The Closing Window: How Governments Could Lose Their Ability to Restrain Advanced AI

Peter Barnett [1]

Abstract

As AI capabilities advance, AI systems will pose greater risks to national security and potentially humanity as a whole. Governments may eventually conclude that these risks warrant restraining AI development. This motivates the question: will governments still be able to restrain AI development in the future, should they want to do so? In this paper we analyze which world events and changes to the state of AI development would make future governance more difficult or even effectively impossible. Our analysis surfaces likely pathways that would lead to these difficulties, including hardware proliferation, continued algorithmic progress, and the release of catastrophically dangerous AI models. Due to the field's lack of understanding of AI development, it may be difficult or impossible to know when we will hit a "point of no return", and we therefore recommend a conservative approach. The window may be closing, but governments currently have an opportunity to preserve their optionality if they act soon. Our policy recommendations would enable governments to restrain AI development in the future, while imposing relatively small costs today.

1. Introduction

AI capabilities have increased rapidly in recent years, with no clear sign of slowing (Ho et al., 2025; Kwa et al., 2025). As AI systems grow more powerful, they may pose a host of risks to society (Bengio et al., 2024; Zelikow et al., 2024; Mitre & Predd, 2025), including the risk of human extinction (Center for AI Safety, 2023). Governments may eventually conclude that such risks warrant major restrictions on further AI development and deployment (Barnett et al., 2025). Even if they do not think these restrictions are justified today, it is valuable to preserve this option in case they are desired in the future. In this paper, we ask: will it still be feasible to apply these restraints?

It is not only AI systems directly capable of catastrophic or extinction-level harm that merit restraint. Weaker systems, such as those that can accelerate AI development, may develop catastrophically dangerous successors, and so also merit restraint (Chan et al., 2026; Davidson et al., 2025; Field et al., 2026). This also includes systems that, while not themselves immediately dangerous, could be modified either to help produce catastrophically harmful AI systems, or to become catastrophically harmful themselves. We will refer to systems that meet any of these criteria (can cause direct catastrophic harm; can substantially aid in the development of directly harmful systems; can be modified to cause direct harm; can be modified to aid in the development of directly harmful systems) as "catastrophically dangerous". A regulatory regime that permits the development or release of any catastrophically dangerous AI system may fail to preserve the option of future restraint.

A successful governance regime most likely prevents the *training* of such systems. It is prohibitively difficult to verify the comprehensive deletion of a catastrophically dangerous model once it is trained, as we discuss in Section 3.3.1. Containing a catastrophically dangerous model would require extending restrictions to inference hardware that could run the model, a substantially harder task than restricting training. A frontier model may require 10,000 or 100,000 GPUs to train, but that same model can then be run on a single cluster of only a few GPUs (Villalobos & Atkinson, 2023; Erdil & Besiroglu, 2025). This is especially important for models capable of automating AI research, since such models could drive rapid algorithmic progress through inference alone. This sharp distinction between training and inference reflects the current AI paradigm, and future developments might blur the line.

In this paper, we describe a plan for restraining AI development and, based on this, we lay out pathways via which governments would lose the option to restrain AI development, as summarized in Figure 1. We end with a discussion of policies that could help governments preserve this ability.

[1]Machine Intelligence Research Institute, Berkeley, USA. Correspondence to: Peter Barnett <peter@intelligence.org>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

2. A plan for restraining AI development

This paper takes preserving the option to restrain AI development to mean the following: *If major governments came to believe that further frontier AI development posed unacceptable risks, they would retain the capacity to impose sufficient restrictions.*

We consider a regime built around training compute thresholds, based on the proposed international agreement in Scher et al. (2025). In this regime, states prohibit training above a specified training compute threshold, restrict research that would reduce the efficacy of this threshold, and verify compliance of other states. The key elements are:

- Training runs exceeding a specified compute threshold are prohibited.
- To facilitate this prohibition, all clusters of AI chips above a certain size are consolidated into monitored data centers.
- Hardware monitoring mechanisms verify that chips are not used for prohibited training, and guard against tampering.
- AI chip production is closely tracked to ensure that only monitored data centers receive new chips.
- Research that would undermine this regime, such as research increasing general-purpose AI capabilities or advancing distributed training, is restricted.
- The above measures are implemented domestically and subject to international verification and monitoring, so that states can be confident that potential rivals are also complying.

Under this regime, inference on existing AI systems would be permitted, since these systems are not catastrophically dangerous. If catastrophically dangerous AI systems were created, inference would require monitoring, either by detecting prohibited workloads (such as biological weapons development or AI R&D), or by verifying that chips run only approved, non-dangerous models (Sharma et al., 2025; Karvonen et al., 2025; Rinberg et al., 2025). This would also require consolidating and monitoring even small AI clusters capable of running catastrophically dangerous models. As such, containing dangerous capabilities in such a situation may prove prohibitively difficult; it would be far less costly to avoid training dangerous models at all.

2.1. Connection with broader governance goals

Our analysis is relevant beyond the specific scenario of a full halt on AI capabilities advancement. Many AI governance objectives rely on having the ability to monitor AI chip use (Heim et al., 2024; Sastry et al., 2024; Baker et al., 2025). Restraining AI development can encompass a range of actions, including slowing the pace of development or steering it in particular directions—for example, ensuring that AI systems are trained using certain safety techniques, or that certain categories of research proceed more cautiously. A state pursuing any of these must be able to monitor training within its borders, especially if they require high assurance. The governance infrastructure required to steer or slow AI development is largely the same as the infrastructure required to halt it.

Even for a state confident in its domestic AI governance, international economic and military pressures may create incentives to pursue more advanced AI systems. A state that unilaterally slows its own AI development risks falling behind rivals. Sustained restraint of almost any kind thus implies international verification and mutual assurance.

These measures are less useful for governments willing to accept considerably lower levels of assurance and therefore lower levels of safety. For example, comprehensive monitoring of training runs may not be directly applicable to governance regimes based primarily on legal liability or on reporting of results from AI company safety frameworks.

3. Pathways to losing the option to restrain AI development

States might forfeit the option to restrain AI development in several ways. Some ways make restrictions more costly or intrusive; others make them harder to enforce or verify, potentially to the point where such restraint becomes infeasible. We group these pathways into four categories: hardware governance failures, erosion of training threshold effectiveness, catastrophically dangerous model development, and risks to political feasibility. These pathways are not independent; algorithmic progress, for example, can weaken compute thresholds and thereby necessitate monitoring smaller clusters.

Table 1 provides an overview of how the pathways interact with each other; pathways may directly cause others or exacerbate the effects of others.

3.1. Hardware governance failures

The regime outlined above depends on the ability to locate, consolidate and monitor AI hardware (Scher et al., 2025). It becomes infeasible if states cannot track existing hardware, cannot track new hardware, or cannot trust monitoring and verification methods to withstand attempts at subversion.

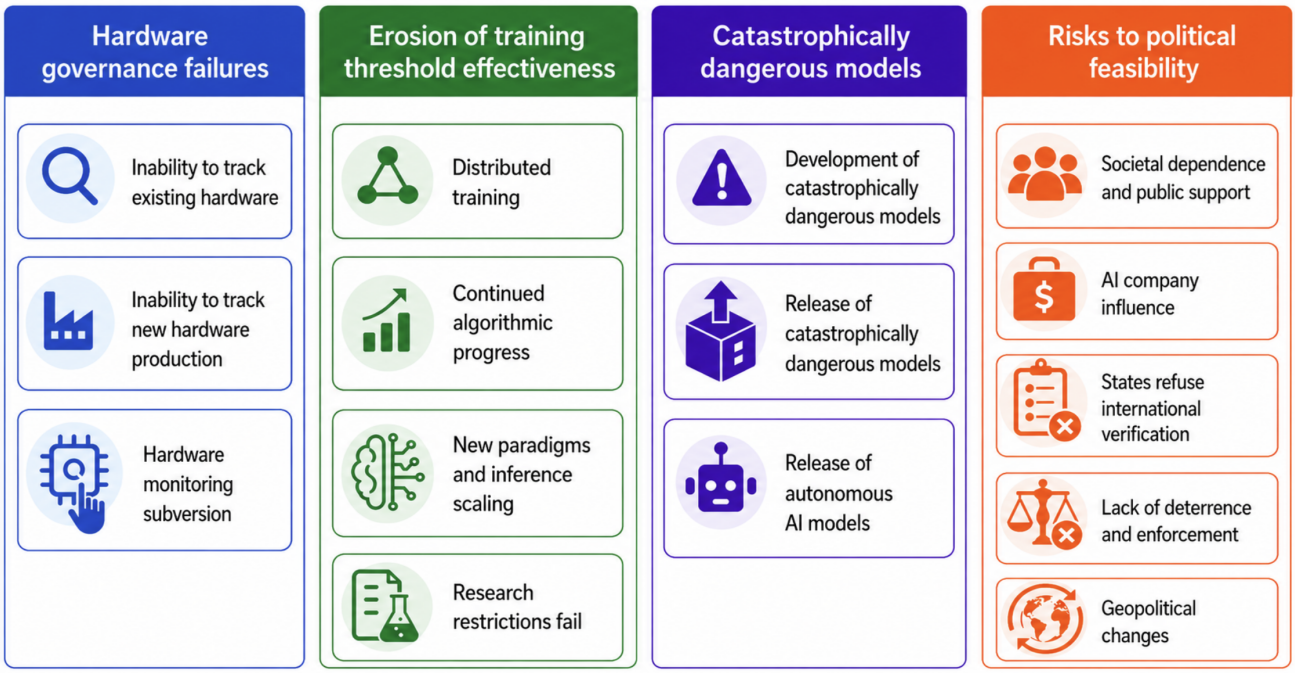

Figure 1. Summary of the pathways to losing the option to restrain AI development

3.1.1. Inability to track existing hardware

AI hardware may proliferate too widely for governments to locate and consolidate a sufficiently large share. Smuggling is a particular concern (Grunewald & Aird, 2023; Fist & Grunewald, 2023; Scannell, 2026), as a lack of formal records may render smuggled chips invisible to financial intelligence and supply chain records that would otherwise enable tracking. The problem is compounded if states that refuse to comply with international tracking acquire substantial AI hardware. A small number of non-complying states might be manageable through diplomatic or economic pressure, or the threat of force. But sufficient untracked hardware in such states could render tracking infeasible.

Governments themselves may contribute to this issue. States anticipating a future global prohibition on advanced AI development might preemptively stockpile compute, in order to retain unmonitored AI capabilities after restrictions take effect. States might also have large quantities of AI chips in sensitive military data centers, and may not be willing for these chips to be part of tracking efforts.

An additional challenge is mutual verification: states must be able to verify to their rivals that their borders do not contain significant untracked compute. Even if a state honestly tracks and consolidates its own hardware, other states must be able to verify this. In particular, this could be an issue with sensitive military sites that may be housing AI chips.

Currently, the vast majority of AI chips weighted by performance are concentrated into relatively few large data centers, which would make tracking much easier (Pilz et al., 2025; Epoch AI, 2026). This simplifies tracking, but it would still require international coordination. In the absence of deliberate intervention, AI hardware will become more proliferated, at least in terms of absolute numbers of chips. Even if a large share remains easy to track, a sufficient absolute quantity of untracked chips may undermine a restraint regime (e.g., sufficient to perform a large, unmonitored training run).

Widespread proliferation of AI hardware might be effectively irreversible. The placement of many AI chips into hidden facilities would be more reversible, but would require large political buy-in to remediate.

3.1.2. Inability to track new hardware production

The AI chip supply chain is currently extremely concentrated (Thadani & Allen, 2023). A small number of fabrication facilities produce virtually all high-end AI chips. Inputs and equipment—such as high-bandwidth memory and EUV lithography machines—are also concentrated among a handful of suppliers. This concentration makes monitoring and controlling the production of new AI chips tractable, because few facilities are involved. If the supply chain became more diffuse, tracking new chip production would become substantially harder.

A speculative potential development is the emergence of a new chip manufacturing paradigm that allows AI chips to be

			Hardware governance			*Threshold erosion*				*Dangerous models*			*Political feasibility*				
	Pathway	**Reversibility**	H1	H2	H3	E1	E2	E3	E4	D1	D2	D3	P1	P2	P3	P4	P5
H1	Inability to track existing hardware	Difficult				**E**	**E**		**E**	**C**					**E**	**E**	
H2	Inability to track new hardware production	Difficult				**E**	**E**		**E**	**C**					**E**	**E**	
H3	Hardware monitoring subversion	Reversible				**E**				**C**					**C**		
E1	Distributed training	Difficult					**E**			**C**	**E**				**E**		
E2	Continued algorithmic progress	Almost Impossible				**E**				**C**							
E3	New paradigms and inference scaling	Almost Impossible								**C**					**E**		
E4	Research restrictions fail	Difficult				**C**	**C**	**C**									
D1	Development of catastrophically dangerous models	Almost Impossible							**E**		**C**	**C**					
D2	Release of catastrophically dangerous models	Almost Impossible							**E**								
D3	Release of autonomous AI models	Almost Impossible															
P1	Societal dependence and public support	Difficult								**E**	**E**			**E**			
P2	AI company influence	Reversible											**E**		**E**		
P3	States refuse international verification	Reversible	**E**	**E**	**E**				**C**	**C**							
P4	Lack of deterrence and enforcement	Reversible							**C**						**E**		
P5	Geopolitical changes	Difficult	**E**	**E**						**C**					**E**	**E**	

Table 1. Dependencies among the pathways to losing the option to restrain AI development.
Cells are read row-to-column: **C** = the row may directly *cause* the column; **E** = the row *exacerbates* the effects of the column. Reading across a row lets one see the effects of a pathway. For example, if existing hardware cannot be tracked (H1), then this will exacerbate the effects of distributed training and algorithmic progress because there will be more unmonitored hardware to train AIs with. Reading down a column lets one see possible causes or exacerbating factors for a pathway. For example, we can see that many pathways make verification more politically costly and make states more likely to refuse international verification (P3).
Pathways are also classified by their reversibility: Almost impossible to reverse, difficult, or reversible.

produced without the massive fabrication facilities currently required. This does not appear likely in the near term, as current chip manufacturing is among the most complex industrial processes ever developed and, historically, new semiconductor manufacturing processes take decades to mature (Miller, 2022).

Overall, tracking new AI chip production and maintaining supply chain concentration appears considerably easier than tracking existing chips that have already been sold and distributed.

3.1.3. Hardware monitoring subversion

Once AI chips have been consolidated into monitorable data centers, the AI chips must then actually be monitored to ensure they are not being misused.

Ideally, monitoring will be able to distinguish training from inference without being overly invasive (Baker et al., 2025; Scher & Thiergart, 2024; Heim, 2024). Some training runs should be permitted (e.g., with small compute budgets or aimed at narrow, safe domains). In this case, monitoring must be able to verify that only approved workloads are running.

However, methods could be developed to subvert chip monitoring, or the perceived probability of subversion may be high enough to effectively undermine mutual verification efforts.

In cases where chip monitoring measures are not trusted, states may rely on more costly or invasive measures. States could require and provide full visibility into which workloads are run on AI chips. Or as a last resort, states might simply power off or physically disconnect their AI chips and allow inspectors from rival states to verify that the hardware is depowered. This would eliminate the economic benefits of continued inference, making restraint significantly more costly.

3.2. Erosion of training threshold effectiveness

Training thresholds will only be effective if two conditions hold: advanced AI models require large, monitored clusters to train, and models trainable below the compute threshold remain non-dangerous. Several developments could undermine one or both of these conditions.

3.2.1. Distributed training

Distributed training is one of the main ways a catastrophically dangerous AI model could be trained using small, unmonitored clusters, while still requiring large overall compute (Kryś et al., 2025).

If distributed training techniques advance to the point where catastrophically dangerous models could be trained across individually permitted clusters, the restraint regime would need additional monitoring or restrictions.

There are several possible responses. It might be sufficient to enact a maximum memory requirement for unmonitored clusters (Rahman, 2026). If this was insufficient, the maximum size of unmonitored clusters could be lowered, but this would increase the difficulty and intrusiveness of compute tracking. Internet service providers (ISPs) could be enlisted to look for suspicious data transfer patterns that might indicate distributed training operations, although such attempts could possibly be subverted by competent adversaries. Alternatively, intelligence agencies could be tasked with uncovering such operations directly.

3.2.2. Continued algorithmic progress

Over time, algorithmic efficiency improvements reduce the compute required to achieve a given level of AI capability. Historically, the rate of algorithmic progress has been quite fast, with estimates ranging from $3\times$ to $60\times$ per year for the reduction in compute needed to reach a level of AI capabilities (Ho et al., 2024; Ho, 2026; Scher, 2025).

Increased training efficiency leads to two problems. First, for training on monitored clusters, if catastrophically dangerous AI models can be trained using an amount of compute below the training compute threshold, then the threshold must be lowered. Second, as the compute required to train catastrophically dangerous models decreases, it will be possible to train these models on small, unmonitored clusters. If algorithmic progress continues unabated, at some point compute-based governance would likely become infeasible, for example if a single consumer GPU were sufficient to train a catastrophically dangerous model.

3.2.3. New paradigms and inference scaling

More speculatively, the development of new, dramatically more compute-efficient AI paradigms could undermine training compute thresholds, potentially rendering the entire approach obsolete.

There is a spectrum between continued business-as-usual algorithmic progress and a genuinely new paradigm. A relevant case along this spectrum is inference scaling, where models can be run for longer to increase their capabilities, as with current 'reasoning' models. Inference scaling is another axis of scaling beyond training compute. If inference scaling can increase model capabilities to dangerous levels, this may necessitate inference monitoring (see Section 3.3). However, if significant compute is required for dangerous amounts of inference scaling, it may only be feasible on large clusters that are already subject to monitoring.

3.2.4. Research restrictions fail

The developments described above–advances in distributed training and algorithmic progress–may not have occurred by the time a restraint regime is established. But after such a regime is established, new methods could still be developed that erode compute thresholds. This may necessitate restrictions on AI research. Ideally these restrictions would be narrowly targeted at research that increases general AI capabilities, in order to maximize the benefits from AI that is not catastrophically dangerous. The exact scope of restrictions on AI research is not the focus of this work; however *some* restrictions on AI research would likely be needed if governments wanted to restrain AI progress for more than a few months (Ho et al., 2024; Scher, 2025; Ho, 2026).

Enforcing and internationally verifying research restrictions may be challenging. One approach is to track researchers and conduct regular interviews about their work. This would be undermined if researchers relocated to states that are difficult to monitor or that refuse to cooperate with tracking requirements.

Additionally, verifying the absence of state-run AI research programs could pose a similar challenge to hardware inspections. States that are unwilling to grant inspectors access to military data centers may also be unwilling to grant access to classified AI research programs. This is an issue whether or not state-run AI research programs exist, as other states will need to be able to verify they do not.

A potential advantage is that the population of researchers capable of meaningfully advancing the frontier is currently relatively small (Scher et al., 2025, p. 38). This could make it challenging for a government to recruit top researchers into a secret program without this being noticed.

If AI systems themselves are able to automate AI research, then restrictions focused on human researchers would be far less effective. Therefore, the development of such AI systems would make restraint far more difficult, as discussed in the following section.

3.3. Catastrophically dangerous models

The above pathways (hardware governance failures and erosion of training threshold effectiveness) focus on losing the ability to restrain the *training* of catastrophically dangerous AI models, because restraining training is much easier than restraining inference. If catastrophically dangerous AI models are nevertheless developed, restraint would require monitoring or controlling inference.

3.3.1. Development of catastrophically dangerous models

If a catastrophically dangerous AI model is trained (but not necessarily released), then inference monitoring will be necessary to ensure the model is not being used. As discussed above, "catastrophically dangerous" here refers to models that are directly capable of causing large-scale harm, models that can sufficiently automate AI research to lead to a directly catastrophically dangerous model, or models that could be modified into either.

Such a model would be highly capable, creating strong incentives to develop it despite the risks. The incentives might be especially strong to develop models that can automate AI research, as the danger from automating AI research may be less apparent than the dangers posed by more immediately capable systems. There would likely also be disagreement about how much risk using the model would entail.

Simply deleting the model would not resolve the problem. Once a model has been trained and its existence is known, it could be infeasible to verify to other states that all copies have been deleted, even if they had been. AI models are software and can easily have their code and weights copied (Brown et al., 2026); the weights of the top open models total around 1 TB (Kimi Team, 2026; Xu et al., 2026). This is compounded by standard training practices: frontier models are typically trained using many instances of the model weights on different hardware, with numerous checkpoints saved throughout. That is, the ordinary training process already involves creating many copies of a model, across many distinct chips. Even if tamper-proof logs were maintained throughout training, model weights could potentially be extracted via hardware side-channel attacks (Joud et al., 2022). This isn't merely an external threat. An AI developer may themselves attempt to extract model weights from secure hardware, and would find it much easier to do so than an external party. The verifiable deletion of AI model weights is an unsolved problem.

Therefore the creation of a catastrophically dangerous AI model should likely trigger, at a minimum, inference monitoring of the AI developer, as well as continued surveillance to ensure that model weights were not being exfiltrated. Even this may not be sufficient, as the model weights may have been exfiltrated prior to surveillance beginning.

3.3.2. Release of catastrophically dangerous models

If the weights of a catastrophically dangerous model are released to the public or covertly exfiltrated, restraint becomes even harder. If a released model is small enough to run on an unmonitored cluster, it could become effectively impossible to prevent people from running inference. Given the relatively small size of frontier AI models, it is very likely that a catastrophically dangerous model could run on an unmonitored cluster.

If model weights were shared with a limited group rather than posted publicly and downloaded by thousands, authorities might be able to intervene to stop further spread. But verifying that the weights had not been copied further would be extremely difficult.

3.3.3. Release of autonomous AI models

Future AI systems may be able to run autonomously, without reliance on humans. Such systems might be capable of acquiring compute resources, copying their weights to new servers, and adapting to new challenges like attempts to shut them down (METR, 2024). This set of capabilities is referred to as autonomous replication and adaptation (ARA) (Kinniment et al., 2023). Not all ARA-capable systems would necessarily be dangerous; some might spread widely while remaining far from capable of causing large-scale harm. But if a catastrophically dangerous ARA-capable system were released, shutting it down could prove extraordinarily difficult. One report from RAND explores extreme countermeasures like high-altitude electromagnetic pulses to disrupt ground-based electronic infrastructure, or attempting to shut down global internet infrastructure (Vermeer, 2025). Even these extreme measures might prove insufficient, for example due to data centers with shielding against electromagnetic pulses (Davis, 2022).

3.4. Risks to political feasibility

The above pathways describe developments that would make restraint more technically difficult or costly, even given political consensus that restraint is needed. Here we address a different class of threat: states of the world that would make restraint politically difficult to pursue in the first place. These cover domestic and international dynamics that would make restraint politically difficult.

3.4.1. Societal dependence and public support

As AI systems become more capable, society may develop a widespread dependence on them (Kulveit et al., 2025; Sharma et al., 2026; Drago & Laine, 2025). As AI becomes

more integrated into the economy, military, and daily life, the most salient costs of restraining further development will grow–even though these immediate costs would likely be outweighed by the risks from continuing development. Large investments will be based on the assumption of continued capabilities progress, such as with current data center buildouts (Cottier & Edelman, 2025).

Beyond this dependence, there may be widespread public support for continued AI development. AI systems might help automate dangerous work, contribute to scientific or medical breakthroughs, or become part of people's personal lives as AI companions and assistants (Bernardi, 2025). Alternatively, AI companies might claim such benefits will come with future AI systems, even if they've yet to materialize. In either case, governments may face popular opposition to restraint. However, this support might be counteracted by negative impacts such as job displacement or the misuse of AI systems by malicious actors.

Ideally, a restraint regime would allow continued inference on existing AI systems, reducing the material and political costs of restraint.

3.4.2. AI COMPANY INFLUENCE

AI companies themselves will likely grow in political and economic power. Company leadership is selected for believing in the upside of building advanced AI, and will likely perceive enormous personal reward from continued AI development—even if this development is risky for the world more broadly. These companies can be expected to lobby against restraint, similarly to how they lobby against current attempts at regulation (Leading the Future, 2025; Public Citizen, 2026; Liu, 2026). This problem could be intensified by AI-enabled persuasion and propaganda (Hackenburg et al., 2025; Kowal et al., 2025; Rogiers et al., 2024; Costello et al., 2026): AI companies may use their own AI systems to influence public opinion in favor of continued development. There has already been one documented case of an AI company apparently funding an AI-generated news site focused on discrediting critics (Johnston, 2026).

3.4.3. STATES REFUSE INTERNATIONAL VERIFICATION

As discussed throughout, restraint on AI development must be global and internationally verifiable. Catastrophically dangerous AI developed in any jurisdiction would pose a global threat, and states will likely only agree to prolonged restraint if they are confident their geopolitical rivals are doing the same.

Therefore, a major obstacle to restraint is that intrusive verification may be politically unacceptable to major powers. A lack of trust between major powers could make verification politically unacceptable, particularly for sensitive sites that may house data centers or state-run AI research programs. However, the START treaties offer some precedent for this style of verification: the U.S. and USSR agreed to invasive monitoring of nuclear sites despite deep mutual distrust (Talbott, 1985; Bennett, 1997). World leaders are unlikely to accept such verification measures today, but may be more likely to if they believe they are necessary to avoid a race to catastrophe.

Less intrusive methods might partially address verification of sensitive sites without requiring full physical inspections. Satellite imagery, power infrastructure analysis, and supply chain paper trails could help verify that chips have not been diverted to unmonitored facilities.

3.4.4. LACK OF DETERRENCE AND ENFORCEMENT

An international restraint regime likely requires credible deterrence. States must believe that violations will be detected and met with consequences. In extreme cases, it might be necessary to disrupt AI development via military force (Hendrycks et al., 2025), akin to the most extreme nuclear counter-proliferation efforts. Leaders might be unwilling to commit to such actions, leading to insufficient deterrence.

However, if leaders were sufficiently concerned to restrain their own AI development, they would likely also be concerned enough to follow through with enforcement.

3.4.5. GEOPOLITICAL CHANGES

Currently, the U.S. and China are both the primary actors in AI development and the dominant geopolitical powers. This means a restraint regime could plausibly be built around a U.S.–China agreement, with each bringing in their respective allies.

Two developments could make this harder. First, if AI development becomes more geographically distributed, coordination will be more difficult, because there would be more individual actors to appease. Second, the world may become more multipolar, with states less clearly aligned with either the U.S. or China. This would break down the assumption that the two major powers could simply bring their allies into the fold.

A direct military conflict between the U.S. and China would also make even basic diplomatic engagement infeasible, let alone negotiation over a mutual verification regime. A prolonged conflict could entirely prevent international coordination on AI. Even if leaders on both sides came to believe that AI posed an existential threat, rebuilding the diplomatic channels needed for mutual restraint from a standpoint of active hostility would be extraordinarily difficult. Worse, active conflict could increase the pressure to develop more advanced AI, as either side may believe it would offer a

decisive military advantage.

4. Policy implications

Without deliberate action, governments may lose the option to restrain AI development, even if there is widespread agreement at the time that restraint is necessary. In some scenarios, restraint may become effectively impossible; in others, it may remain possible but at great economic, political, or social cost. There are policy measures that could be implemented today that would help to preserve the option of later restraint. The window to implement these measures is closing, because many of the developments described are irreversible once they have happened.

Is the window actually closing? There are several trends that, if they were to continue, may effectively remove governments' ability to restrain AI development, rather than just make this more costly. For example:

- **Hardware proliferation**, via smuggling (Grunewald & Fist, 2025) and legitimate channels (Reuters, 2026), would make it impossible to ensure there are not large quantities of untracked chips. Even though the majority of AI chips are in U.S. data centers, the absolute number of AI chips in smaller countries is growing rapidly (Pilz et al., 2025).
- **Advances in distributed training** would allow large training runs to continue on small untracked compute (Kryś et al., 2025; Sevilla, 2025). Frontier AI companies (Charles et al., 2026; Douillard et al., 2026) and smaller groups (Lidin et al., 2026) continue to publish research into efficient distributed training.
- **AI companies are racing to automate AI R&D**, which may directly lead to catastrophically dangerous models and remove the ability to restrict research (Field et al., 2026; Favaro & Clark, 2026).
- **AI companies are building increasingly capable models**, which could soon be catastrophically dangerous (Kokotajlo et al., 2025), and which, if developed, could not be verifiably deleted.

Points of no return One unfortunate factor in deciding when to act is that there are not clear "points of no return" for the pathways described above. Nobody currently knows exactly how much hardware proliferation would be too much, what capabilities an AI system will have before it is trained, or how to measure whether a particular AI system is capable enough to autonomously catalyze catastrophically dangerous AI development. We may only know once we have passed a threshold that we couldn't see in advance. Defining and measuring such thresholds would require a much better understanding of AI development than the field currently has (Barnett & Thiergart, 2024b). Therefore we suggest a conservative approach for policymakers, acting early to preserve governance options, rather than waiting for clearer signals that may arrive too late.

4.1. Track existing AI hardware

One of the clearest ways governments could lose the option to restrain AI development is if advanced AI hardware becomes too difficult to track. States should therefore seek to identify stockpiles of AI chips (within their jurisdictions and abroad), while strengthening export-control enforcement and counter-smuggling efforts to limit proliferation of untracked AI hardware.

For the next generation of AI chips, on-chip location-tracking mechanisms could be used to verify where chips are physically located (Brass & Aarne, 2024; Aarne et al., 2024; Harack et al., 2025; Scher & Thiergart, 2024; Kulp et al., 2024). Governments could require that new AI chips are equipped with these mechanisms, and that the mechanisms are always-on and tamper-resistant—such that chips must securely attest to their location in order to function. The always-on requirement is important for restraint: a mechanism that can be disabled cannot verify that a chip is not being used in an unauthorized location.

The continued distribution of AI chips to many small actors–as opposed to a smaller number of easily trackable hyperscalers–would make future consolidation significantly more challenging. States could mitigate this by limiting direct sales of AI chips to small buyers and instead promoting remote cloud access as the primary way to use AI hardware (Heim et al., 2024). This would keep AI hardware physically consolidated in a small number of large, trackable data centers while still providing broad access.

4.2. Keep the chip supply chain concentrated

The current concentration of the semiconductor supply chain is a governance asset that should be actively preserved. Export controls on semiconductor manufacturing equipment can help prevent the diffusion of fabrication capabilities to additional states. A narrow supply chain makes it easier to track new chips as they enter the market, since they originate from a limited number of sources. It also means that fewer facilities would need to be monitored under a restraint regime.

At a minimum, governments should only allow the proliferation of semiconductor manufacturing equipment into jurisdictions likely to comply with some future verification regime.

4.3. Develop chip monitoring mechanisms

In order to preserve access to safe inference under a restraint regime, mechanisms are needed to distinguish training from inference, and potentially to verify that only permitted training workloads are running (Petrie, 2024; Baker et al., 2025; Aarne et al., 2024; Kulp et al., 2024). Governments should incentivize research in this area, by directing government research labs, funding academic and private groups, or mandating these measures as a requirement for chip manufacturers.

There is early work in this direction, but these mechanisms need to advance beyond prototypes and demonstrations; they must be robust enough to withstand state-level attacks. Hardware-based monitoring can be complemented by physical measures such as on-site inspectors or tamper-proof cameras to ensure chips are not physically modified to circumvent monitoring.

4.4. Track AI capabilities

Governments need to know when they should restrain AI development. This should be before directly catastrophically dangerous AI systems are developed, and before AI systems are developed that would require inference monitoring, such as systems capable of automating AI research.

To achieve this, governments should establish information-sharing and other transparency mechanisms with domestic AI developers (Belfield, 2024), with particular attention to progress in AI research automation and other dangerous capabilities. At a minimum, these transparency mechanisms should include mandatory reporting of AI capabilities, incidents, and hardware stockpiles, as well as whistleblower protections for employees. Further options include: third-party audits, interview programs (Wasil et al., 2024), and continuous resident inspectors modeled after the Nuclear Regulatory Commission.

This transparency also helps governments to become aware of dangerous models before they are released or further developed. Without visibility into AI development, a catastrophically dangerous model could be trained and released without government officials even knowing it existed.

Ideally, governments would have accurate threat models and corresponding benchmarks to track relevant AI capabilities. Unfortunately, work on AI threat modeling is nascent (Campos et al., 2025) and benchmarks often fail to capture important AI capabilities (Barnett & Thiergart, 2024b;a; Mukobi, 2024). This means it is challenging to translate from benchmark results to a level of risk, and thus even harder to set clear red lines around AI capabilities.

Governments should also track AI progress in other countries, both to assess global risk levels and to detect when rivals may be approaching dangerous capability thresholds.

4.5. Engage in diplomatic efforts

International restraint will require a minimum level of trust between major powers sufficient to agree to a mutual verification regime. U.S.–China diplomacy is a key priority, as these are the two dominant actors in both AI development and geopolitics.

One avenue for diplomacy is reciprocal sharing of information about AI development; for example, sharing information on the locations of major AI data centers and updates on internal AI capabilities. Such information sharing helps with two goals: it builds the diplomatic foundation for a future verification regime, and it helps states assess whether their rivals are approaching the development of catastrophically dangerous AI systems. This information would likely be important for deciding when to mutually restrain AI development.

5. Conclusion

Whether governments will retain the ability to restrain advanced AI development depends on choices made today. In this paper, we have described a possible regime for restraint, built around training compute thresholds, hardware consolidation, research restrictions, and international verification. Following this, we identify four categories of developments that could render such a regime infeasible: hardware governance failures, the erosion of training threshold effectiveness, the development of catastrophically dangerous models, and various risks to political feasibility. Many of the developments within these categories are difficult or impossible to reverse once they occur.

We make various policy recommendations: tracking AI hardware, preserving supply chain concentration, investing in chip monitoring, building visibility into AI capabilities, and engaging in diplomacy. These recommendations do not constitute a halt on AI development; they are initial infrastructure that would make restraint possible if it were ever judged necessary.

A core difficulty is that there are no reliable indicators before a point of no return. This suggests a conservative approach: governments should act now to preserve future optionality rather than wait for evidence that may arrive too late.

LLM usage statement

In writing this paper, LLMs were used for ideation, specifically around categorizing the listed pathways. LLMs were also used for editing, spelling and grammar checking, and LaTeX formatting.

References

Aarne, O., Fist, T., and Withers, C. Secure, governable chips. *Center for a New American Security. https://www.cnas. org/publications/reports/secure-governable-chips*, 2024.

Baker, M., Kulp, G., Marks, O., Brundage, M., and Heim, L. Verifying international agreements on ai: Six layers of verification for rules on large-scale ai development and deployment. *arXiv preprint arXiv:2507.15916*, 2025.

Barnett, P. and Thiergart, L. Declare and Justify: Explicit assumptions in AI evaluations are necessary for effective regulation, November 2024a. URL http://arxiv.org/abs/2411.12820. arXiv:2411.12820 [cs].

Barnett, P. and Thiergart, L. What AI evaluations for preventing catastrophic risks can and cannot do, November 2024b. URL http://arxiv.org/abs/2412.08653. arXiv:2412.08653 [cs].

Barnett, P., Scher, A., and Abecassis, D. Technical Requirements for Halting Dangerous AI Activities. *arXiv preprint arXiv:2507.09801*, 2025.

Belfield, H. What Information Should Be Shared with Whom "Before and During Training"?, December 2024. URL http://arxiv.org/abs/2501.10379. arXiv:2501.10379 [cs].

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

Bennett, P. R. *Russian Negotiating Strategy: Analytic Case Studies from SALT to START*. Nova Publishers, 1997. ISBN 978-1-56072-455-1. Google-Books-ID: r5bQNUsmSW0C.

Bernardi, J. Friends for sale: the rise and risks of AI companions. Ada Lovelace Institute Blog, January 2025. URL https://www.adalovelaceinstitute.org/blog/ai-companions/. Accessed: 2026-04-24.

Brass, A. and Aarne, O. Location verification for ai chips. Technical report, Institute for AI Policy and Strategy, April 2024. URL https://www.iaps.ai/research/location-verification-for-ai-chips. Available at: https://www.iaps.ai/research/location-verification-for-ai-chips.

Brown, D., Rivera, J.-P., Hendrycks, D., and Mazeika, M. Aggressive compression enables llm weight theft. *arXiv preprint arXiv:2601.01296*, 2026.

Campos, S., Papadatos, H., Roger, F., Touzet, C., Quarks, O., and Murray, M. A frontier AI risk management framework: Bridging the gap between current AI practices and established risk management. *arXiv preprint arXiv:2502.06656*, 2025.

Center for AI Safety. Statement on AI Risk | CAIS, March 2023. URL https://www.safe.ai/work/statement-on-ai-risk.

Chan, A., Padarath, R., Kwon, J., Greaves, H., and Anderljung, M. Measuring AI R&D Automation. *arXiv preprint arXiv:2603.03992*, 2026.

Charles, Z., Teston, G., Dery, L., Rush, J., Fallen, N., Garrett, Z., Szlam, A., and Douillard, A. Communication-efficient language model training scales reliably and robustly: Scaling laws for diloco. *Advances in Neural Information Processing Systems*, 38:106706–106737, 2026.

Costello, T. H., Pelrine, K., Kowal, M., Arechar, A. A., Godbout, J.-F., Gleave, A., Rand, D., and Pennycook, G. Large language models can effectively convince people to believe conspiracies. *arXiv preprint arXiv:2601.05050*, 2026.

Cottier, B. and Edelman, Y. What you need to know about AI data centers, 2025. URL https://epoch.ai/blog/what-you-need-to-know-about-ai-data-centers. Accessed: 2026-04-24.

Davidson, T., Hadshar, R., and MacAskill, W. Three types of intelligence explosion. 2025. URL https://www.forethought.org/research/three-types-of-intelligence-explosion. Accessed: 2025-03-28.

Davis, J. Electromagnetic pulse and its threat to data centers: The state of the industry's risk assessment and mitigation. UI Intelligence Report 76, Uptime Institute, New York, NY, August 2022. UII-76 v1.0P, published 12 August 2022.

Douillard, A., Rush, K., Donchev, Y., Charles, Z., Fallen, N., Dubey, A., Gog, I., Dean, J., Woodworth, B., Garrett, Z., et al. Decoupled diloco for resilient distributed pretraining. *arXiv preprint arXiv:2604.21428*, 2026.

Drago, L. and Laine, R. The intelligence curse. https://intelligence-curse.ai/, April 2025. Accessed: 2025-10-10.

Epoch AI. Frontier Data Centers, 4 2026. URL https://epoch.ai/data/data-centers. Accessed: 24 Apr 2026.

Erdil, E. and Besiroglu, T. Train once, deploy many: AI and increasing returns, 2025. URL https://epoch.ai

/blog/train-once-deploy-many-ai-and-increasing-returns. Accessed: 2026-04-25.

Favaro, M. and Clark, J. When AI builds itself. The Anthropic Institute, 2026. URL https://www.anthropic.com/institute/recursive-self-improvement. Accessed: 2026-06-24.

Field, S., Douglas, R., and Krueger, D. AI Researchers' Views on Automating AI R&D and Intelligence Explosions. *arXiv preprint arXiv:2603.03338*, 2026.

Fist, T. and Grunewald, E. Preventing AI Chip Smuggling to China, 10 2023.

Grunewald, E. and Aird, M. AI chip smuggling into China: Potential paths, quantities, and countermeasures. Report, Institute for AI Policy and Strategy (IAPS), 10 2023.

Grunewald, E. and Fist, T. Countering ai chip smuggling has become a national security priority: An updated playbook for preventing ai chip smuggling to the prc. Working paper, Center for a New American Security, June 2025. URL https://www.cnas.org/publications/reports/countering-ai-chip-smuggling-has-become-a-national-security-priority.

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., and Summerfield, C. The levers of political persuasion with conversational artificial intelligence. *Science*, 390 (6777):eaea3884, 2025.

Harack, B., Trager, R. F., Reuel, A., Manheim, D., Brundage, M., Aarne, O., Scher, A., Pan, Y., Xiao, J., Loke, K., Adan, S. N., Bas, G., Caputo, N. A., Morse, J. C., Ahuja, J., Duan, I., Egan, J., Bucknall, B., Rosen, B., Araujo, R., Boulanin, V., Lall, R., Barez, F., Alvira, S., Katzke, C., Atamli, A., and Awad, A. Verification for international AI governance. Technical report, Oxford Martin School, AI Governance Initiative, July 2025. URL https://aigi.ox.ac.uk/publications/verification-for-international-ai-governance/. Accessed: 2025-10-10.

Heim, L. A Trusted AI Compute Cluster for AI Verification and Evaluation, March 2024. URL https://blog.heim.xyz/a-trusted-ai-compute-cluster/.

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., Osborne, M. A., and Zilberman, N. Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation, March 2024. URL http://arxiv.org/abs/2403.08501. arXiv:2403.08501 [cs].

Hendrycks, D., Schmidt, E., and Wang, A. Superintelligence Strategy: Expert Version, March 2025. URL http://arxiv.org/abs/2503.05628. arXiv:2503.05628 [cs].

Ho, A. The least understood driver of ai progress, 2026. URL https://epoch.ai/gradient-updates/the-least-understood-driver-of-ai-progress. Accessed: 2026-04-19.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic progress in language models, March 2024. URL http://arxiv.org/abs/2403.05812. arXiv:2403.05812 [cs].

Ho, A., Denain, J.-S., Atanasov, D., Albanie, S., and Shah, R. A rosetta stone for ai benchmarks. *arXiv preprint arXiv:2512.00193*, 2025.

Johnston, T. The reporters at this news site are AI bots. OpenAI's super PAC appears to be funding it. Model Republic, April 2026. URL https://www.modelrepublic.org/articles/the-reporters-at-this-news-site-are-ai-bots.-openai%E2%80%99s-super-pac-appears-to-be-using-it-to-advance-its-political-agenda. Accessed: 2026-04-24.

Joud, R., Moëllic, P.-A., Pontié, S., and Rigaud, J.-B. A practical introduction to side-channel extraction of deep neural network parameters. In *International Conference on Smart Card Research and Advanced Applications*, pp. 45–65. Springer, 2022.

Karvonen, A., Reuter, D., Rinberg, R., Marks, L., Garriga-Alonso, A., and Warr, K. DiFR: Inference Verification Despite Nondeterminism. *arXiv preprint arXiv:2511.20621*, 2025.

Kimi Team. Kimi k2.5 technical report. https://github.com/MoonshotAI/Kimi-K2.5/blob/master/tech_report.pdf, 2026. Moonshot AI.

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*, 2023.

Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., and Dean, R. AI 2027. https://ai2027.com/, April 2025. URL https://ai2027.com/. Accessed: 2025-04-07.

Kowal, M., Timm, J., Godbout, J.-F., Costello, T., Arechar, A. A., Pennycook, G., Rand, D., Gleave, A., and Pelrine, K. It's the Thought that Counts: Evaluating the Attempts of Frontier LLMs to Persuade on Harmful Topics. *arXiv preprint arXiv:2506.02873*, 2025.

Kryś, J., Sharma, Y., and Egan, J. Distributed and decentralised training: Technical governance challenges in a shifting AI landscape. *arXiv preprint arXiv:2507.07765*, 2025.

Kulp, G., Gonzales, D., Smith, E., Heim, L., Puri, P., Vermeer, M. J. D., and Winkelman, Z. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/WRA3056-1.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint arXiv:2501.16946*, 2025.

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., et al. Measuring AI Ability to Complete Long Tasks. *arXiv preprint arXiv:2503.14499*, 2025.

Leading the Future. AI industry launches "leading the future" to drive U.S. AI leadership, economic growth, national security, and innovation. PR Newswire, August 2025. URL https://www.prnewswire.com/news-releases/ai-industry-launches-leading-the-future-to-drive-us-ai-leadership-economic-growth-national-security-and-innovation-302537548.html. Accessed: 2026-04-24.

Lidin, J., Sarfi, A., Miahi, E., Anthony, Q., Chauhan, S., Pappas, E., Thérien, B., Belilovsky, E., and Dare, S. Covenant-72b: Pre-training a 72b llm with trustless peers over-the-internet. *arXiv preprint arXiv:2603.08163*, 2026.

Liu, P. AI's biggest builders are now its biggest lobbyists. Forbes, February 2026. URL https://www.forbes.com/sites/phoebeliu/2026/02/20/ais-biggest-builders-openai-anthropic-among-biggest-government-lobbyists/. Accessed: 2026-04-24.

METR. The Rogue Replication Threat Model. https://metr.org/blog/2024-11-12-rogue-replication-threat-model/, 11 2024.

Miller, C. *Chip War: The Fight for the World's Most Critical Technology*. Scribner, New York, NY, 2022. ISBN 978-1982172008. 464 pages.

Mitre, J. and Predd, J. B. *Artificial General Intelligence's Five Hard National Security Problems*. RAND Corporation, Santa Monica, CA, 2025. doi: 10.7249/PEA3691-4.

Mukobi, G. Reasons to doubt the impact of AI risk evaluations. *arXiv preprint arXiv:2408.02565*, 2024.

Petrie, J. Near-term enforcement of AI chip export controls using a firmware-based design for offline licensing. *arXiv preprint arXiv:2404.18308*, 2024.

Pilz, K. F., Sanders, J., Rahman, R., and Heim, L. Trends in ai supercomputers. *arXiv preprint arXiv:2504.16026*, 2025.

Public Citizen. One in four federal lobbyists now work on AI. Public Citizen, February 2026. URL https://www.citizen.org/news/one-in-four-federal-lobbyists-now-work-on-ai/. Accessed: 2026-04-24.

Rahman, R. Catching illicit distributed training operations during an AI pause. MIRI Technical Governance Team Blog, April 2026. URL https://techgov.intelligence.org/blog/catching-illicit-distributed-training-operations-during-an-ai-pause. Accessed: 2026-04-24.

Reuters. Us clears h200 chip sales to 10 china firms as nvidia ceo looks for breakthrough. Reuters, May 2026. URL https://www.reuters.com/business/retail-consumer/us-clears-h200-chip-sales-10-china-firms-nvidia-ceo-looks-breakthrough-2026-05-14/. Exclusive.

Rinberg, R., Karvonen, A., Hoover, A., Reuter, D., and Warr, K. Verifying LLM Inference to Detect Model Weight Exfiltration. *arXiv preprint arXiv:2511.02620*, 2025.

Rogiers, A., Noels, S., Buyl, M., and De Bie, T. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*, 2024.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing Power and the Governance of Artificial Intelligence, February 2024. URL http://arxiv.org/abs/2402.08797. arXiv:2402.08797 [cs].

Scannell, K. Co-founder of tech company charged with diverting $2.5 billion in Nvidia AI chips to China in violation of export laws. CNN, March 2026. URL https://www.cnn.com/2026/03/19/politics/super-micro-computer-founder-charged-ai-chips-china. Accessed: 2026-04-24.

Scher, A. Catch-Up algorithmic progress might actually be 60× per year. MIRI Technical Governance Team Blog, December 2025. URL https://techgov.intelligence.org/blog/catch-up-algorithmic

-progress-might-actually-be-60x-per-year.

Scher, A. and Thiergart, L. Mechanisms to Verify International Agreements About AI Development, November 2024. URL https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development.

Scher, A., Abecassis, D., Barnett, P., and Abeyta, B. An international agreement to prevent the premature creation of artificial superintelligence. *arXiv preprint arXiv:2511.10783*, 2025.

Sevilla, J. How far can decentralized training over the internet scale?, 2025. URL https://epoch.ai/gradient-updates/how-far-can-decentralized-training-over-the-internet-scale. Accessed: 2026-06-24.

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., Agarwal, R., Anil, C., et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025.

Sharma, M., McCain, M., Douglas, R., and Duvenaud, D. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage. *arXiv preprint arXiv:2601.19062*, 2026.

Talbott, S. *Deadly Gambits: The Reagan Administration and the Stalemate in Nuclear Arms Control*. Vintage Books, 1985. ISBN 978-0-394-74009-6. Google-Books-ID: zDQiAQAAIAAJ.

Thadani, A. and Allen, G. C. Mapping the semiconductor supply chain. *Center for Strategic and International Studies*, 11, May 2023.

Vermeer, M. J. D. *Evaluating Select Global Technical Options for Countering a Rogue AI*. RAND Corporation, Santa Monica, CA, 2025. doi: 10.7249/PEA4361-1.

Villalobos, P. and Atkinson, D. Trading off compute in training and inference, 2023. URL https://epoch.ai/blog/trading-off-compute-in-training-and-inference. Accessed: 2025-03-28.

Wasil, A., Berglund, L., Reed, T., Plueckebaum, M., and Smith, E. Understanding frontier AI capabilities and risks through semi-structured interviews, July 2024. URL https://papers.ssrn.com/abstract=4881729.

Xu, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., Ling, C., et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zelikow, P., Cuéllar, M.-F., Schmidt, E., and Matheny, J. Defense against the AI dark arts: Threat assessment and coalition defense. Report, Hoover Institution, Stanford, CA, 12 2024. A Publication of the Hoover Institution.