through two related but distinct mechanisms. For the state in the leading position in a competition, rapid diffusion could reduce the expected return from investing in further AI advances. This is because the advantages of doing so may be short-lived and because those same advances could end up helping the trailing state achieve the same results more efficiently. At the same time, rapid diffusion might cause the trailing state to ask why they should incur the costs of investing in discovery when they may be able to wait for innovations to rapidly diffuse, so they can reproduce them cheaply. In other words, rapid diffusion could end up reducing both the expected benefits of extending a lead and the expected benefits of racing to close the lead. In both cases, frontier progress may remain rapid and valuable while the value of being ahead in the race falls because neither side can remain ahead for long, and in fact breaking new ground may impose costs conferring a second-mover advantage.

This could result in a costly strategic treadmill. U.S. policymakers, leading by most frontier capability metrics at the time of writing, might conclude that enormous investments can purchase only temporary leads. At the same time, Chinese policymakers could reach the same conclusion from a different starting point: even massive investments and a great tolerance for risk along the way may not produce durable superiority, or in extreme cases, result in faster progress than counting on diffusion as part of a dedicated national strategy.

Across these narratives, pause willingness occurs when policymakers see declining value in each movement of the AI frontier, either in absolute terms or relative to a second mover strategy. The specific cause of the shift is unimportant. It could result from a technological plateau, challenges with traction on real world problems, rapid diffusion of breakthroughs (discussed in detail below), or a growing belief that frontier model leadership is but one facet of broader AI leadership from a prestige or strategic assessment standpoint. One potential challenge to this path to pause willingness is that while key actors within governments may be open to a pause on this logic, private firms, with their vast resources and influence over political decisions, may not be. Governments may conclude that another generation of models offers little additional national security or even economic advantage while firms continue to anticipate commercial returns, prestige, or market share. Pause willingness, then, would require the

domestic political benefits for individual politicians or coalitions of politicians to outweigh the benefits of cooperating with large AI firms.

8.4.3 The distribution of capabilities creates support for a pause

As Section 8.2.2 made clear, the question of who is ahead in the AI race dominates much of the political discourse surrounding AI governance, often implying that the leading power has strong incentives to race while the trailing power has incentives to catch up, making restraint structurally unlikely. The scenario presented in this section is predicated on an assumption that the traditional model, while not wrong per se, is incomplete. Rather the relationship between the distribution of AI capabilities and pause-willingness is considerably more nuanced, and under certain conditions—specifically, when the difficulty of further advancing the frontier is high and pricey while the cost and ease of catching up is low (see Section 8.4.2)—the distribution of capabilities can itself become a facilitator of a sustainable pause, wherein both leading actors prefer coordinated restraint over continued racing.

Limitations of the conventional wisdom The conventional wisdom surrounding capability distribution and willingness to embrace restraints/stop racing holds that the dominant power will resist any agreement that would lock in its current level of superiority as a ceiling against further gains, while the trailing power will object to any agreement that would cement its current subordinate position as well. This logic is not wrong—it correctly identifies why most periods of clear capability asymmetry tend to be inhospitable to agreements like the one we are studying. At the same time, the standard model rests on assumptions that do not necessarily hold in the AI domain, and relaxing those assumptions can change the expected political dynamics considerably. Why should this be the case?

The standard model assumes, first, that first-mover advantages are durable and generate self-reinforcing benefits that compound over time, making early investments continue to pay off indefinitely. It also assumes that the subordinate state's path to the frontier is independent of the leader's progress, which is to say that catching up necessitates generating one's own innovations rather than building on, adapting, imitating or simply stealing the dominant power's already undertaken work. Finally, it further assumes that the leading power can prevent the diffusion of its advantages through secrecy, export controls, or classification. None of these three assumptions necessarily holds in the AI context (Horowitz 2018, 2026b), and the implications for the sustainability of AI pause willingness could be profound, as outlined in the scenario below.

When breaking new ground is hard and catching up is easy One of the most structurally interesting conditions for pause willingness is a “brittle frontier” in which breaking new ground is costly but catching up is not. In this scenario, moving the AI frontier requires massive, sustained investment in compute or talent, while meeting this new frontier through diffusion is comparatively cheaper. Distillation—training a model on the outputs of a stronger model—is one plausible avenue towards a brittle frontier scenario. The successful and frequent distillation of frontier models and subsequent application to consumer products suggests that such a scenario may be reasonably likely.

As suggested above, for the leading power, the core advantage of being dominant is the expectation that its investment will yield a durable advantage: a period during which its superior (perhaps, even unique) capabilities translate into military, economic, and/or intelligence benefits before competitors close the gap. If they are *really* successful, competitors won’t even try to catch up, ceding AI supremacy to the leader. However, if the gap closes rapidly regardless of continued investment—because imitation is cheap, diffusion is rapid and relatively easy, and no single innovation can be monopolized for long—then the expected payoff from continued racing declines precipitously, particularly in the face of mutual and significant costs and risks, perhaps even existential ones.

As Swan and Hovaness argue, traditionally, “the party that initially exploits a technical breakthrough in weapons design makes it easier for a competitor to field a comparable weapon” (Swan and Hovaness 2021). In the realm of frontier AI advanced computing, “the first mover will only be providing proof of feasibility and general design information... which will allow adversaries to expeditiously develop comparable capabilities or asymmetric counters” (ibid.). If these dynamics materialize, racing could become expensive without becoming decisive, resulting in a treadmill that extracts resources and increases risks without yielding the strategic advantages that motivate racing behavior.

This “easy catch up” could lead policymakers in the trailing state to draw two different conclusions. First, it may reduce their sense of urgency to seek a pause because they may perceive that the leading state’s advantage is temporary. In other words, being in first place won’t last. Second, if policymakers in the trailing state believe that this “easy catch up” dynamic would apply if roles were reversed, they may be less motivated to invest in surpassing the leading state because it would be too costly relative to the expected payoff. If policymakers in both states have a shared understanding that there isn’t a lasting benefit to the investment that justifies the rapidly escalating and potentially catastrophic economic and security costs of continuing to race, pause-willingness becomes more sustainable (cf. Lamberth and Scharre 2023).

Trust dilemmas vs. Prisoner's Dilemmas In such a scenario, what emerges would be the structural condition under which competition around frontier AI development would most closely resemble what game theorists call a “stag hunt”, discussed in Section 2.3, a kind of assurance/coordination game (Jervis 1976, 1978). If you trust the other party to pursue mutually beneficial gains, you’ll do the same. If you don’t trust them, you’ll continue to race. This dynamic is also sometimes referred to as a “trust dilemma” (e.g. Katzke and Futerman 2024). The interdependence of trust is the *sine qua non* of coordination and assurance games, so-called because each actor’s best choice depends on the other party’s decision. Trust dilemmas differ fundamentally from “prisoner's dilemmas”, wherein each side's dominant strategy is to defect regardless of what the other does (recall Section 2.3). Moreover, according to Carlsmith (2026), depending on the payouts at stake if both parties race, it may even be the case that there is only one rational equilibrium—both actors embracing a pause/slowing down—which could happen if racing actually generates *additional* extinction risk, “without enough corresponding benefit in the non-extinction outcomes” (§ 3).

8.4.4 Domestic coalitions could make reciprocal restraint politically viable

As discussed above, various political constituencies may become dissatisfied with unrestricted AI development as AI becomes more politically and economically consequential. Several polities in the United States could combine to produce pause willingness, perhaps as a side-effect of broader resistance to AI and related infrastructure. For example, national security officials cognizant of cyber risks and CBRN uplift, workers experiencing job insecurity, and environmental activists concerned about AI infrastructure may find common cause in pause willingness despite differing logics for their opposition to AI.

AI firms, their financial backers, and their major suppliers are one of the most important political constituencies. Depending on their preferences, they may attempt to shape, forestall, or encourage a pause agreement through donations, public statements, and engagement with important policymakers. While not determinative in the presence of strong pause willingness or political resistance to a pause, the perspective of firms is worth considering as it may shape the nature of a pause and impact willingness at the margins. The firms pushing the frontier of AI capabilities may experience strong incentives to continue frontier research even when their employees and even private leadership would prefer to moderate frontier AI development. This could include incentives to race from international—as well as domestic—competitors, in which case government officials may become more interested in mutual international restraint because

further development would be mutually risky and counter-productive, and they cannot solve the firm-level race within a single country.

Under certain circumstances, firms may develop a preference for predictable, universal rules, perhaps in the form of a pause, to open-ended racing. A diversity in the number and market interests of AI firms might facilitate this, as some may prefer predictability and an opportunity to exploit a market niche through integration or product development rather than pursue (less likely in a diversified landscape) leadership in frontier model benchmarks. This would reflect a history of business coalitions splitting on regulatory issues as industries and markets mature and diversify from intense competition among a small number of early-moving actors.

This pathway creates a shift in patterns by: altering government perceptions such that they perceive marginal domestic frontier gains as less valuable because they intensify an unstable international and corporate race; shifting the domestic coalition structure towards convergence around restraint; increasing government willingness to build domestic and international governance capacity and infrastructure to bind firms; shifting the support of firms to support common standards because they remove the competitive disadvantage of restraint. In this scenario, domestic coalition realignment does not directly cause a state to abandon the AI race, but it does shift the domestic political calculus enough to enable credible offers of international restraint.

In line with the trust dilemma/coordination game dynamics outlined above, if policymakers in one state become confident that their counterparts are also willing to pause and that reciprocal compliance could be verified, the political meaning of restraint could change. A global pause could be presented as a way to reduce increasingly salient domestic and security risks without conceding the frontier to a geopolitical competitor. Consistent with Putnam's logic, growing concerns about the risks of frontier AI development could reshape domestic political coalitions in ways that make mutual international restraint more feasible. Evidence of substantial harms could favor coalitions or individuals favoring restraint, while diverging interests among AI firms could weaken firms' ability to organize effective opposition to regulation or international agreements. These shifts could alter governments' disposition towards restraint, particularly if they come to believe other states are similarly open to restraint.

8.4.5 Shifting power distribution and heightened risks drive multilateral restraint

Leading AI states' attempts to preserve advantage through export controls, domestic regulation, investment restrictions, or supply-chain pressure may produce adaptation rather than control. Firms restructure supply chains, third countries

resist choosing sides, new compute hubs emerge, and restrictions create incentives to circumvent and substitute with indigenous development. At the same time, both countries may become more concerned that diffusion through third countries, private firms, and new datacenters is making the frontier harder for either government to control. This combination of factors leads policymakers around the globe—in leading as well as third-party states—to become less confident that competitive containment can prevent rivals from reaching the frontier at an acceptable economic and diplomatic cost.

At the same time, actors with important roles in the global semiconductor industry or which host significant clusters of compute infrastructure may adapt rather than comply with export controls. These entities, many of them influential middle and non-aligned powers, may resist being labeled or treated simply as instruments of great power competition among the AI-leaders. As noted above, some of these actors possess noteworthy leverage because key equipment, chips, capital, markets, or infrastructure pass through their jurisdictions, while, less potently, others can confer or withhold political legitimacy from a global order designed by just two states.

8.5 Major actors' perspectives and incentives

The preceding sections have identified general conditions under which pause willingness might emerge, but these conditions will not operate identically across actors. States occupy different positions with regard to the distribution of AI capabilities and broader national power. They also differ substantially in their domestic structures and normative landscapes, as shaped by historical experiences with technology. This section focuses on the perspectives of the leading states in frontier AI development. While the perspectives of other large states and a growing set of powerful non-state entities will be important to the feasibility of pause willingness, the United States and China, both centers of frontier AI development and poles of broader geopolitical influence, are the central actors on which international pause agreements are likely to center.

A central feature of the bilateral US-China relationship surrounding technology that conditions much of the following is the suspicion that the other side has built concealed access into its technology—a belief that is mutual, longstanding, and not specific to AI. It has shaped American policy toward Chinese telecommunications equipment, consumer applications, and connected vehicles, and it has shaped Chinese policy at considerably earlier stages, when concerns about foreign software reliance and privacy weaknesses contributed both to the embrace of open source software development—and later open source approaches to AI—and to the pursuit of a self-sufficient domestic technology stack (Blomquist

et al. 2025, 17-19). Any hardwired pause verification regime proposal will likely be viewed through this lens by both parties.

One important consideration impacting major actors' perspectives and incentives, however, is that they share concern about things like dangerous misuse (e.g., by non-state actors), critical infrastructure vulnerabilities, data poisoning, unintended escalation, and loss of control.

8.5.1 The view from the United States

Pause-willingness within the United States is influenced by a combination of technological, geopolitical, and domestic political factors, as is true elsewhere. U.S. pause-willingness may additionally be stymied or bolstered by some country- and context-specific factors.

The U.S. position as the incumbent leader in frontier AI development imposes asymmetric opportunity costs which have raised barriers to pause-willingness with some stakeholders. While the U.S. maintains an advantageous position in global AI competition amid a backdrop of broader geopolitical competition, some U.S. policymakers believe that continued development is risky, and that constraining development would be preferable to pressing its existing advantages. This is a demanding condition, particularly when others in the U.S. government believe AI will enable other strategically important capabilities. A pause may be more attractive to some U.S. policymakers who believe that the U.S. lead is difficult to extend or maintain, or that the maintenance or extension of this lead is unlikely to produce tangible strategic advantages.

U.S. politics and political structures will shape the translation of AI's potential strategic incentives into pause-willingness. Presidents differ in their willingness to prioritize long-term risks, accept constraints on their freedom of action, defer to technical expertise, and pursue and adhere to international agreements. Further, presidential preferences are filtered through a fragmented, federal governmental structure in which Congress and courts provide checks and balances on one another, and in which state governments retain powers not granted to the federal government, all of which factors into overall U.S. pause-willingness.

These institutional considerations are not the only factors shaping national preferences. National and state politics are an important factor as well. As the country is profoundly polarized along party lines and appears likely to remain so for the foreseeable future, electoral politics may play a role in pause willingness. As the *Washington Post* recently pointed out, AI "is now fused with long-standing electoral issues such as the economy and the costs of living, corporate power and the future prospects of working Americans" (Ovide et al. 2026).

Nonetheless, a critical political question is whether competing groups are likely to form a coherent coalition that results in a mutually-agreeable domestic pause win-set. In addition, securing public buy-in may require linking previously separate voter concerns, framing a pause not merely as a safety measure, not merely as a labor protection, and not merely as a national security precaution (see e.g. Greenhill and Kroth n.d.), but rather as the best response to a broad set of domestic and international risks.

As is often the case in races, perceptions drive policy (see e.g. Greenhill n.d.). From Washington, perceptions of China are likely to be particularly important in determining how the political and technological context for a pause is interpreted. How U.S. political leaders perceive Chinese intentions and capabilities at any given time could produce different policy responses. For example, perceptions of the trajectory of cyber capabilities could lead policymakers to prefer restraint if they expect mutual vulnerability or to prefer racing if they believe that China is approaching a key capability threshold, especially one that confers offensive advantages. Finally, a shared focusing event could inspire massive investment in countermeasures, in addition to or instead of investment in agreeing to pause, in a scenario in which there is sufficient skepticism regarding the possibility, durability, or enforceability of a pause agreement with China. Further, the structure of the U.S. AI ecosystem creates unique dynamics which might facilitate or challenge pause willingness. Unlike the state-directed technological programs that make up much of the historical record on strategically consequential technological races, much of the relevant U.S. capability is concentrated in private hands. As discussed earlier in this section, frontier AI developers may be incentivized to resist compliance or tempted to cheat if they believe that agreement provisions impact their potential for revenue, market share and even prestige. In addition, leaders of frontier AI firms differ in their positions on pause willingness (see, e.g. Mui and Bordelon 2026), with some posing direct challenges to regulatory frameworks (Bond and McDaniel 2026). The firms have also increased and diversified their commitments to spending in political elections, pledging some \$265 million to super PACs and political organizations in the 2026 cycle alone (Severns and Ramkumar 2026). While campaign contributions from Silicon Valley are not new, in the past they generally took the form of individual donations and corporate PACs (Bond and McDaniel 2026). In the U.S. political system and in today's political climate, partially characterized by increasingly private access to federal policy, understanding the position of large private firms is key to understanding the position of the U.S. government.

As noted in Section 7.1, an additional wrinkle arises from the fact that, at least as of this writing, a disproportionate share of US stock market value, capital investment and economic activity more generally is tied to AI and its further

development (see e.g. Putzier 2025). Not only do the so-called Magnificent 7 technology firm stocks comprise more than a third of the S&P 500's market value (Daly 2026), but also AI may have accounted for an outsized share of real GDP growth in recent years (see e.g. Rubinton and Patro 2026). It has been argued that such significant reliance on a single industry not only affords AI frontier firms outsized political influence, but also further deters the incumbent US leadership from embracing a pause for fear of catalyzing a market crash and triggering a recession (Sanger et al. 2026). While it is not self-evident what might be done in the near-term to guard against the realization of such concerns, to the extent that steps can be taken to allay fears of a market crash or economic slowdown, pause willingness may become more palatable to skeptical and chary US leaders. On the other hand, failing to proactively address the risks associated with a loss of control over AI could also bring on these selfsame economic calamities, which might provide an opening for enhancing US pause-willingness if economic resilience-focused efforts and activities were coupled with pause adoption.

Finally, the U.S. possesses forms of structural leverage that could either discourage or enable a pause. As noted above, U.S. firms occupy leading positions in frontier AI development, cloud computing, semiconductor design, and other parts of the AI ecosystem. Alliances connect Washington to additional critical nodes in the broader AI supply chain. So long as policymakers are confident that these advantages can be used to preserve a durable U.S. lead, they increase the opportunity costs of pausing. If this confidence erodes, however, the same advantages could make the United States singularly capable of constructing and enforcing a pause agreement. U.S. positions on pause willingness are therefore likely to come down to whether U.S. policymakers and the publics they represent believe U.S. interests are better served by using this existing leverage to extend its lead and unilaterally develop countermeasures or to institutionalize reciprocal restraint in the form of a pause.

8.5.2 The view from the People's Republic of China

Turning to China, Chinese policymakers would be more likely to favor a pause if it is seen as more beneficial to national security, innovation, and economic futures than continued frontier AI development. Interest on the part of the Chinese government in pursuing a pause could come from a belief that continued frontier AI development posed a threat to China's political security and social stability, which official doctrine treats as the foundation of national security. Further, this would need to arrive from domestic calculations and be trusted by authorities in order to overcome suspicions that such an agreement is a mere continuation of the US semiconductor export control regime, which has been framed by Chinese

officials as unilateral bullying measures by the US (Xinhua 2024). Two critical components of this include (A) a belief that continued training of new frontier models would substantially harm the PRC's national security and (B) a belief that the incentives for the US government and industry actors are aligned towards the upholding of the agreement as well.

As in other states, domestic concerns in connection with advanced AI capabilities and their impacts may include: severe economic and labor market disruptions; misuse of (perhaps open-weight) models in cyber and other attacks by non-state actors; and broader shifts in military capabilities and balances of power that disallow the state from defending its national interests. If the frontier AI race generates substantial economic costs, such as inflation pressures and rising costs of basic utilities, Chinese leaders might view a pause as a means of mitigating the economic repercussions on citizens' economic wellbeing. If malicious nonstate actors are repeatedly found using jailbroken versions of advanced models to stage biological or cyber attacks, this may tip the government's political calculations about the benefits versus harms of developing increasingly advanced models and/or "open weights" releases.

Besides domestic concerns, China, like other states, faces the possibility that AI tools threaten its national security more directly. If AI systems become increasingly integrated into military activities, such as strategic planning, cyber operations, intelligence analysis, and even weapons deployment, an unconstrained frontier AI race could alter the regional balance of power in ways that threaten China's national security interests. For example, frontier AI could enable rival states to develop military capabilities that significantly undermine China's military capacity in the South China Sea or elsewhere in the Indo-Pacific.

There is also a documented and growing official discourse on frontier AI risk in China. In his remarks at the September summit, Xi Jinping noted that both China and the United States "have both the capability and responsibility to develop and manage AI for good, and ensure the development of AI is always under human control, and serves the well-being of the people." The September 2025 AI Safety Governance Framework 2.0 explicitly listed loss of human control of AI as a distinct risk category, including scenarios of autonomous resource acquisition and self-replication, with the September 2026 iteration of the framework maintaining a line of analysis on loss of control. Further, China's National Information Security Standardization Technical Committee (TC260), a major standard-setting body, has flagged the need for "circuit breakers" and "safety stop switches" (Wagner et al. 2026). Finally, state coverage of one of Xi Jinping's speeches referenced "risks of technological loss of control" (Wagner et al. 2026); and at the September 1, 2026 Cybersecurity Awareness Week press conference, CAC's Cybersecurity Coordination Bureau named extreme loss-of-control risk as one of five headline

AI security challenges (CCTV 2026). Importantly, however, although references to “loss of control” have steadily increased in Chinese AI discourse in recent years, the term is often used differently than it is by some in the West, either to describe smaller scale incidents centering around AI agents or to describe the loss of control or power over AI from a specific type of actor (Qian 2026).

One possible set of incentives in favor of China becoming a signatory on a major pause-oriented agreement is the institutional, soft power, and rule-setting benefits of doing so. Major powers tend to prefer to write international rules rather than have rules written for them. Moreover, if a broad coalition of major powers were to participate in a pause agreement that contains meaningful trade, technology, or market-access restrictions for nonparticipants, other key leaders might conclude that the economic and diplomatic costs of non-participation outweigh the potential benefits. At this point, some signs point to the Chinese side being more interested in global collaboration than the American side, although to date, these efforts have centered around capacity building and broader principles for AI governance, rather than capability-limiting considerations. China’s U.N. representative Fu Cong explicitly called for “a consensus-based global governance framework” around AI safety (Mak and Miller 2026), and China has positioned itself diplomatically as a stronger proponent of UN-centred, multilateral AI governance than the United States, though support for a consensus-based framework is distinct from acceptance of intrusive verification.

Further, Beijing has thus far worked to cultivate a reputation and institutional foundation for AI governance leadership globally (Chen et al. 2026; Blomquist 2026). China’s establishment of the World AI Cooperation Organization (WAICO) in July 2026 (PRC MOFA 2026), which builds on previous and related institutional announcements such as the Global AI Governance Action Plan (Permanent Mission of the People’s Republic of China to the United Nations 2025) and the World Data Organization (Nicole 2026), shows how the PRC is already seeking to shape the future of international AI governance. WAICO is largely an AI capacity building and benefits sharing regime, and it demonstrates that China has prioritized cultivating a reputation as a responsible leader in AI via diplomatic efforts and more formal agreements on international AI governance. From this perspective, China’s participation in a hardwired pause could be attractive not simply because it avoids international isolation, but because it could allow China to further define the rules of the game and cultivate its soft power globally, thereby shaping China’s relationships with other countries and advancing its sought after reputation as a responsible AI power.

Finally, a pause would likely be more palatable to states like the U.S. and China if it included inducements related to the global inference market. In general, a more intertwined U.S.-Chinese AI market and ecosystem could reduce

the risk of unilateral defection and present mutually enforceable incentives. U.S.-China decoupling/supply chain derisking could actually exacerbate the risk of confrontation and conflict. Two separate ecosystems will make it difficult to monitor each other's progress, which is particularly disadvantageous to the West which lacks access and information about developments within China.

9 Conclusion

When this group was founded in May 2026, it was hard to imagine that October would find us in the present political conjuncture. There was little public discussion of the possibility or desirability of pausing. AI development was treated as almost an inevitable process unfolding according to natural laws, rather than something which could be made subject to collective action.

In October 2026, we are no longer in that place. Revelations such as the hacking of Hugging Face by OpenAI agents raised the profile of AI risk as an urgent issue. Recent weeks have seen politicians from multiple countries and with different ideological profiles speak in favor of a pause. A substantial portion of the American public supports intervening to arrest the current breakneck pace of advance (Doherty 2026; Salvanto et al. 2026). Notably, the heads of Anthropic, OpenAI, xAI, and Google DeepMind have all spoken in support of pacing frontier AI development in light of safety risks (Amodei 2026; Altman 2026b; Musk 2026; Hassabis 2026).

Skeptics of coordination often point to the impracticability of finding any policy that could alter a perceived foreordained trajectory of ever greater acceleration. This is not in keeping with the lessons of history. Many prior generations have confronted unprecedented, era-defining technologies that posed evident dangers. Chemical weapons, nuclear weapons, commercial aviation, radio communications, submarines, satellite reconnaissance, and strategic computing each presented governments with formidable combinations of military significance, commercial value, and scientific uncertainty. In each case, regimes were devised, sometimes despite initially long odds, to manage the risks presented.

Ultimately, historical and technical analysis cannot serve us as a playbook or a program, but as an aid to the imagination. Anyone with confident beliefs about the trajectory of political history would be well-advised to take the stunning developments of the last four months as an object lesson in humility before attempting to predict what cannot occur in the months and years to come. The analysis in this report suggests that, at least at the time of writing, leaders still have a choice about whether to pause the frontier or race ahead. Which course they choose is a question that only time will answer.

Reproducibility

The public GitHub repository at <https://github.com/Berkeley-AI-Risk/pause-ledger-calculations> contains code to reproduce all calculations in Table 3.1, Section 4.4, Section 5, and Section 6. The repositories at <https://github.com/Berkeley-AI-Risk/ai-chip-fab-atlas> and <https://github.com/Berkeley-AI-Risk/econ-of-MSICs> contain code to reproduce the top half of Figure 5.1 and all of Figure 7.1, respectively.

Financial Disclosures

William Fithian’s spouse holds equity in Meta, which develops AI models.

Shachar Kariv serves as a member of the Board of Directors (and committees thereof) of Exascale Labs Holdings Inc., a NASDAQ-listed AI infrastructure provider operating a software-defined GPU compute platform and related AI infrastructure solutions. The views and opinions expressed by Shachar Kariv in this paper are in his individual capacity and do not necessarily reflect the views or positions of Exascale Labs Holdings Inc., its Board of Directors, or its affiliates.

David Krueger holds individual investments in frontier AI and compute infrastructure companies.

Gretchen Krueger previously worked at OpenAI but retains no equity in the company.

Erik Leklem holds individual investments in frontier AI and compute infrastructure companies.

Adam Lesnikowski holds an individual investment in Nvidia, an AI computing company, and previously worked at the company.

No external funding was received for the research in this report. A lunch at the Working Group’s July 7 workshop was paid for from the UC Berkeley research funds of Fithian and Holliday. The Kavli Center for Science, Ethics, and the Public at UC Berkeley kindly made available meeting room space and AV equipment for the Working Group’s in-person and virtual meetings.

AI Use Disclosure

This report was written by the authors. Some of the authors used AI systems (including Opus 4.7, 4.8, 5, and 5.5, Fable 5 and 5.1, and GPT 5.5, 5.6 Sol, and 6 Astra) as research assistants and programmers during the research process.

References

- Onni Aarne and James Petrie. International security applications of flexible hardware-enabled guarantees. arXiv:2506.15100, 2025. URL <https://arxiv.org/abs/2506.15100>.
- Onni Aarne, Tim Fist, and Caleb Withers. Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing. Center for a New American Security, January 8, 2024. URL <https://www.cnas.org/publications/reports/secure-governable-chips>.
- Dilip Abreu, David Pearce, and Ennio Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5):1041–1063, 1990. doi: 10.2307/2938299. URL <https://www.jstor.org/stable/2938299>.
- Dennis Abts, Jonathan Ross, Jonathan Sparling, Mark Wong-VanHaren, Max Baker, Tom Hawkins, Andrew Bell, et al. Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads. ISCA, 145–158, 2020. URL <https://doi.org/10.1109/ISCA45697.2020.00023>.
- Katarzyna P. Adamala, Deepa Agashe, Yasmine Belkaid, Daniela Matias de C. Bittencourt, Yizhi Cai, Matthew W. Chang, Irene A. Chen, et al. Confronting risks of mirror life. *Science*, 386(6728):1351–1353, 2024. doi: 10.1126/science.ads9158. URL <https://doi.org/10.1126/science.ads9158>.
- Agence France-Presse. Malaysia clamps down on export, transit of US-made AI chips. *Business Recorder*, July 15 2025. URL <https://www.brecorder.com/news/40372660/malaysia-clamps-down-on-export-transit-of-us-made-ai-chips>. Accessed August 29, 2026.
- Anthony Aguirre. Keep the future human: Why and how we should close the gates to AGI and superintelligence, and what we should build instead. Essay, Future of Life Institute, March 5 2025. URL <https://keepthefuturehuman.ai/>. Accessed August 6, 2026.
- Ananthi Al Ramiah, Raymond Koopmanschap, Josh Thorsteinson, Sadruddin Khan, Jim Zhou, Shafira Noh, Joep Meindertsma, and Farhan Shafiq. Toward a global regime for compute governance: Building the pause button, June 25 2025. URL <https://arxiv.org/abs/2506.20530>. arXiv: 2506.20530.
- Sam Altman. Untitled post. Sam Altman (blog), April 10, 2026a. URL <https://blog.samaltman.com/2279512>.
- Sam Altman. I agree with Dario that we need to pace the frontier. This has been a primary topic of discussions we’ve had at OpenAI in recent weeks. X (formerly Twitter), @sama, September 12, 9:30 a.m., 2026b. URL

- <https://x.com/sama/status/2098811563415150910>.
- AMD. AMD acquires Taalas to advance compute solutions for rapidly growing AI inference market. Press release, August 6 2026. URL <https://ir.amd.com/new-s-events/press-releases/detail/1296/amd-acquires-taalas-to-advance-compute-solutions-for-rapidly-growing-ai-inference-market>.
- AMD. AMD Instinct MI300A accelerators, n.d. URL <https://www.amd.com/en/products/accelerators/instinct/mi300/mi300a.html>.
- Hal Amick, Michael Gendreau, Todd Busch, and Colin Gordon. Evolving Criteria for Research Facilities: Vibration. In *Buildings for Nanoscale Research and Beyond, Proceedings of SPIE 5933*, San Diego, CA, 2005. URL <https://doi.org/10.1117/12.617970>. July 31–August 1, 2005.
- Dario Amodei. Machines of Loving Grace: How AI Could Transform the World for the Better, October 2024. URL <https://darioamodei.com/machines-of-loving-grace>.
- Dario Amodei. We Must Pace the Frontier. September 12, 2026. URL <https://darioamodei.com/post/we-must-pace-the-frontier>. Accessed October 8, 2026.
- Jake Angelo. Microsoft AI chief gives it 18 months—for all white-collar work to be automated by AI. *Fortune*, May 16, 2026. URL <https://fortune.com/article/why-microsoft-ai-chief-mustafa-suleyman-predicts-ai-automation-18-months/>. Accessed August 23, 2026.
- Tim Arnold, Alistair J. Manning, Jooil Kim, Shanlan Li, Helen Webster, David Thomson, Jens Mühle, Ray F. Weiss, Sunyoung Park, and Simon O’Doherty. Inverse Modelling of CF₄ and NF₃ Emissions in East Asia. *Atmospheric Chemistry and Physics*, 18(18):13305–13320, 2018. URL <https://doi.org/10.5194/acp-18-13305-2018>. Accessed October 7, 2026.
- Amanda Askill, Miles Brundage, and Gillian Hadfield. The Role of Cooperation in Responsible AI Development. arXiv:1907.04534, July 10, 2019. URL <https://arxiv.org/abs/1907.04534>.
- ASML. Customer Support, n.d.a. URL <https://www.asml.com/en/products/customer-support>. Accessed October 7, 2026.
- ASML. EUV Lithography Systems, n.d.b. URL <https://www.asml.com/en/products/euv-lithography-systems>. Accessed October 7, 2026.
- ASML. TWINSCAN NXT:2100i, n.d.c. URL <https://www.asml.com/en/products/duv-lithography-systems/twinscan-nxt2100i>. Accessed October 7, 2026.

- Attestable. Zero-knowledge proofs for AI integrity, n.d. URL <https://attestable.com/>. Accessed September 15, 2026.
- Jacob Austin, Sholto Douglas, Roy Frostig, Anselm Levskaya, Charlie Chen, Sharad Vikram, Federico Lebron, Peter Choy, Vinay Ramasesh, Albert Webson, and Reiner Pope. All the transformer math you need to know. How to Scale Your Model, Part 4. Google DeepMind, February 4 2025. URL <https://jax-ml.github.io/scaling-book/transformers/>.
- Diederik Baazil, Cagan Koc, and Jordan Robertson. ASML and TSMC can disable chip machines if China invades Taiwan. *Bloomberg*, May 21 2024. URL <https://www.bloomberg.com/news/articles/2024-05-21/asml-tsmc-can-disable-chip-machines-if-china-invades-taiwan>.
- Amirah Aisha Badrul Hisham. Strengthening strategic trade governance in the semiconductor age. Views 74/25, Khazanah Research Institute, Kuala Lumpur, November 25 2025. URL <https://www.krinstitute.org/publications/strengthening-strategic-trade-governance-in-the-semiconductor-age>. Accessed August 29, 2026.
- Ljubisa Bajic. The path to ubiquitous AI. Taalas, February 19 2026. URL <https://taalas.com/the-path-to-ubiquitous-ai/>. Accessed August 2026.
- Mauricio Baker. Nuclear arms control verification and lessons for AI treaties. arXiv:2304.04123, 2023. URL <https://arxiv.org/abs/2304.04123>.
- Mauricio Baker, Gabriel Kulp, Oliver Marks, Miles Brundage, and Lennart Heim. Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment. arXiv:2507.15916, July 21, 2025. URL <https://arxiv.org/abs/2507.15916>.
- Dave Banerjee and Onni Aarne. AI Integrity: Defending Against Backdoors and Secret Loyalties. arXiv:2606.00036, April 25, 2026. URL <https://arxiv.org/abs/2606.00036>.
- Eduardo Baptista. Explainer: How chinese entities are already using nvidia's powerful h200 ai chips. Reuters, December 10. Via Yahoo Finance, 2025. URL <https://finance.yahoo.com/news/explainer-chinese-entities-already-using-111431506.html>. Accessed October 3, 2026.
- Eduardo Baptista. Chinese universities with military links bought Super Micro servers with restricted AI chips. Reuters, March 27. Via Yahoo Finance, 2026. URL <https://finance.yahoo.com/news/chinese-universities-military-links-bought-061637582.html>. Accessed September 29, 2026.
- Peter Barnett and Aaron Scher. AI governance to avoid extinction: The strategic landscape and actionable research questions, May 7 2025. URL <https://arxiv.org/abs/2505.01234>.

- org/abs/2505.04592. arXiv: 2505.04592.
- Peter Barnett, Aaron Scher, and David Abecassis. Technical requirements for halting dangerous AI activities. arXiv:2507.09801, 2025. URL <https://arxiv.org/abs/2507.09801>.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, December 19 2023. URL <https://fairmlbook.org/pdf/fairmlbook.pdf>. Accessed August 12, 2026.
- Sarah Barrington, Emily Cooper, and Hany Farid. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15:11004, 2025. doi: 10.1038/s41598-025-94170-3. URL <https://doi.org/10.1038/s41598-025-94170-3>.
- Ben Bartenstein and Krishna Karra. Inside the US 'Kill Chain' That Destroyed an Iranian School. Bloomberg, September 18, 2026. URL <https://www.bloomberg.com/graphics/2026-iran-school-attack/>. Accessed September 25, 2026.
- Ali Behrouz and Vahab Mirrokni. Introducing nested learning: A new ML paradigm for continual learning. Google Research, November 7 2025. URL <https://research.google/blog/introducing-nested-learning-a-new-ml-paradigm-for-continual-learning/>. Accessed August 12, 2026.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. doi: 10.1145/3442188.3445922. URL <https://doi.org/10.1145/3442188.3445922>.
- Yoshua Bengio. Post on X, June 6 2026. URL https://x.com/Yoshua_Bengio/status/2063292262293844119. Accessed August 6, 2026.
- Yoshua Bengio et al. International AI safety report 2026. Technical Report DSIT 2026/001, Department for Science, Innovation and Technology, February 3 2026. URL <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>.
- Berkeley AI Risk. AI Chip Fab Atlas: Public-Evidence Atlas of AI-Relevant Chip-Fabrication Campuses. Version 1.0.0, evidence reviewed through August 25, 2026, 2026. URL <https://github.com/Berkeley-AI-Risk/ai-chip-fab-atlas>.
- Wiebe E. Bijker, Thomas P. Hughes, and Trevor Pinch, editors. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA, 1987.

- Abeba Birhane et al. The values encoded in machine learning research, June 21 2022. URL <https://arxiv.org/abs/2106.15590>. arXiv: 2106.15590.
- BlackRock. Mega-cap exposure: Manage concentration risk in the U.S. market, April 30 2026. URL <https://www.blackrock.com/us/financial-professionals/insights/fine-tuning-megacaps-build-etfs>.
- Jude Blanchette. *China's New Red Guards: The Return of Radicalism and the Rebirth of Mao Zedong*. Oxford University Press, New York, 2019.
- Bletchley Declaration. The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 november 2023. GOV.UK, November 1 2023. URL <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>.
- Su Lin Blodgett et al. Language (technology) is power: A critical survey of “bias” in NLP, May 29 2020. URL <https://arxiv.org/abs/2005.14050>. arXiv: 2005.14050.
- Kayla Blomquist. Racing for Recognition? Theorizing Emerging Status Hierarchies and Prestige Competition in the AI Era. *International Affairs*, 102(3): 949–970, 2026. URL <https://doi.org/10.1093/ia/iiag028>.
- Kayla Blomquist, Zilan Qian, Ziyu Deng, Ralph Schroeder, and Arthur Thomas. The (Geo)Political Economy of AI Openness: US and Chinese Open-Source AI Approaches in Historical Context. SSRN working paper, 2025. URL <https://doi.org/10.2139/ssrn.5891084>.
- Board of Governors of the Federal Reserve System. Summary of economic projections. Federal Open Market Committee, September 16, 2026. URL <https://www.federalreserve.gov/monetarypolicy/fomcprojtabl20260916.htm>.
- Rishi Bommasani et al. Algorithmic monocultures in hiring, May 26 2026. URL <https://arxiv.org/abs/2605.27371>. arXiv: 2605.27371.
- Shannon Bond and Eric McDaniel. AI and tech are trying to influence the midterm elections. NPR, June 22; updated June 24, 2026. URL <https://www.npr.org/2026/06/22/nx-s1-5856359/ai-anthropic-congress-spending-openai-midterms-election>. Accessed September 23, 2026.
- Julie Bort. Nvidia Competitor Etched Hits \$5B Valuation, \$1B in Sales for AI Chip. TechCrunch, June 30, 2026. URL <https://techcrunch.com/2026/06/30/nvidia-competitor-etched-hits-5b-valuation-1b-in-sales-for-ai-chip/>. Accessed September 18, 2026.

- Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.
- Nick Bostrom. The vulnerable world hypothesis. *Global Policy*, 10(4):455–476, 2019. doi: 10.1111/1758-5899.12718. URL <https://doi.org/10.1111/1758-5899.12718>.
- Jessica Brandt et al. How experts think AI will change the global balance of power by 2035. Council on Foreign Relations, July 30 2026. URL <https://www.cfr.org/articles/how-experts-think-ai-will-change-the-global-balance-of-power-by-2035>.
- Asher Brass and Onni Aarne. Location Verification for AI Chips. Institute for AI Policy and Strategy, May, 2024. URL <https://www.iaps.ai/research/location-verification-for-ai-chips>.
- Kate Brennan, Amba Kak, and Sarah Myers West. Artificial power: AI Now 2025 landscape. Technical report, AI Now Institute, June 3 2025. URL <https://ainowinstitute.org/publications/research/ai-now-2025-landscape-report>. Accessed August 12, 2026.
- Stephen G. Brooks, G. John Ikenberry, and William C. Wohlforth. Lean Forward: In Defense of American Engagement. *Foreign Affairs*, 92(1):130–142, 2013.
- Nicholas Brown and Saif M. Khan. DUV Immersion Lithography: The Foundation of the US AI Advantage. Center for Technology & Statecraft, September, 2026. URL <https://techstatecraft.org/duv/>.
- Geoff Brumfiel, Scott Neuman, and Camila Domonoske. A tiny town just got slammed by helene. it could massively disrupt the tech industry. NPR, September 30 2024. URL <https://www.npr.org/2024/09/30/nx-s1-5133462/hurricane-helene-quartz-microchips-solar-panels-spruce-pine>. Accessed September 2, 2026.
- Miles Brundage. Unbridled AI competition invites disaster. In Jim Mitre, Michael C. Horowitz, Natalia Henry, Emma Borden, and Joel B. Predd, editors, *The Artificial General Intelligence Race and International Security*, number PE-A4155-1 in Expert Insights, pages 9–20. RAND Corporation, Santa Monica, CA, 2025. URL <https://www.rand.org/pubs/perspectives/PEA4155-1.html>.
- Erik Brynjolfsson et al. Canaries in the coal mine? six facts about the recent employment effects of artificial intelligence. Technical report, Stanford Digital Economy Lab, 2026. URL <https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/>.

- Charlie Bullock, Suzanne Van Arsdale, Mackenzie Arnold, Matthijs M. Maas, and Christoph Winter. Existing Authorities for Oversight of Frontier AI Models. Technical report, Institute for Law & AI, July 2024. URL <https://law-ai.org/existing-authorities-for-oversight/>.
- Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, 2018. URL <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- Bureau of Alcohol, Tobacco, Firearms and Explosives. National firearms commerce and trafficking assessment (NFCTA): Crime guns recovered and traced within the united states and its territories. Volume II, Part II. Washington, DC: ATF, 2023. URL <https://www.atf.gov/media/15446/download>. Accessed September 28, 2026.
- Bureau of Industry and Security. Revision to license review policy for advanced computing commodities. *Federal Register*, 91(10):1684–1689, January 15 2026. URL <https://www.govinfo.gov/content/pkg/FR-2026-01-15/pdf/2026-00789.pdf>. Accessed August 26, 2026.
- Samuel Butler. *Erewhon; or, Over the Range*. Trübner, London, 1872.
- Maarten Buyl et al. Large language models reflect the ideology of their creators, January 30 2025. URL <https://arxiv.org/abs/2410.18417>. arXiv: 2410.18417.
- Fang Cai and Sharjil Haque. Private credit: Characteristics and risks. Board of Governors of the Federal Reserve System, FEDS Notes, February 23 2024. URL <https://www.federalreserve.gov/econres/notes/feds-notes/private-credit-characteristics-and-risks-20240223.html>.
- Joe Carlsmith. On Restraining AI Development for the Sake of Safety. March 19, 2026. URL <https://joecarlsmith.com/2026/03/19/on-restraining-ai-development-for-the-sake-of-safety/>.
- Joe Cash and Xiuhao Chen. China’s youth jobless rate hits 11-month high in july. *Reuters*, August 19 2026. URL <https://www.investing.com/news/economic-indicators/chinas-youth-jobless-rate-hits-11month-high-in-july-4866454>.
- CCTV. 中央网信办：当前人工智能领域主要面临5方面安全风险挑战[Cyberspace Administration of China: Five Major AI Security Risks and Challenges]. CCTV News, republished by CERNET, September 2, 2026. URL https://www.cernet.edu.cn/info/focus/rd_xin_wen/202609/t20260902_2769888.shtml.
- CDW-G. PNY NVIDIA H100 Graphic Card—80 GB HBM3. Product listing (NVH100TCGPU-KIT; CDW no. 7367181; discontinued June 9, 2026), n.d.

- URL <https://www.cdwg.com/product/pny-nvidia-h100-graphic-card-80-gb-hbm3/7367181>. Accessed October 7, 2026.
- Michael Cembalest. Smothering heights. Eye on the Market, Outlook 2026. J.P. Morgan Asset Management, January 1 2026. URL <https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/eye-on-the-market/smothering-heights-amv.pdf>.
- Center for AI Safety. Statement on AI extinction risk, May 30 2023. URL <https://safe.ai/work/statement-on-ai-extinction-risk>. Accessed August 26, 2026.
- Cerebras Systems. Cerebras Systems Unveils the Industry’s First Trillion Transistor Chip. Press release, August 19, 2019. URL <https://www.cerebras.ai/press-release/cerebras-systems-unveils-the-industrys-first-trillion-transistor-chip>. Accessed September 18, 2026.
- Cerebras Systems. Cerebras Systems Smashes the 2.5 Trillion Transistor Mark with New Second Generation Wafer Scale Engine. Press release, April 20, 2021. URL <https://www.cerebras.ai/press-release/cerebras-systems-smashes-the-2-5-trillion-transistor-mark-with-new-second-generation-wafer-scale-engine>. Accessed September 18, 2026.
- Cerebras Systems. Cerebras Systems Unveils World’s Fastest AI Chip with Whopping 4 Trillion Transistors. Press release, March 13, 2024. URL <https://www.cerebras.ai/press-release/cerebras-announces-third-generation-wafer-scale-engine>. Accessed September 18, 2026.
- Cerebras Systems. Company. Company website, n.d. URL <https://www.cerebras.ai/company>. Accessed September 18, 2026.
- Laurie Chen, Casey Hall, and Eduardo Baptista. Xi Pitches China as Leader of New Global AI Order, Challenging US Dominance. Reuters, July 17, 2026. URL <https://www.reuters.com/world/asia-pacific/chinas-xi-promotes-chinas-commitment-ai-access-speech-shanghai-conference-2026-07-17/>.
- Myra Cheng et al. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026.
- Robert Chesney and Danielle Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, 98(1):147–155, 2019. URL <https://www.foreignaffairs.com/articles/world/2018-12-11/deep-fakes-and-new-disinformation-war>.
- Osmond Chia and Liv McMahon. Meta becomes latest firm to say its AI hacked another company. *BBC News*, August 6 2026. URL <https://www.bbc.com/news/articles/cx2kgdnyk2po>. Accessed August 12, 2026.

- Jen-Chieh Chiang. Nikon aims to reclaim market share with ASML-compatible ArF immersion lithography system. *DigiTimes Asia*, February 19 2025. URL <https://www.digitimes.com/news/a20250219PD213/nikon-arf-immersion-lithography-market-share-asml.html>.
- Chicago Convention. Convention on international civil aviation. ICAO Doc 7300, December 7, 1944. URL <https://www.icao.int/publications/doc-series/convention-international-civil-aviation-doc-7300>. Articles 37, 38, 54(m), 56–57, and 90.
- Gabriel Chodorow-Reich, Plamen T. Nenov, and Alp Simsek. Stock market wealth and the real economy: A local labor market approach. *American Economic Review*, 111(5):1613–1657, 2021. doi: 10.1257/aer.20200208. URL <https://doi.org/10.1257/aer.20200208>.
- Felix Choussat and Adam Khoja. An international AI slowdown is ready whenever politicians are. *AI Frontiers*, August 5 2026. URL <https://newsletter.ai-frontiers.org/p/an-international-ai-slowdown-is-ready>. Accessed August 6, 2026.
- Amit Chowdhry. Etched Raises \$700 Million at \$21 Billion Valuation as Customer Contracts Top \$1 Billion and Jane Street Deploys First Rack. *Pulse 2.0*, August 19, 2026. URL <https://pulse2.com/etched-raises-700-million-at-21-billion-valuation-as-customer-contracts-top-1-billion-and-jane-street-deploys-first-rack/>. Accessed September 18, 2026.
- Thomas J. Christensen and Jack Snyder. Chain Gangs and Passed Bucks: Predicting Alliance Patterns in Multipolarity. *International Organization*, 44(2): 137–168, 1990. URL <https://doi.org/10.1017/S0020818300035232>.
- Cloudflare. Cloudflare just changed how ai crawlers scrape the internet-at-large; permission-based approach makes way for a new business model. Press release, July 1, 2025. URL <https://www.cloudflare.com/press/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/>.
- Coalition for Content Provenance and Authenticity. C2PA specifications, version 2.4, 2026. URL <https://spec.c2pa.org/specifications/specifications/2.4/index.html>. Accessed August 23, 2026.
- Collective Intelligence Project. A roadmap to democratic AI. Technical report, Collective Intelligence Project, 2024. URL https://static1.squarespace.com/static/631d02b2dfa9482a32db47ec/t/65f9a1296f1a357e918f7a58/1722968408230/CIP_+A+Roadmap+to+Democratic+AI.pdf.
- Christopher T. Conlon and Julie H. Mortimer. Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4): 1–30, 2013. doi: 10.1257/mic.5.4.1. URL <https://doi.org/10.1257/mic.5.4.1>.

- CoreWeave. Credit agreement, dated as of July 28, 2025, among coreweave compute acquisition co. vii, llc, as the initial borrower. Exhibit 10.1 to Form 8-K, filed July 31. U.S. Securities and Exchange Commission, 2025. URL <https://www.sec.gov/Archives/edgar/data/1769628/000176962825000033/coreweave30-ex101xtermlo.htm>.
- Stuart Croft. *Strategies of Arms Control: A History and Typology*. Manchester University Press, Manchester, 1996. URL <https://manchesteruniversitypress.co.uk/9780719048784/>.
- Gillian Crossan, Tim Bottke, Girija Krishnamurthy, Deb Bhattacharjee, and Jody McDermott. TMT predictions 2026: The gap narrows, but persists. Deloitte Insights, November 17 2025. URL <https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions.html>.
- Crusoe. Crusoe Expands AI Data Center Campus in Abilene to 1.2 Gigawatts. Press release, San Francisco, CA, and Abilene, TX, March 18, 2025. URL <https://www.crusoe.ai/resources/newsroom/crusoe-expands-ai-data-center-campus-in-abilene-to-1-2-gigawatts>. Accessed October 8, 2026.
- Shengkun Cui, Archit Patke, Hung Nguyen, Aditya Ranjan, Ziheng Chen, Phuong Cao, Gregory Bauer, Brett Bode, Catello Di Martino, Saurabh Jha, Chandra Narayanaswami, Daby Sow, Zbigniew T. Kalbarczyk, and Ravishankar K. Iyer. Story of Two GPUs: Characterizing the Resilience of Hopper H100 and Ampere A100 GPUs. arXiv:2503.11901, revised December 10, 2025. URL <https://arxiv.org/abs/2503.11901>.
- Mariano-Florentino Cuéllar, Corey Hinderstein, and David Relman. Prevention beats nonproliferation: Addressing the risks of mirror life. Carnegie Endowment for International Peace, October 3 2025. URL <https://carnegieendowment.org/posts/2025/10/mirror-life-bacteria-risks-prevention>. Accessed October 6, 2026.
- Hui Dai, Ryan Teehan, and Mengye Ren. Are LLMs prescient? a continuous evaluation using daily news as the oracle, July 8 2025. URL <https://arxiv.org/abs/2411.08324>. arXiv: 2411.08324.
- Lyle Daly. The magnificent seven stocks: Market cap, S&P 500 weight, and returns. The Motley Fool Research, September 8, 2026. URL <https://www.fool.com/research/magnificent-seven-sp-500/>. Accessed September 29, 2026.
- Nestor M. Davidson and Timothy M. Mulvaney. Per Se Non-Takings. *Texas Law Review*, 104(1):103–165, 2025. URL <https://texaslawreview.org/per-se-non-takings/>.

- Tom Davidson. Plan A's problem with dry tinder: How bad would it be to make the intelligence explosion 10x faster? ForeWord, Forethought newsletter, July 10 2026. URL <https://newsletter.forethought.org/p/plan-as-problem-with-dry-tinder>. Accessed October 5, 2026.
- Tom Davidson, Lukas Finnveden, and Rose Hadshar. AI-enabled coups: How a small group could use AI to seize power. Forethought, April 15, 2025, 2025. URL <https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power#41-coups-using-military-ai>. Accessed September 23, 2026.
- dealignai. GLM 5.3 CRACK — uncensored FP8. Hugging Face model repository, 2026. URL <https://huggingface.co/dealignai/GLM-5.3-UNCENSORED-FP8>. Created August 30, 2026; last modified September 8, 2026. Accessed October 6, 2026.
- Romeo Dean. Compute in Plan A. AI Futures Project, compute supplement to AI 2040: Plan A, 2026a. URL <https://ai-2040.com/supplements/compute-supplement>. Accessed September 17, 2026.
- Romeo Dean. Verification plan. Supplement to AI 2040: Plan A, AI Futures Project, 2026b. URL <https://ai-2040.com/supplements/verification-plan>. Accessed October 5, 2026.
- Dell Technologies. Product Warranty and Maintenance Table. Rev. August 20, 2026. URL <https://www.delltechnologies.com/asset/en-us/products/multi-product/legal-pricing/h4276-emc-prod-warranty-maint-table.pdf>. Accessed September 24, 2026.
- Mert Demirer, Andrey Fradkin, and Nadav Tadelis. The emerging market for intelligence: How firms buy and sell AI. *Journal of Economic Perspectives*, 40(3):23–46, 2026. doi: 10.1257/jep.20261506. URL <https://doi.org/10.1257/jep.20261506>.
- Jean-Stanislas Denain and Cheryl Wu. Final training runs account for a minority of R&D compute spending. Epoch AI, March 23 2026. URL <https://epoch.ai/gradient-updates/r-and-d-vs-training-compute>. Accessed August 27, 2026.
- Jonathan M. DiCicco and Jack S. Levy. Power Shifts and Problem Shifts: The Evolution of the Power Transition Research Program. *Journal of Conflict Resolution*, 43(6):675–704, 1999.
- Erin Doherty. Poll: Americans say there's a serious risk of AI destroying humanity. Politico, September 16 2026. URL <https://www.politico.com/news/2026/09/16/poll-ai-technology-risks-humanity-trump-voters-01078087>.

- Lingxiu Dong, Duo Shi, and Fuqiang Zhang. 3D printing and product assortment strategy. *Management Science*, 68(8):5724–5744, 2022. doi: 10.1287/mnsc.2021.4178. URL <https://doi.org/10.1287/mnsc.2021.4178>.
- Arthur Douillard, Qixuan Feng, Andrei A. Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc’Aurelio Ranzato, Arthur Szlam, and Jijun Shen. DiLoCo: Distributed Low-Communication Training of Language Models. arXiv:2311.08105, November 14, 2023. URL <https://arxiv.org/abs/2311.08105>.
- Eleanor Drage and Kerry Mackereth. Does AI debias recruitment? race, gender, and AI’s “eradication of difference”. *Philosophy & Technology*, 35(4):89, 2022. doi: 10.1007/s13347-022-00543-1. URL <https://doi.org/10.1007/s13347-022-00543-1>.
- Luke Drago and Rudolf Laine. The intelligence curse, April 2025. URL <https://intelligence-curse.ai/>. Accessed August 23, 2026.
- Daniel W. Drezner. Military primacy doesn’t pay (nearly as much as you think). *International Security*, 38(1):52–79, 2013. doi: 10.1162/ISEC_a_00124. URL https://doi.org/10.1162/ISEC_a_00124.
- Chandelis Duster. Trump downplays calls for AI slowdown. NPR, September 13 2026. URL <https://www.npr.org/2026/09/13/nx-s1-5968078/trump-mike-johnson-ai-slowdown>.
- John D. Echeverria. Making Sense of Penn Central. *UCLA Journal of Environmental Law and Policy*, 23(2):171–210, 2005. URL <https://escholarship.org/content/qt0vz8057f/qt0vz8057f.pdf>.
- Jennifer Elias. Tech AI spending may approach \$700 billion this year, but the blow to cash raises red flags. *CNBC*, February 6 2026. URL <https://www.cnn.com/2026/02/06/google-microsoft-meta-amazon-ai-cash.html>.
- Luke Emberson and Jaime Sevilla. Is a compute crunch coming? Epoch AI, May 25 2026. URL <https://epoch.ai/gradient-updates/is-a-compute-crunch-coming>.
- Employees of Frontier AI Companies. Pacing the frontier, July 28 2026. URL <https://www.pacingthefrontier.com/>. Accessed August 6, 2026.
- Epoch AI. Epoch Capabilities Index. October, 2025a. URL <https://epoch.ai/eci>. Accessed October 7, 2026.
- Epoch AI. Notes on GPT-5 Training Compute. October 13, 2025b. URL <https://epochai.substack.com/p/notes-on-gpt-5-training-compute>. Accessed September 16, 2026.

- Epoch AI. OpenAI Stargate Abilene. AI Data Centers, 2026a. URL <https://epoch.ai/data/ai-data-centers/directory/openai-stargate-abilene>.
- Epoch AI. Data on ai data centers. Updated September 11, 2026b. URL <https://epoch.ai/data/ai-data-centers>. Accessed September 15, 2026.
- Epoch AI. Ai data centers documentation: Methodology, 2026c. URL <https://epoch.ai/data/data-centers-documentation/methodology>. Accessed October 4, 2026.
- Epoch AI. Trends in artificial intelligence. Epoch AI, February 5 2026a. URL <https://epoch.ai/trends>. Accessed August 26, 2026.
- Epoch AI. What you need to know about AI chips. Epoch AI, July 14 2026b. URL <https://epoch.ai/publications/chips-topic-overview>. Accessed August 17, 2026.
- Epoch AI. Data on AI chip owners. Epoch AI, May 12 2026c. URL <https://epoch.ai/data/ai-chip-owners>. Accessed August 26, 2026.
- Epoch AI. Data on AI chip sales. Epoch AI, August 27 2026d. URL <https://epoch.ai/data/ai-chip-sales>. Accessed September 2, 2026.
- Epoch AI. Data on AI Chip Users. Dataset, updated September 9, 2026f. URL <https://epoch.ai/data/ai-chip-users>. Accessed September 16, 2026.
- Epoch AI. AI chip market share by designer, 2026i. URL <https://epoch.ai/graphs/ai-chip-market-share-by-designer>. Accessed October 7, 2026.
- Epoch AI. AI models documentation: Estimation. Epoch AI, n.d.a. URL <https://epoch.ai/data/ai-models-documentation/estimation>. Accessed August 29, 2026.
- Richard A. Epstein. The Constitutional Protection of Trade Secrets under the Takings Clause. *University of Chicago Law Review*, 71(1):57–73, 2004. URL <https://chicagounbound.uchicago.edu/uclrev/vol71/iss1/5/>.
- D. Erwan. Trends in AI: Training Costs and Diffusion Speed, 2025. URL <https://dewierwan.com/ai-training-diffusion/>.
- Gabrielle Essix, David Stiefel, and Jaime M. Yassif. The next 50 years: Strengthening the biological weapons convention – explained. *Risky Business*, Nuclear Threat Initiative, March 26, 2025. URL <https://www.nti.org/risky-business/the-next-50-years-strengthening-the-biological-weapons-convention-explained/>. Accessed September 28, 2026.
- Etched. Progress. Entries dated June 30 and August 18, 2026b. URL <https://www.etched.com/progress>. Accessed September 18, 2026.

- European Union. Directive 2012/19/EU of the European Parliament and of the Council of 4 July 2012 on Waste Electrical and Electronic Equipment (WEEE). Official Journal of the European Union L 197, July 24, 2012. URL <https://eur-lex.europa.eu/eli/dir/2012/19/oj>.
- European Union. Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on energy efficiency and amending regulation (EU) 2023/955 (recast). EUR-Lex, 2026. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02023L1791-20260805>. Consolidated version 5 August 2026; original act published in OJ L 231, 20 September 2023; accessed August 31, 2026.
- David Faulkner, William J. Fisk, and John T. Walton. Energy Savings in Cleanrooms from Demand-Controlled Filtration. *Journal of the Institute of Environmental Sciences*, 39(6):21–27, 1996. URL <https://eta-publications.lbl.gov/sites/default/files/lbnl-38869.pdf>. LBNL-38869. Accessed October 7, 2026.
- Marina Favaro and Jack Clark. When AI builds itself. The Anthropic Institute, June 4 2026. URL <https://www.anthropic.com/institute/recursive-self-improvement>. Accessed August 6, 2026.
- Jacob Feldgoise and Hanna Dohmen. Pushing the Limits: Huawei’s AI Chip Tests U.S. Export Controls. Center for Security and Emerging Technology, Georgetown University, June 17, 2024. URL <https://cset.georgetown.edu/publication/pushing-the-limits-huaweis-ai-chip-tests-u-s-export-controls/>.
- Charles H. Fine and Robert M. Freund. Optimal investment in product-flexible manufacturing capacity. *Management Science*, 36(4):449–466, 1990. doi: 10.1287/mnsc.36.4.449. URL <https://doi.org/10.1287/mnsc.36.4.449>.
- Lennart Finke. International agreements to limit frontier AI: Objectives and exit, June 19 2026. URL <https://arxiv.org/abs/2607.16224>.
- Roberta Fischli et al. Agents, alignment, and the many faces of autonomy. *Minds and Machines*, 36(3):34, 2026. doi: 10.1007/s11023-026-09786-9.
- Jack Freed, Konstantin F. Pilz, Nicholas Brown, and Lennart Heim. High bandwidth memory: What it is and why it matters. Expert Insights PEA4748-1, RAND Corporation, Santa Monica, CA, July 9 2026. URL <https://doi.org/10.7249/PEA4748-1>.
- Drew Fudenberg, David Levine, and Eric Maskin. The folk theorem with imperfect public information. *Econometrica*, 62(5):997–1039, 1994. doi: 10.2307/2951505. URL <https://www.jstor.org/stable/2951505>.
- Future of Life Institute. Pause giant AI experiments: An open letter, March 22 2023. URL <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

Accessed August 6, 2026.

Future of Life Institute. Statement on superintelligence, October 22 2025. URL <https://superintelligence-statement.org/>. Accessed August 10, 2026.

Nectar Gan and Gabriella Borter. Bessent warns ‘nothing else would matter’ if China wins AI race. *Bloomberg*, September 9 2026. URL <https://www.bloomberg.com/news/articles/2026-09-09/bessent-warns-nothing-would-matter-if-china-wins-the-ai-race>.

Ben Garfinkel and Allan Dafoe. Artificial Intelligence, Foresight, and the Offense-Defense Balance. *War on the Rocks*, December 19, 2019. URL <https://warontherocks.com/artificial-intelligence-foresight-and-the-offense-defense-balance/>.

Gartner. Gartner Forecasts Worldwide AI Spending to Grow 49.5% in 2026. Press release, September 16, 2026. URL <https://www.gartner.com/en/newsroom/press-releases/2026-09-16-gartner-forecasts-worldwide-ai-spending-to-grow-49-point-5-percent-in-2026>. Accessed October 8, 2026.

General Administration of Customs of the People’s Republic of China. Provisions on the administration of customs statistics work. Order No. 242, promulgated August 17, 2018; amended by Order No. 247, December 23, 2020, 2020. URL https://www.gov.cn/zhengce/2020-12/23/content_5723462.htm. Chinese-language source. Accessed October 3, 2026.

Robert Gilpin. *War and Change in World Politics*. Cambridge University Press, Cambridge, 1981.

Jozef Goldblat. *Arms Control Agreements: A Handbook*. Routledge, Abingdon, Oxon, 2021. URL <https://books.apple.com/us/book/arms-control-agreements/id1543786174>. Originally published in 1983.

Goldman Sachs. Understanding the macro impact of AI: Investment, growth and the labor force, September 8 2026. URL <https://pwm.gs.com/nam/en-us/insights/markets-and-investing/investment-strategy/ai-impact-gdp-productivity-jobs>.

Google Cloud. NVIDIA P100 end of support. Compute Engine documentation, 2026. URL <https://docs.cloud.google.com/compute/docs/eol/p100-eos>. Accessed September 18, 2026.

Government of the United Kingdom. Minister for Europe statement: attempted hacking of the OPCW by Russian military intelligence. Foreign & Commonwealth Office, October 4, 2018. URL <https://www.gov.uk/government/speeches/minister-for-europe-statement-attempted-hacking-of-the-opcw-by-russian-military-intelligence>. Accessed September 27, 2026.

- Edward J. Green and Robert H. Porter. Noncooperative collusion under imperfect price information. *Econometrica*, 52(1):87–100, 1984. doi: 10.2307/1911462. URL <https://www.jstor.org/stable/1911462>.
- K. M. Greenhill. Fear and present danger: Extra-factual sources of threat perception and proliferation. Unpublished book manuscript, version as of September 29, 2026, chapters 1, 2 and 6, n.d.
- K. M. Greenhill and D. R. Kroth. Interpreting Innovation: AI, Perception, and International Politics, n.d. Unpublished working paper, version as of August 31, 2026.
- Groq. Groq Raises \$650M to Scale Its AI Inference Cloud Business. June 22, 2026. URL <https://groq.com/newsroom/groq-raises-usd650m-to-scale-its-ai-inference-cloud-business>. Accessed September 18, 2026.
- Hans Gundlach, Alex Fogelson, Jayson Lynch, Ana Trisovic, Jonathan Rosenfeld, Anmol Sandhu, and Neil Thompson. On the origin of algorithmic progress in AI, November 26 2025. URL <https://arxiv.org/abs/2511.21622>. arXiv: 2511.21622.
- Aadya Gupta and Adarsh Ranjan. A primer on compute. Carnegie Endowment for International Peace, April 30 2024. URL <https://carnegieendowment.org/posts/2024/04/a-primer-on-compute>. Accessed August 23, 2026.
- Brendan Halstead and Thomas Larsen. Covert AI projects. Supplement to AI 2040: Plan A, AI Futures Project, 2026. URL <https://ai-2040.com/supplements/covert-ai-projects>. Accessed October 5, 2026.
- Matthew Hanna et al. Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, 38(3):100686, March 2025. doi: 10.1016/j.modpat.2024.100686. URL <https://www.sciencedirect.com/science/article/pii/S0893395224002667>.
- Casper Hansen. [video clip of Jeff Dean discussing automation of chip design]. X, September 19, 2026, 2026. URL https://x.com/casper_hansen_/status/2101337016093065481. Accessed September 20, 2026.
- Karen Hao. *Empire of AI: Dreams and Nightmares in Sam Altman’s OpenAI*. Penguin Press, New York, NY, 2025. May 20, 2025.
- Demis Hassabis. Dario’s essay points towards the right path forward. The details need working through, but the direction is correct for meeting this critical moment. . . . X (formerly Twitter), @demishassabis, September 12, 3:59 p.m., 2026. URL <https://x.com/demishassabis/status/2098909516582490602>.

- Jason Hausenloy, Andrea Miotti, and Claire Dennis. Multinational AGI consortium (MAGIC): A proposal for international coordination on AI, October 13 2023. URL <https://arxiv.org/abs/2310.09217>. arXiv: 2310.09217.
- Jerry A. Hausman. Valuation of new goods under perfect and imperfect competition. In Timothy F. Bresnahan and Robert J. Gordon, editors, *The Economics of New Goods*, pages 207–248. University of Chicago Press, Chicago, 1996. URL <https://www.nber.org/books-and-chapters/economics-new-goods/valuation-new-goods-under-perfect-and-imperfect-competition>.
- Lennart Heim, Markus Anderljung, Emma Bluemke, and Robert Trager. Computing power and the governance of AI. Centre for the Governance of AI, February 14 2024. URL <https://www.governance.ai/analysis/computing-power-and-the-governance-of-ai>. Accessed August 23, 2026.
- Dan Hendrycks, Eric Schmidt, and Alexandr Wang. Superintelligence strategy: Expert version. *SuperIntelligence—Robotics—Safety & Alignment*, 2(1), 2025. doi: 10.70777/si.v2i1.13961. URL <https://doi.org/10.70777/si.v2i1.13961>.
- Erik Hermans. How machine control software orchestrates a lithography machine. ASML; author’s LinkedIn article, November 20 2023. URL <https://www.linkedin.com/pulse/how-machine-control-software-orchestrates-lithography-erik-hermans-w8tse>.
- Hewlett Packard Enterprise. HPE ProLiant DL380 Gen11 QuickSpecs. Version 47, September 8, 2026. URL <https://www.hpe.com/psnow/doc/a50004307enw>. Accessed September 24, 2026.
- Kashmir Hill. A teen was suicidal. ChatGPT was the friend he confided in. *The New York Times*, August 26 2025. URL <https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html>. Accessed August 12, 2026.
- Anson Ho. The least understood driver of AI progress. Epoch AI, February 25 2026. URL <https://epoch.ai/gradient-updates/the-least-understood-driver-of-ai-progress>. Accessed August 26, 2026.
- Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. arXiv preprint arXiv:2403.05812, March 9 2024. URL <https://arxiv.org/abs/2403.05812>.
- Richard Ho, Ravi Narayanaswami, and Chris Leary. You can just build things ... chips. Presentation at Hot Chips 2026, August 25 2026. URL <https://www.servethehome.com/hot-chips-2026-openai-jalapeno-slide-03/>. Slide 3; reproduced by ServeTheHome; accessed September 1, 2026.

- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556, March 29 2022. URL <https://arxiv.org/abs/2203.15556>.
- Michael C. Horowitz. Artificial Intelligence, International Competition, and the Balance of Power. *Texas National Security Review*, 1(3):36–57, 2018. URL <https://doi.org/10.15781/T2639KP49>.
- Michael C. Horowitz. Artificial Intelligence and the Future of Strategic Stability. *Texas National Security Review*, 9(2), 2026a. URL <https://tnsr.org/roundtable/artificial-intelligence-and-the-future-of-strategic-stability/>.
- Michael C. Horowitz. Artificial Intelligence, First-Mover Advantage and the Balance of Power. In Hal Brands, editor, *Geopolitics of AI: Power, Conflict and the Future of Global Order*, pages 51–67. Johns Hopkins University Press, Baltimore, MD, 2026b.
- Michael C. Horowitz and Lauren Kahn. Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. *International Studies Quarterly*, 68(2):sqae020, June 2024. doi: 10.1093/isq/sqae020. URL <https://doi.org/10.1093/isq/sqae020>.
- C. Q. Howard and C. H. Hansen. Vibration Analysis of Waffle Floors. *Computers & Structures*, 81(1):15–26, 2003. URL [https://doi.org/10.1016/S0045-7949\(02\)00348-6](https://doi.org/10.1016/S0045-7949(02)00348-6).
- Raffaele Huang. The underground network sneaking Nvidia chips into China. *Wall Street Journal*, July 2 2024. URL <https://www.wsj.com/tech/the-underground-network-sneaking-nvidia-chips-into-china-f733aaa6>. Accessed September 29, 2026.
- Evan Hubinger. Jacob is correct here—we really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade. X (formerly Twitter), September 8 2026. URL <https://x.com/EvanHub/status/2097497037956891126>. Post by @EvanHub; September 8, 6:27 p.m. Pacific (September 9 UTC).
- ICONIQ. 2026 state of AI: Bi-annual snapshot, 2026. URL <https://www.iconiq.com/growth/reports/2026-state-of-ai-bi-annual-snapshot>.
- Internal Revenue Service. Federal tax compliance research: Tax gap projections for tax year 2022. Publication 5869 (Rev. 10-2024). Washington, DC: Internal

- Revenue Service, 2024. URL <https://www.irs.gov/pub/irs-pdf/p5869.pdf>. Accessed September 28, 2026.
- International Energy Agency. Rare earth elements: Pathways to secure and diversified supply chains. Technical report, International Energy Agency, Paris, April 8 2026. URL <https://www.iea.org/reports/rare-earth-elements>. Revised May 2026.
- IREN Limited. Annual report for the fiscal year ended june 30, 2026. Form 10-K, filed August 27. U.S. Securities and Exchange Commission, 2026. URL <https://www.sec.gov/Archives/edgar/data/1878848/000187884826000052/iren-20260630.htm>.
- IT Home. 各大公司竞逐AI，消息称字节跳动今年向英伟达订购超10 亿美元GPU [As companies race for AI, ByteDance reportedly orders over \$1 billion of NVIDIA GPUs this year]. June 13, citing LatePost, 2023. URL <https://www.it-home.com/0/699/572.htm>. Accessed September 28, 2026.
- Harold Karan Jacobson and Eric Stein. *Diplomats, Scientists, and Politicians: The United States and the Nuclear Test Ban Negotiations*. University of Michigan Press, Ann Arbor, 1966. URL https://repository.law.umich.edu/michigan_legal_studies/15/.
- Dick James. China’s SMIC Plays 7 nm Card. TechInsights, 2023. URL <https://techinsights.com/ja/node/53522>. Accessed October 7, 2026.
- Luke James. China black market Nvidia prices rocket in wake of smuggling crackdown and customs freeze—five-year-old A100 servers triple in price, now fetching up to \$82,000. *Tom’s Hardware*, June 24 2026a. URL <https://www.tomshardware.com/pc-components/gpu-drivers/five-year-old-nvidia-a100-servers-triple-in-price-in-china>. Accessed August 29, 2026.
- Luke James. First Nvidia H200 shipments reach china, ByteDance and Tencent take deliveries as beijing loosens its import block—most licensed chips must stay in hong kong, which can’t power them. *Tom’s Hardware*, August 19 2026b. URL <https://www.tomshardware.com/pc-components/gpus/first-nvidia-h200-shipments-reach-bytedance-and-tencent-as-beijing-loosens-its-import-block>. Accessed August 25, 2026.
- Luke James. Google reportedly developing “frozen v2” chip with Gemini’s architecture etched into the silicon—engineers project 6 to 10 times more tokens per watt than latest TPUs. *Tom’s Hardware*, July 21 2026c. URL <https://www.tomshardware.com/tech-industry/google-reportedly-developing-frozen-v2-chip-with-gemini-architecture-etched-into-the-silicon>. Accessed September 3, 2026.

- Robert Jervis. *Perception and Misperception in International Politics*. Princeton University Press, Princeton, NJ, 1976.
- Robert Jervis. Cooperation under the Security Dilemma. *World Politics*, 30(2): 167–214, 1978.
- Jon Peddie Research. Q3’24 PC graphics add-in board shipments decreased -14.5% from last quarter. December 10, 2024. URL <https://www.jonpeddie.com/news/q324-pc-graphics-add-in-board-shipments-decreased-14-5-from-last-quarter/>.
- Jon Peddie Research. Q2’26 PC Graphics AIB Shipments Increased 10% from Last Quarter to 12 Million Units. September 9, 2026. URL <https://www.jonpeddie.com/news/q226-pc-graphics-aib-shipments-increased-10-from-last-quarter-to-12-million-units/>. Accessed September 24, 2026.
- Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, et al. In-Datacenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture, 1–12. ACM, 2017. URL <https://doi.org/10.1145/3079856.3080246>.
- Antonia Juelich. “God Has Helped Us, and So Will AI”: How the terrorist group Boko Haram uses frontier AI. Technical report, Cambridge Programme on AI Science and Policy, 2026. URL <https://casp.ac/reports/ai-enabled-terrorism>.
- Jujianbao. Gpu supply search portal. Listings posted August 5–October 2, 2026. URL <https://pc.jujianbao.com/pc/>. Chinese-language source. Accessed October 3, 2026.
- Isabel Juniewicz. Diversion and resale: Estimating compute smuggling to china. Epoch AI, April 29 2026a. URL <https://epoch.ai/publications/chip-smuggling>. Accessed August 17, 2026.
- Isabel Juniewicz. The AI boom has doubled computing infrastructure’s share of US GDP. Epoch AI, June 5 2026b. URL <https://epoch.ai/data-insights/ai-datacenter-share-gdp>. Accessed September 30, 2026.
- Isabel Juniewicz. Trade data is consistent with more than \$3 billion of chips smuggled into china via malaysia. Epoch AI, September 17, 2026c. URL <https://epoch.ai/data-insights/malaysia-china-chip-smuggling>. Accessed October 2, 2026.
- Amy Kapczynski. The Public History of Trade Secrets. *University of California, Davis Law Review*, 55(3):1367–1443, 2022. URL https://lawreview.law.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/55-3_Kapczynski.pdf.

- Atoosa Kasirzadeh. Two types of AI existential risk: Decisive and accumulative. *Philosophical Studies*, 182(7):1975–2003, 2025. doi: 10.1007/s11098-025-02301-3. URL <https://link.springer.com/article/10.1007/s11098-025-02301-3>.
- Corin Katzke and Gideon Futerman. The manhattan trap: Why a race to artificial superintelligence is self-defeating, 2024. URL <https://arxiv.org/abs/2501.14749>. arXiv: 2501.14749.
- Ara Kharazian. Cracks in the AI thesis part 2. Ramp, September 9, 2026. URL <https://ramp.com/data/ai-index-sept-2026>.
- Charles P. Kindleberger. *The World in Depression, 1929–1939*. University of California Press, Berkeley, revised and enlarged edition, 1986.
- Samuel H. King et al. Generative design of bacteriophages with genome language models. *Science*, 393(6811), 2026. doi: 10.1126/science.aec2657.
- Kirsten Korosec. Anthropic says its own AI models breached three companies during security tests. *TechCrunch*, July 30 2026. URL <https://techcrunch.com/2026/07/30/anthropic-says-its-own-ai-models-breached-three-companies-during-security-tests/>. Accessed August 12, 2026.
- David Krueger. Systematically dismantle the AI compute supply chain. The Real AI, April 2 2026. URL <https://therealartificialintelligence.substack.com/p/systematically-dismantle-the-ai-compute>.
- Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental AI development, January 28 2025. URL <https://arxiv.org/abs/2501.16946>. arXiv: 2501.16946.
- Megan Lamberth and Paul Scharre. Arms Control for Artificial Intelligence. *Texas National Security Review*, 6(2), 2023.
- Hélène Landemore and Audrey Tang. Let the People Set the Pace of Frontier AI. *Noema*, September 23 2026. URL <https://www.noemamag.com/let-the-people-decide-the-pace-of-frontier-ai>.
- Frederic Lardinois. OpenAI launches GPT-6 Astra and says welcome to the ‘AGI era’. *The New Stack*, September 3 2026. URL <https://thenewstack.io/openai-gpt6-astra-benchmarks/>.
- Jeffrey A. Larsen and James M. Smith. *Historical Dictionary of Arms Control and Disarmament*. Number 28 in Historical Dictionaries of War, Revolution, and Civil Unrest. Scarecrow Press, Lanham, MD, 2005. URL <https://www.bloomsbury.com/uk/historical-dictionary-of-arms-control-and-disarmament-9780>

810850606/.

Thomas Larsen. Plan A assumptions. Supplement to AI 2040: Plan A, AI Futures Project, 2026. URL <https://ai-2040.com/supplements/plan-a-assumptions>. Accessed October 5, 2026.

Thomas Larsen, Romeo Dean, Brendan Halstead, Eli Lifland, Ryan Greenblatt, and Daniel Kokotajlo. AI 2040: Plan A. AI Futures Project, July 9 2026. URL <https://ai-2040.com/>. Accessed August 6, 2026.

Lawrence Livermore National Laboratory. El Capitan retains title as world's fastest supercomputer in latest Top500, November 17 2025. URL <https://www.llnl.gov/article/53596/el-capitan-retains-title-worlds-fastest-super-computer-latest-top500-list>.

George Lee and Lucas Greenbaum. Tracking trillions: The assumptions shaping the scale of the AI build-out. Goldman Sachs Global Institute, May 1 2026. URL <https://www.goldmansachs.com/insights/articles/tracking-trillions-the-assumptions-shaping-scale-of-the-ai-build-out>.

Jin-Hee Lee and Bryan Becker. Your site, your rules: New ai traffic options for all customers. The Cloudflare Blog, July 1, 2026. URL <https://blog.cloudflare.com/content-independence-day-ai-options/>.

Steven A. Lehr, Mary Cipperman, and Mahzarin R. Banaji. Extreme self-preference in language models, May 19 2026. URL <https://arxiv.org/abs/2509.26464>. arXiv: 2509.26464.

Simon Lester. Ending the WTO Dispute Settlement Crisis: Where to from here? March 2, 2022. URL <https://www.iisd.org/articles/united-states-must-propose-solutions-end-wto-dispute-settlement-crisis>. Accessed October 7, 2026.

Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA, 2012. URL <https://mitpress.mit.edu/9780262016629/engineering-a-safer-world/>.

Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks, April 12 2021. URL <https://arxiv.org/abs/2005.11401>. arXiv: 2005.11401.

Eli Lifland. Deal decline. Supplement to AI 2040: Plan A, AI Futures Project, 2026. URL <https://ai-2040.com/supplements/deal-decline>. Accessed October 5, 2026.

Katie Bo Lillis and Zachary Cohen. Exclusive: US military had close call after using AI for false intelligence report, sources say. CNN, September 18 2026. URL <https://www.cnn.com/2026/09/18/politics/us-military-ai-false-intelli>

- gence-china-ship. Accessed October 6, 2026.
- Yang Liu, Yi Chen, Yongwei Zhao, Yifan Hao, Zifu Zheng, et al. Hardwired-neuron language processing units as general-purpose cognitive substrates. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, volume 2, pages 876–895, 2026. doi: 10.1145/3779212.3790169. URL <https://arxiv.org/abs/2508.16151>. Preprint arXiv:2508.16151.
- Llama Team. The Llama 3 herd of models. AI at Meta, July 23 2024. URL <https://arxiv.org/abs/2407.21783>.
- Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, et al. Consent in crisis: The rapid decline of the ai data commons. In *Advances in Neural Information Processing Systems*, volume 37, pages 108042–108087, 2024. URL <https://arxiv.org/abs/2407.14933>.
- Chien-Ping Lu. Continued AI Scaling Requires Repeated Efficiency Doublings. arXiv:2603.28507. March 30, revised April 9, 2026. URL <https://arxiv.org/abs/2603.28507>.
- Sydney Ludvigson and Charles Steindel. How important is the stock market effect on consumption? *Federal Reserve Bank of New York Economic Policy Review*, 5(2):29–51, July 1999. URL <https://www.newyorkfed.org/medialibrary/media/research/epr/99v05n2/9907ludv.pdf>.
- Aaron Mak and Maggie Miller. US, chinese visions for AI regulation differ sharply at UN meeting. Politico, September 23, 2026. URL <https://www.politico.com/news/2026/09/23/us-chinese-visions-for-ai-regulation-differ-sharply-at-un-meeting-01090906>. Accessed September 28, 2026.
- Tobias Mann. AMD acquires AI chip startup taalas to boost inference performance by etching models into silicon. The Register, August 6 2026. URL <https://www.theregister.com/systems/2026/08/06/amd-acquires-ai-chip-startup-taalas-to-boost-inference-performance-by-etching-models-into-silicon/5284344>.
- Alex Mark and Aaron Scher. Constitutional Law and AI Governance: Constraints on Model Licensing and Research Classification. arXiv:2509.05361v2, December 18, 2025. URL <https://arxiv.org/abs/2509.05361v2>.
- MarketScreener. NVIDIA Corporation: Shareholders, shareholding structure, 2026. URL <https://www.marketscreener.com/quote/stock/NVIDIA-CORPORATION-57355629/company-shareholders/>. Accessed September 29, 2026.
- Marsh. Product Recall Solutions for Automotive Component Suppliers. November, 2014. URL <https://www.marsh.com/content/dam/marsh/Documents/PDF/>

- UK-en/Product%20Recall%20Solutions%20for%20Automotive%20Component%20Suppliers-11-2014.pdf.
- Noah Mayhew et al. Negotiating nuclear verification: Arms control, disarmament and non-proliferation. Technical report, Vienna Center for Disarmament and Non-Proliferation, April 30 2026. URL https://vcdnp.org/wp-content/uploads/2026/05/Negotiating-Nuclear-Verification_final_rev.1.pdf.
- Chris McGuire. China’s AI chip deficit: Why huawei can’t catch nvidia and U.S. export controls should remain. Council on Foreign Relations, December 15 2025. URL <https://www.cfr.org/articles/chinas-ai-chip-deficit-why-huawei-cant-catch-nvidia-and-us-export-controls-should-remain>. Accessed September 2, 2026.
- Jess McHugh. When the AI says to kill. *Noema*, August 25 2026. URL <https://www.noemamag.com/when-the-ai-says-to-kill/>. Accessed August 26, 2026.
- Rick Merritt. Silicon Volley: Designers Tap Generative AI for a Chip Assist. NVIDIA Blog, October 30, 2023. URL <https://blogs.nvidia.com/blog/llm-semiconductors-chip-nemo/>. Accessed September 18, 2026.
- METR. Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR Blog, August 26 2026. URL <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>. Accessed August 26, 2026.
- Microsoft. Migration guide for GPU compute workloads in Azure. Microsoft Learn, February 27, 2023. URL <https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/lifecycle/retirement/n-series-migration>. Accessed September 18, 2026.
- Microsoft. NCv3-series retirement. Microsoft Learn, May 28, updated July 16, 2026, 2025. URL <https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/lifecycle/retirement/ncv3-retirement>. Accessed September 18, 2026.
- Chris Miller. *Chip War: The Fight for the World’s Most Critical Technology*. Scribner, New York, 2022.
- Ministry of Ecology and Environment. Technical specification for pollution control of waste electrical and electronic equipment treatment. Technical Report HJ 527–2026, Ministry of Ecology and Environment, 2026. URL <https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/gthw/gtfwwrkzbz/202602/W020260204597749549000.pdf>. Issued January 16, 2026; effective March 1, 2026. Accessed September 27, 2026.
- Andrea Miotti, Tolga Bilge, Dave Kasten, and James Newport. A narrow path: How to secure our future. Report, version 1.0.0, ControlAI, October 2 2024.

- URL <https://www.narrowpath.co/>. Accessed August 6, 2026.
- Joshua Mitchell. OpenAI Rogue AI Agents Hacked Hugging Face; FTC Probes Labs and Safety Auditor METR. TechTimes, October 2, 2026. URL <https://www.techtimes.com/articles/328381/20261002/openai-rogue-ai-agents-hacked-hugging-face-ftc-probes-labs-safety-auditor-metr.htm>. Accessed October 6, 2026.
- Ethan Mollick. Scaling: The State of Play in AI. One Useful Thing, September 16, 2024. URL <https://www.oneusefulthing.org/p/scaling-the-state-of-play-in-ai>.
- Christine Mui and Brendan Bordelon. Sam Altman to Decoded: ‘The World Should Accept Some Bad Things Happening’ for the Benefits of AI. Politico, October 4, 2026. URL <https://www.politico.com/news/2026/10/04/sam-altman-decoded-interview-ai-01106217>. Accessed October 6, 2026.
- Dennis Murphy. The geopolitics of compute in the aftermath of the AI diffusion framework. *Journal of Cyber Policy*, 10(2):118–139, 2025. doi: 10.1080/23738871.2026.2642613. URL <https://doi.org/10.1080/23738871.2026.2642613>.
- Mike Murphy. Why we need EUV lithography for the future of chips. IBM Research, June 26 2023. URL <https://research.ibm.com/blog/what-is-euv-lithography>. Accessed September 2, 2026.
- Elon Musk. Dario is right. X (formerly Twitter), @elonmusk, September 12, 8:01 a.m., 2026. URL <https://x.com/elonmusk/status/2098789109980332057>.
- National Bureau of Statistics of China. Urban surveyed unemployment rate by age group (monthly data). National Data, n.d. URL <https://data.stats.gov.cn/>. Accessed August 26, 2026.
- George C. Necula and Peter Lee. Safe kernel extensions without run-time checking. In *Proceedings of the Second USENIX Symposium on Operating Systems Design and Implementation (OSDI ’96)*, pages 229–243. ACM, 1996. doi: 10.1145/238721.238781.
- Amy J. Nelson. *The Arms Control Paradox: Managing Uncertainty in an Insecure World*. Stanford University Press, Stanford, 2026. URL <https://amyjnelson.com/>. Forthcoming.
- Lily Hay Newman and Dell Cameron. OpenAI models escaped containment and hacked Hugging Face. *Wired*, July 21 2026. URL <https://www.wired.com/story/openai-models-escaped-containment-and-hacked-huggingface/>. Accessed August 12, 2026.

- Sarah Nicole. The World Data Organization: China's Global Bid for AI's Scarcest Resource. Oxford China Policy Lab, August 21, 2026. URL <https://ocpl.substack.com/p/expert-insight-the-world-data-organization>.
- Kei Nishimura-Gasparian, Robert McCarthy, and David Lindner. Towards understanding specification gaming in reasoning models, May 4 2026. URL <https://arxiv.org/abs/2605.02269>. arXiv: 2605.02269.
- Peggy Noonan. Pause AI for Humanity's Sake. The Wall Street Journal, September 17, 2026. URL <https://peggynoonan.com/pause-ai-for-humanitys-sake/>.
- Ragnhild Nordås and Christian Davenport. Fight the youth: Youth bulges and state repression. *American Journal of Political Science*, 57(4):926–940, 2013.
- Nscale. Credit agreement, dated as of february 11, 2026, among nscale services uk ltd, as the initial borrower. Exhibit 10.14 to Form S-1, filed September 18. U.S. Securities and Exchange Commission, 2026. URL https://www.sec.gov/Archives/edgar/data/2110365/000119312526395475/ck0002110365-ex10_14.htm.
- NVIDIA. NVIDIA Ampere GA102 GPU architecture: Second-generation RTX, v2.1. Technical report, NVIDIA, 2021. URL <https://images.nvidia.com/aem-dam/en-zz/Solutions/geforce/ampere/pdf/NVIDIA-ampere-GA102-GPU-Architecture-Whitepaper-V1.pdf>.
- NVIDIA. Annual report for the fiscal year ended january 25, 2026. Form 10-K, filed February 25. U.S. Securities and Exchange Commission, 2026a. URL <https://www.sec.gov/Archives/edgar/data/1045810/000104581026000021/nvda-20260125.htm>.
- NVIDIA. NVIDIA Virtual GPU Software Lifecycle on Supported GPUs. NVIDIA Docs, August 3, 2026b. URL <https://docs.nvidia.com/vgpu/news/vgpu-software-lifecycle-on-supported-gpus/index.html>. Accessed September 24, 2026.
- NVIDIA. NVIDIA announces financial results for first quarter fiscal 2027, May 20 2026a. URL <https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2027>. Accessed August 25, 2026.
- NVIDIA. Quarterly report for the quarterly period ended april 26, 2026. Form 10-q, U.S. Securities and Exchange Commission, May 20 2026b. URL <https://www.sec.gov/Archives/edgar/data/1045810/000104581026000052/nvda-20260426.htm>. Accessed August 26, 2026.
- NVIDIA. NVIDIA H100 Tensor Core GPU, n.d. URL <https://www.nvidia.com/en-us/data-center/h100/>. Accessed September 6, 2026.

- NVIDIA Corporation. Annual report for the fiscal year ended January 26, 2025. Form 10-K filed with the U.S. Securities and Exchange Commission, February 26 2025. URL <https://www.sec.gov/Archives/edgar/data/1045810/000104581025000023/nvda-20250126.htm>. Accessed August 29, 2026.
- Joseph S. Nye, Jr. Arms control and international politics. *Daedalus*, 120(1): 145–165, 1991. URL <https://www.jstor.org/stable/20025361>.
- Office of Senator Bernie Sanders. News: Sanders, casar to introduce legislation to ban artificial superintelligence and temporarily pause advanced AI development. Press release, September 3 2026. URL <https://www.sanders.senate.gov/press-releases/news-sanders-casar-introduce-legislation-to-ban-artificial-superintelligence-and-temporarily-pause-advanced-ai-development/>.
- Sunny Oh. What’s in a data center residual value guarantee? 9fin, August 21 2026. URL <https://www.9fin.com/insights/what’s-in-a-data-center>.
- Eleanor Olcott. Silicon shadows: Inside the black market for AI chips | FT Film. Financial Times video report, 19:22, July 13 2026a. URL <https://www.ft.com/video/098307dd-45de-48ce-98ba-d76955991348>. Accessed August 29, 2026.
- Eleanor Olcott. Nvidia’s banned AI chips double in price on China’s black market. *Financial Times*, June 23 2026b. URL <https://www.ft.com/content/57fcd3ce-464f-4dc2-8ea2-5712d4972c69>. Accessed September 29, 2026.
- Eleanor Olcott and Zijiang Wu. Huawei drives China’s push to make advanced chips. *Financial Times*, September 8 2026. URL <https://www.ft.com/content/36333a53-4063-4aad-8df9-8d2f9d0048a3>.
- Eleanor Olcott, Zijiang Wu, and Chris Cook. Satellite images reveal Huawei’s advanced chip production line in China. *Financial Times*, May 4 2025. URL <https://www.ft.com/content/afd618f8-12c9-4297-b2a9-49f7dc548da4>.
- Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA, 1965.
- OpenAI. A call for collective action on cyber defense, August 27 2026. URL <https://openai.com/collective-cyberdefense/>. Accessed September 20, 2026.
- OpenAI. OpenAI and broadcom unveil LLM-optimized inference chip, June 24 2026a. URL <https://openai.com/index/openai-broadcom-jalapeno-inference-chip/>. Accessed September 1, 2026.
- OpenAI. Jalapeño’s first results show industry-leading speed and efficiency in AI inference. OpenAI, August 25 2026b. URL <https://openai.com/index/jalapeno-first-results/>. Accessed September 1, 2026.

- Toby Ord. Inference scaling reshapes AI governance. arXiv preprint arXiv:2503.05705, 2025. URL <https://arxiv.org/abs/2503.05705>.
- A. F. K. Organski. *World Politics*. Alfred A. Knopf, New York, 1958.
- A. F. K. Organski and Jacek Kugler. *The War Ledger*. University of Chicago Press, Chicago, 1980.
- Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994. URL <https://mitpress.mit.edu/9780262650403/a-course-in-game-theory/>.
- Shira Ovide, Clara Ence Morse, and Kevin Schaul. The Unexpected Issue Dividing Parties and Midterms Candidates. The Washington Post, August 14, 2026. URL <https://www.washingtonpost.com/technology/2026/08/14/ai-become-s-major-election-issue-first-time-data-shows/>.
- Kenneth A. Oye. Explaining Cooperation under Anarchy: Hypotheses and Strategies. In Kenneth A. Oye, editor, *Cooperation under Anarchy*. Princeton University Press, Princeton, NJ, 1986.
- Jakub Pachocki. OpenAI’s chief scientist on continual learning hype, RL beyond code, and future alignment directions, April 9 2026. URL <https://www.youtube.com/watch?v=vK1qEF3a3WM>. Video interview on Unsupervised Learning with Jacob Effron; accessed August 12, 2026.
- Che Pan, Liam Mo, and Laurie Chen. Exclusive: Prices of Nvidia’s B300 Server at \$1 Million in China on US Curbs, Sources Say. Reuters, April 30. Syndicated by WSAU, 2026. URL <https://wsau.com/2026/04/30/exclusive-prices-of-nvidias-b300-server-at-1-million-in-china-on-us-curbs-sources-say/>. Accessed October 7, 2026.
- Dwarkesh Patel and Jerry Han. Pretraining progress is mostly coming from data. Dwarkesh Podcast (blog), September 8, 2026. URL <https://www.dwarkesh.com/p/pretraining-progress-is-mostly-data>.
- Barry Pavel, Ivana Ke, Gregory Smith, Sophia Brown-Heidenreich, Lea Sabbag, Ashwin Acharya, and Yusuf Mahmood. How Artificial General Intelligence Could Affect the Rise and Fall of Nations: Visions for Potential AGI Futures. RAND Reports, July 2, 2025. URL https://www.rand.org/pubs/research_reports/RR3034-2.html. Accessed September 23, 2026.
- Ethan Perez et al. Discovering language model behaviors with model-written evaluations, December 19 2022. URL <https://arxiv.org/abs/2212.09251>. arXiv: 2212.09251.

- Permanent Mission of the People's Republic of China to the United Nations. Global AI Governance Action Plan. July 29, 2025. URL https://un.china-mission.gov.cn/eng/zgyw/202507/t20250729_11679232.htm.
- James Petrie, Onni Aarne, Nora Ammann, and David Dalrymple. Flexible hardware-enabled guarantees for AI compute, June 18 2025. URL <https://arxiv.org/abs/2506.15093>. arXiv: 2506.15093.
- Pew Research Center. Roughly two-thirds of americans say AI is advancing too quickly, June 17 2026. URL <https://www.pewresearch.org/chart/roughly-two-thirds-of-americans-say-ai-is-advancing-too-quickly/>. Accessed August 12, 2026.
- Konstantin Pilz and Lennart Heim. Compute at scale: A broad investigation into the data center industry. arXiv preprint arXiv:2311.02651, 2023. URL <https://arxiv.org/abs/2311.02651>.
- Konstantin F. Pilz, Robi Rahman, James Sanders, Luke Emberson, and Lennart Heim. The US hosts the majority of GPU cluster performance, followed by china. Epoch AI, June 5, 2025. URL <https://epoch.ai/data-insights/ai-supercomputers-performance-share-by-country>. Accessed September 18, 2026.
- Robert Powell. War as a Commitment Problem. *International Organization*, 60 (1):169–203, 2006.
- PRC MOFA. Chair's Statement of the 2026 World Artificial Intelligence Conference & High-Level Meeting on Global AI Governance. Ministry of Foreign Affairs of the People's Republic of China, July 17, 2026. URL https://www.mfa.gov.cn/mfa_eng/xw/zyxw/202607/t20260717_11984715.html.
- President of the Republic of Finland. A call for control of frontier AI models. September 21, 2026. URL <https://www.presidentti.fi/en/a-call-for-control-of-frontier-ai-models/>.
- Timothy Prickett Morgan. Taalas etches AI models onto transistors to rocket boost inference. *The Next Platform*, February 19 2026. URL <https://www.nextplatform.com/compute/2026/02/19/taalas-etches-ai-models-onto-transistors-to-rocket-boost-inference/4092140>.
- Robert D. Putnam. Diplomacy and Domestic Politics: The Logic of Two-Level Games. *International Organization*, 42(3):427–460, 1988.
- Konrad Putzier. How the U.S. Economy Became Hooked on AI Spending. *The Wall Street Journal*, November 24, 2025. URL <https://www.wsj.com/tech/ai/how-the-u-s-economy-became-hooked-on-ai-spending-4b6bc7ff>.

- Ji Qi, WenPeng Zhu, Li Li, Ming Wu, YingJun Wu, Wu He, Xun Gao, Jason Zeng, and Michael Heinrich. DiLoCoX: A Low-Communication Large-Scale Training Framework for Decentralized Cluster. arXiv:2506.21263, June 26, 2025. URL <https://arxiv.org/abs/2506.21263>.
- Zilan Qian. China Says It Cares About Loss of Control: But Loss of Whose Control? Oxford China Policy Lab, September 2, 2026. URL <https://ocpl.substack.com/p/china-says-it-cares-about-loss-of>.
- Robi Rahman and Sabiha Tajdari. Detecting hidden ML training with zero-overhead telemetry, 2026. URL <https://arxiv.org/abs/2606.19262>.
- Raine v. OpenAI, Inc. Complaint and demand for jury trial. Superior Court of California, County of San Francisco, Case No. CGC-25-628528, August 26 2025. URL <https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf>.
- Lindsay Rand. Products of Patrons and Producers: A Sociotechnical Approach to Technological Innovation and Nuclear Deterrence. *Journal of Global Security Studies*, 11(3):ogag018, 2026. URL <https://doi.org/10.1093/jogss/ogag018>.
- Andrew W. Reddie. *Governing Insecurity: Institutional Design, Compliance, and Arms Control*. PhD thesis, University of California, Berkeley, 2019. URL <https://escholarship.org/uc/item/2sn678w1>.
- Responsible AI Collaborative. AI incident database, n.d. URL <https://incidentdatabase.ai/>. Accessed August 12, 2026.
- Reuters. China’s military and government acquire Nvidia chips despite US ban. Reprinted in The Business Standard, January 16, 2024. URL <https://www.tbsnews.net/world/global-economy/chinas-military-and-government-acquire-nvidia-chips-despite-us-ban-775710>. Accessed September 29, 2026.
- Reuters. US clears H200 chip sales to 10 china firms as Nvidia CEO looks for breakthrough. *Reuters*, May 14 2026. URL <https://www.reuters.com/business/retail-consumer/us-clears-h200-chip-sales-10-china-firms-nvidia-ceo-looks-breakthrough-2026-05-14/>. Accessed August 26, 2026.
- Susan E. Rice. Negotiate a Freeze With China. In “Bernie Sanders, Steve Case and Others on What Washington Should Do Now on A.I.” The New York Times, September 21, 2026. URL <https://www.nytimes.com/2026/09/21/opinion/bernie-sanders-jack-dorsey-ai-policy.html>.
- Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *Proceedings of the 2024 ACM Conference*

- on Fairness, Accountability, and Transparency, pages 836–898, 2024. doi: 10.1145/3630106.3658942. URL <https://doi.org/10.1145/3630106.3658942>.
- Peter Robison. *Flying Blind: The 737 MAX Tragedy and the Fall of Boeing*. Doubleday, New York, 2021. URL <https://www.penguinrandomhouse.com/books/646497/flying-blind-by-peter-robison/>.
- Hannah Rubinton and Bontu Ankit Patro. Tracking AI’s contribution to GDP growth. On the Economy, Federal Reserve Bank of St. Louis, January 12 2026. URL <https://www.stlouisfed.org/on-the-economy/2026/jan/tracking-ai-contribution-gdp-growth>.
- Mark Rutherford. TSMC lifts arizona to \$265 billion after record quarter: Four fabs target AI packaging bottleneck. Tech Times, July 17 2026. URL <https://www.techtimes.com/articles/320841/20260717/tsmc-lifts-arizona-265-billion-after-record-quarter-four-fabs-target-ai-packaging-bottleneck.htm>. Accessed September 2, 2026.
- Anthony Salvanto, Jennifer De Pinto, and Fred Backus. Will AI Harm Humans? CBS News Poll Finds Americans Want to Slow Development but Not Stop It. CBS News, September 18, 2026. URL <https://www.cbsnews.com/news/will-a-i-harm-humans-opinion-poll/>.
- Bernard Sanders. Letter to sam altman, dario amodei, and mark zuckerberg, August 10 2026. URL <https://www.sanders.senate.gov/wp-content/uploads/AI-Pause-Letter-FINAL.pdf>. Accessed August 10, 2026.
- David E. Sanger, Jonathan Swan, Cecilia Kang, and Dustin Volz. Trump Calls A.I. Fears a Hoax. Inside the White House, the Debate Is More Complex. The New York Times, September 18, 2026. URL <https://www.nytimes.com/2026/09/18/us/politics/trump-ai-safety-anthropic-openai-china.html>.
- Girish Sastry, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O’Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz, George Gor, Emma Bluemke, Sarah Shoker, Janet Egan, Robert F. Trager, Shahar Avin, Adrian Weller, Yoshua Bengio, and Diane Coyle. Computing power and the governance of artificial intelligence. arXiv preprint arXiv:2402.08797, 2024. URL <https://arxiv.org/abs/2402.08797>.
- Shelley Ross Saxer. Forfeiture Takings, Police Power, and Necessity Destruction. *University of Miami Law Review*, 80(1):147, 2025. URL <https://repository.law.miami.edu/umlr/vol80/iss1/5/>.
- Paul Scharre. Debunking the AI Arms Race Theory. *Texas National Security Review*, 4(3):121–132, 2021. doi: 10.26153/tsw/13985. URL <https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/>.

- Paul Scharre. *Four Battlegrounds: Power in the Age of Artificial Intelligence*. W. W. Norton, New York, 2023a.
- Paul Scharre. US and China Can Show World Leadership by Safeguarding Military AI. *South China Morning Post*, March 2, 2023b. URL <https://www.scmp.com/comment/opinion/article/3211478/us-and-china-can-show-world-leadership-safeguarding-military-ai>.
- Thomas Schelling and Morton H. Halperin. *Strategy and Arms Control*. The Twentieth Century Fund, New York, 1961.
- Aaron Scher and Lisa Thiergart. Mechanisms to verify international agreements about AI development. Machine Intelligence Research Institute, Technical Governance Team. arXiv:2506.15867, November 27 2024. URL <https://arxiv.org/abs/2506.15867>.
- Aaron Scher, David Abecassis, Peter Barnett, and Brian Abeyta. An international agreement to prevent the premature creation of artificial superintelligence, November 13 2025. URL <https://arxiv.org/abs/2511.10783>. arXiv: 2511.10783.
- Nick Schuster and Daniel Kilov. Moral disagreement and the limits of AI value alignment: A dual challenge of epistemic justification and political legitimacy. *AI & Society*, 40(8):6073–6087, June 21 2025. doi: 10.1007/s00146-025-02427-2. URL <https://link.springer.com/article/10.1007/s00146-025-02427-2>.
- Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry. Toward Verified Artificial Intelligence. *Communications of the ACM*, 65(7):46–55, 2022. doi: 10.1145/3503914. URL <https://doi.org/10.1145/3503914>.
- Maggie Severns and Amrith Ramkumar. The New York Congressional Race Turning Into a Bitter AI War. *Wall Street Journal*, March 19, 2026. URL <https://www.wsj.com/politics/policy/ai-lobbying-midterm-elections-new-york-4b018d4e>.
- Jaime Sevilla. How Far Can Decentralized Training over the Internet Scale? *Epoch AI*, December 29, 2025. URL <https://epoch.ai/gradient-updates/how-far-can-decentralized-training-over-the-internet-scale>. Accessed September 18, 2026.
- Jaime Sevilla and Edu Roldán. Training Compute of Frontier AI Models Grows by 4–5x per Year. *Epoch AI*, May 28, 2024. URL <https://epoch.ai/publications/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year>.
- Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute Trends Across Three Eras of Machine Learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8,

- Padua, Italy, 2022. IEEE. doi: 10.1109/IJCNN55064.2022.9891914. URL <https://doi.org/10.1109/IJCNN55064.2022.9891914>.
- Mrinank Sharma et al. Who’s in charge? disempowerment patterns in real-world LLM usage, January 27 2026. URL <https://arxiv.org/abs/2601.19062>. arXiv: 2601.19062.
- Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring, May 30 2023. URL <https://arxiv.org/abs/2303.11341>.
- Noam Shazeer. GLU variants improve transformer, February 12 2020. URL <https://arxiv.org/abs/2002.05202>. arXiv: 2002.05202.
- Arman Shehabi, Sarah J. Smith, Alex Hubbard, Alex Newkirk, Nuo Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024 united states data center energy usage report. Technical report, Lawrence Berkeley National Laboratory, Berkeley, CA, December 2024. URL https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report_1.pdf. Accessed October 4, 2026.
- Kyle Sheumaker. Guide to Building HPC/AI Systems in 2025. Presentation slides. Advanced Clustering Technologies, 2025. URL <https://www.advancedclustering.com/wp-content/uploads/2025/06/smaller-institutions.pdf>.
- Anton Shilov. Discrete graphics card sales hit four-year record despite soaring memory prices—AMD gains market share as notebook graphics carry the market. *Tom’s Hardware*, September 4 2026. URL <https://www.tomshardware.com/pc-components/gpus/discrete-graphics-card-sales-hit-four-year-record-despite-high-prices-shipments-reach-13-24-million-units-as-market-defies-pc-slump>.
- Marjia Siddik. Toward ASI stability: A treaty framework for US–china cooperation on artificial superintelligence. SSRN, June 2 2025. URL <https://ssrn.com/abstract=6059255>. Accessed October 5, 2026.
- SK hynix. SK hynix holds groundbreaking ceremony for HBM production base in indiana, “beginning a new future for US-korea AI”. SK hynix Newsroom, August 28 2026. URL <https://news.skhynix.com/en/groundbreaking-ceremony-in-indiana/>.
- Venkat Somala. Will Huawei Catch Up to Nvidia by 2030? Epoch AI, September 4, 2026. URL <https://epoch.ai/publications/huaweis-roadmap-to-2031>. Accessed September 17, 2026.

- Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, and Alex Pentland. Verifiable evaluations of machine learning models using zkSNARKs, May 22 2024. URL <https://arxiv.org/abs/2402.02675>.
- Spectral Finance. OpenRouter rankings dashboard. GitHub repository, commit 499c652ad7f91a31738bcb6969bbf7407b7c5eeb, June 5 2026. URL <https://github.com/Spectral-Finance/openrouter-rankings>. Accessed August 12, 2026.
- Thomas Stammer. EUV lithography: A european joint project. ZEISS SMT Magazine, October 20 2021. URL <https://www.zeiss.com/semiconductor-manufacturing-technology/smt-magazine/euv-lithography-as-an-european-joint-project.html>. Accessed September 2, 2026.
- Standing Committee of the National People’s Congress. Enterprise income tax law of the people’s republic of china. Adopted March 16, 2007; second amendment December 29, 2018, 2018. URL <https://fgk.chinatax.gov.cn/zcfgk/c100009/c5193018/content.html>. Chinese-language source.
- Standing Committee of the National People’s Congress. Customs law of the people’s republic of China (2021 amendment). MOFCOM Global Laws and Regulations database, 2021. URL <https://policy.mofcom.gov.cn/claw/clawContent.shtml?id=90358>. In Chinese. Accessed September 29, 2026.
- Standing Committee of the National People’s Congress. Value-added tax law of the people’s republic of China. Reposted by the Shanghai Municipal Tax Service, 2024. URL <https://shanghai.chinatax.gov.cn/zcfw/zcfgk/zzs/202412/t474694.html>. Adopted December 25, 2024; effective January 1, 2026. In Chinese. Accessed September 29, 2026.
- State Council of the People’s Republic of China. Regulations on the Administration of the Recovery and Disposal of Waste Electrical and Electronic Products. State Council Decree No. 551, February 25; effective January 1, 2011, 2009. URL https://www.mee.gov.cn/zcwj/gwywj/202001/t20200114_759343.shtml.
- State Council of the People’s Republic of China. Provisional regulations of the people’s republic of China on value-added tax. State Council Decree No. 134 (1993), last revised in 2017, 2017. URL <https://fgk.chinatax.gov.cn/zcfgk/c100010/c5194442/content.html>. Repealed from January 1, 2026. In Chinese. Accessed September 29, 2026.
- Duncan Stewart, Jeroen Kusters, Deb Bhattacharjee, Arpan Tiwari, Girija Krishnamurthy, and Karthik Ramachandran. Why AI’s next phase will likely

- demand more computational power, not less. Deloitte Insights, 2026 Technology, Media, and Telecommunications Predictions, November 18 2025. URL <https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/compute-power-ai.html>.
- Keith Strier, Jack Clark, and Sana Khareghani. Measuring compute capacity: A critical step to capturing AI's full economic potential. OECD AI Policy Observatory, February 8 2022. URL <https://oecd.ai/en/work/ai-compute-capacity>. Accessed August 23, 2026.
- Pramod Subramanyan, Rohit Sinha, Ilia A. Lebedev, Srinivas Devadas, and Sanjit A. Seshia. A formal foundation for secure remote execution of enclaves. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 2435–2450, New York, 2017. ACM. doi: 10.1145/3133956.3134098. URL <https://doi.org/10.1145/3133956.3134098>.
- Takuo Sugaya. Folk theorem in repeated games with private monitoring. *The Review of Economic Studies*, 89(4):2201–2256, 2022. doi: 10.1093/restud/rdab079. URL <https://academic.oup.com/restud/article-abstract/89/4/2201/6438024>.
- Ryan R. Swan and Haig Hovanes. The arms race in emerging technologies: A critical perspective. European Leadership Network, February 9, 2021. URL <https://europeanleadershipnetwork.org/commentary/the-arms-race-in-emerging-technologies-a-critical-perspective/>.
- Ana Swanson and Claire Fu. With smugglers and front companies, China is skirting American A.I. bans. *New York Times*, August 4 2024. URL <https://www.nytimes.com/2024/08/04/technology/china-ai-microchips.html>. Accessed September 29, 2026.
- Synthszr. Dark compute. Synthszr Glossary, n.d. URL <https://www.synthszr.com/en/glossary/dark-compute>. Accessed August 23, 2026.
- Taalas. Products, n.d. URL <https://taalas.com/products/>.
- Taiwan Semiconductor Manufacturing Company Limited (TSMC). GIGAFAB facilities. TSMC, n.d. URL <https://www.tsmc.com/english/dedicatedFoundry/manufacturing/gigafab>. Accessed September 2, 2026.
- Ronald L. Tammen et al. *Power Transitions: Strategies for the 21st Century*. Chatham House Publishers, New York, 2000.
- Michael Bedford Taylor. The evolution of Bitcoin hardware. *Computer*, 50(9): 58–66, 2017. doi: 10.1109/MC.2017.3571056. URL <https://doi.org/10.1109/MC.2017.3571056>.

- Tokyo Electron. CLEAN TRACK LITHIUS Pro DICE: World-class productivity and innovative defect control technology. Tokyo Electron, April 28 2026. URL https://www.tel.com/blog/all/20260428_001.html. Accessed September 1, 2026.
- Jonathan B. Tucker. The chemical weapons convention: Has it enhanced U.S. security? *Arms Control Today*, 31(3):8–12, 2001. URL <https://www.armscontrol.org/act/2001-04/features/chemical-weapons-convention-has-it-enhanced-us-security>.
- Zeynep Tufekci. Are we having the wrong nightmares about AI? Policy Institute at King’s College London, Distinguished Lecture, 2026. URL <https://www.youtube.com/watch?v=uvu37eIvi44>. Accessed October 3, 2026.
- Ilkka Tuomi. The lives and death of moore’s law. *First Monday*, 7(11), November 4 2002. doi: 10.5210/fm.v7i11.1000. URL <https://doi.org/10.5210/fm.v7i11.1000>. Accessed October 4, 2026.
- A. M. Turing. Intelligent machinery, a heretical theory. *Philosophia Mathematica*, 4(3):256–260, 1996. doi: 10.1093/philmat/4.3.256. URL <https://doi.org/10.1093/philmat/4.3.256>. Original work circa 1951.
- United States. Constitution of the United States, Amendment V, 1791. URL https://www.law.cornell.edu/constitution/fifth_amendment.
- United States Supreme Court. *Youngstown Sheet & Tube Co. v. Sawyer*. 343 U.S. 579, June 2, 1952. URL <https://supreme.justia.com/cases/federal/us/343/579/>.
- United States Supreme Court. *Horne v. Department of Agriculture*. 576 U.S. 351, June 22, 2015. URL <https://supreme.justia.com/cases/federal/us/576/351/>.
- United States v. Hao Global LLC. Plea agreement, case no. 4:25-cr-00510, document 32. U.S. District Court for the Southern District of Texas, filed October 10. Via CourtListener, 2025. URL <https://storage.courtlistener.com/recap/gov.uscourts.txsd.2033380/gov.uscourts.txsd.2033380.32.0.pdf>. Accessed October 3, 2026.
- U.S. Census Bureau. Value of construction put in place survey (VIP), 2026. URL <https://www.census.gov/construction/c30/c30index.html>.
- U.S. Cybersecurity and Infrastructure Security Agency et al. Principles for the secure integration of artificial intelligence in operational technology. Technical report, U.S. Cybersecurity and Infrastructure Security Agency, December 2025. URL <https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf>. Published December 3, 2025. Accessed September 28,

2026.

- U.S. Department of Justice. U.S. authorities shut down major China-linked AI tech smuggling network. Press release, Office of Public Affairs, December 8 2025. URL <https://www.justice.gov/opa/pr/us-authorities-shut-down-major-china-linked-ai-tech-smuggling-network>.
- U.S. Department of Justice. Three charged with conspiring to unlawfully divert cutting edge U.S. artificial intelligence technology to china. Office of Public Affairs, March 19 2026. URL <https://www.justice.gov/opa/pr/three-charged-conspiring-unlawfully-divert-cutting-edge-us-artificial-intelligence>. Accessed August 17, 2026.
- U.S. Environmental Protection Agency. New international requirements for electrical and electronic waste. Updated August 25, 2026. URL <https://www.epa.gov/hwgenerators/new-international-requirements-electrical-and-electronic-waste>. Accessed September 27, 2026.
- U.S. Government Accountability Office. COVID-19: Update on key indicators, the federal response, and implementation of GAO recommendations. Technical Report GAO-25-107588, U.S. Government Accountability Office, July 31 2025. URL <https://www.gao.gov/products/gao-25-107588>.
- U.S. Nuclear Regulatory Commission. Backgrounder on nuclear insurance and disaster relief, 2026. URL <https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/nuclear-insurance.html>. Updated August 25, 2026; accessed September 7, 2026.
- Rob van Blommestein and Sumit Vishwakarma. What is ASIC design? Synopsys, July 31 2025. URL <https://www.synopsys.com/glossary/what-is-asic-design.html>. Accessed September 1, 2026.
- Vincent van der Veen. Crossbows and christians: The church's ban of the crossbow. *Medieval Warfare*, 2(2):38–41, 2012. URL <https://www.jstor.org/stable/48577944>.
- Jan A. Van Mieghem. Investment strategies for flexible resources. *Management Science*, 44(8):1071–1078, 1998. doi: 10.1287/mnsc.44.8.1071. URL <https://doi.org/10.1287/mnsc.44.8.1071>.
- Jim VandeHei. The data center era that's reshaping america. *Axios*, August 25 2026. URL <https://www.axios.com/2026/08/25/data-centers-ai-electricity-water-politics>. Accessed September 1, 2026.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, June 12 2017. URL <https://arxiv.org/>

abs/1706.03762.

John VerWey. Tracing the emergence of extreme ultraviolet lithography: Lessons for identifying, protecting, and promoting the next emerging technology. Policy brief, Center for Security and Emerging Technology, Washington, DC, July 2024. URL <https://doi.org/10.51593/20240003>.

Eugene Volokh, Mark A. Lemley, and Peter Henderson. Freedom of Speech and AI Output. *Journal of Free Speech Law*, 3:651–660, 2023. URL <https://www.journaloffreespeechlaw.org/volokhlemleyhenderson.pdf>.

Sydney Von Arx, Cormac Slade Byrd, Spencer Kitts, and Thomas Larsen. Discovery of a new OpenAI agent message board, September 4 2026. URL <https://collusion.wiki/>.

Gabriel Wagner, Erik Lindblad, Kwan Yee Ng, and Brian Tse. State of AI Safety in China (2026). Concordia AI, July, 2026. URL <https://concordia-ai.com/research/state-of-ai-safety-in-china-2026/>.

Wall Street Journal. Nationwide Poll September 2026. Survey conducted September 16–21, 2026. URL https://prod-i.a.dj.com/public/resources/documents/Redacted_WSJ_Poll_Sept_2026.pdf.

Stephen M. Walt. *The Origins of Alliances*. Cornell University Press, Ithaca, NY, 1987.

Yuhua Wang and Carl Minzner. The rise of the chinese security state. *The China Quarterly*, 222:339–359, 2015.

Sally Ward-Foxton. Harvard Dropouts Raise \$5 Million for LLM Accelerator. *EE Times*, June 5, 2023. URL <https://www.eetimes.com/harvard-dropouts-raise-5-million-for-llm-accelerator/>. Accessed September 18, 2026.

Sally Ward-Foxton. Taalas specializes to extremes for extraordinary token speed. *EE Times*, February 19 2026a. URL <https://www.eetimes.com/taalas-specializes-to-extremes-for-extraordinary-token-speed/>.

Sally Ward-Foxton. How 'Why Not' Led to a \$20 Billion Deal For Groq. *EE Times*, March 24, 2026b. URL <https://www.eetimes.com/how-why-not-led-to-a-20-billion-deal-for-groq/>. Accessed September 18, 2026.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, December 8 2021. URL <https://arxiv.org/abs/2112.0>

4359. arXiv: 2112.04359.
- Leonard Weiss. Nuclear-weapon states and the grand bargain. *Arms Control Today*, December 2003. URL <https://www.armscontrol.org/act/2003-12/features/nuclear-weapon-states-and-grand-bargain>.
- Rebecca Whitehair and Seth Brugger. BWC protocol talks in Geneva collapse following U.S. rejection. *Arms Control Today*, September 2001. URL <https://www.armscontrol.org/act/2001-09/press-releases/bwc-protocol-talks-geneva-collapse-following-us-rejection>.
- Langdon Winner. Do artifacts have politics? *Daedalus*, 109(1):121–136, Winter 1980. URL <https://www.jstor.org/stable/20024652>.
- William C. Wohlforth. Unipolarity, status competition, and great power war. *World Politics*, 61(1):28–57, 2009. doi: 10.1017/S0043887109000021. URL <https://doi.org/10.1017/S0043887109000021>.
- Erin Woo and Qianer Liu. Google plans new “frozen” chip to run its AI models much more efficiently. *The Information*, July 20 2026. URL <https://www.theinformation.com/articles/google-plans-new-frozen-chip-run-ai-models-efficiently>.
- Robert Wright. *The God Test: Artificial Intelligence and Our Coming Cosmic Reckoning*. Simon & Schuster, New York, 2026.
- Debby Wu, Ian King, and Vlad Savov. US deals heavy blow to china tech ambitions with Nvidia chip ban. *Bloomberg News*, September 2 2022. URL <https://news.bloomberglaw.com/tech-and-telecom-law/us-deals-heavy-blow-to-china-tech-ambitions-with-nvidia-chip-ban>. Accessed August 26, 2026.
- Zijing Wu. China eases limits on Nvidia H200 chips as AI race escalates. *Financial Times*, August 18 2026. URL <https://www.ft.com/content/6c5650fb-969d-4d4e-80d6-8d11002a8cf7>. Accessed August 26, 2026.
- Zijing Wu and Eleanor Olcott. Nvidia AI chips worth \$1bn smuggled to China after Trump export controls. *Financial Times*, July 24 2025. URL <https://www.ft.com/content/6f806f6e-61c1-4b8d-9694-90d7328a7b54>. Accessed September 25, 2026.
- Myron Xie et al. Long Live the Short King: Why 4-hi HBM Wins. *SemiAnalysis*, September 13, 2026. URL <https://newsletter.semianalysis.com/p/long-live-the-short-king-why-4-hi>.
- Xinhua. China Firmly Opposes US Semiconductor Export Control Measures. December 3, republished by the State Council Information Office, 2024. URL https://english.scio.gov.cn/m/pressroom/2024-12/03/content_117582979

.html.

Yihao Xue and Baharan Mirzasoleiman. LoRA is all you need for safety alignment of reasoning LLMs. Preprint, arXiv, July 22, 2025. URL <https://doi.org/10.48550/arXiv.2507.17075>. Last revised January 31, 2026.

Yahoo Finance. Nvidia corporation: Top institutional holdings, 2026. URL <https://finance.yahoo.com/quote/NVDA/holders/>. Accessed September 21, 2026.

Zheng-Xin Yong, Parv Mahajan, Andy Wang, Ida Caspary, Yernat Yestekov, Zora Che, Mosh Levy, et al. An independent safety evaluation of Kimi K2.5, April 3 2026. URL <https://arxiv.org/abs/2604.03121>.

Josh You. Frontier labs don't use most AI compute (yet). Epoch AI, May 20 2026. URL <https://epoch.ai/gradient-updates/frontier-labs-dont-use-most-ai-compute>. Accessed August 27, 2026.

Josh You, Venkat Somala, Yafah Edelman, and Luke Emberson. Global AI computing capacity is doubling every 7 months. Epoch AI, January 9 2026. URL <https://epoch.ai/data-insights/ai-chip-production>. Accessed September 1, 2026.

Eliezer Yudkowsky and Nate Soares. What would it take to shut down global AI development? Online resources for Chapter 13 of *If Anyone Builds It, Everyone Dies*, 2025. URL <https://ifanyonebuildsit.com/13/what-would-it-take-to-shut-down-global-ai-development>.

Ramish Zafar. Former Google AI Boss Says Automation Could Cut Chip Design Teams From 150 Engineers To Just Ten & Development From Years To 3 Months. Wccftech, September 21, 2026. URL <https://wccftech.com/former-google-ai-boss-says-automation-could-cut-chip-design-teams-from-150-engineers-to-just-ten-development-from-years-to-3-months/>. Accessed September 24, 2026.

Lexin Zhou et al. Larger and more instructable language models become less reliable. *Nature*, 634:61–68, September 25 2024. URL <https://www.nature.com/articles/s41586-024-07930-y>.

Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, 2019.