Pacing the Frontier

A Framework & Research Agenda

by Raymond Douglas, Charles Dillon, Nikola Moore, Gavin Leech, Shahar Avin

...and Mathias Kirk Bonde, Rohit Krishnan, Noah Perez, Nathan Young, Cormac Slade Byrd, Stephen Casper, Jan Kulveit, & David Duvenaud[1]

Abstract

Companies and governments are already haphazardly pacing AI. To make good decisions about the pace of AI progress, practitioners, policymakers, and the public need a richer understanding of the options and dynamics that pacing interventions create. This piece surveys the range of interventions aimed at moderating the pace of AI development, deployment, and diffusion, and proposes a detailed research agenda to clarify the tradeoffs and likely effects of each. Now is the time for a dedicated field of pacing research, moving beyond one-off proposals and scenarios toward a flexible capacity to quickly and effectively intervene.

---

1. Raymond Douglas, Jan Kulveit — ACS Research
   Raymond Douglas, David Duvenaud — University of Toronto
   Charles Dillon, Nikola Moore, Gavin Leech, Noah Perez — Arb Research
   Gavin Leech — Paradigm 3 Institute
   Rohit Krishnan — The Wharton School, University of Pennsylvania
   Mathias Kirk Bonde — Independent
   Nathan Young — Goodheart Labs
   Cormac Slade Byrd — Trajectory Institute
   Stephen Casper — Harvard University and Harvard Kennedy School
   Shahar Avin (senior author) — University of Cambridge

This work is funded by ACS Research and the Paradigm 3 Institute. ↩

Contents

Executive Summary

"Pacing AI" usually refers to how to conclusively handle the most extreme risks in the face of race dynamics. However, even for the goal of handling these highest-stakes cases, it's useful to take a broad view of pacing—one that encompasses all interventions aimed at moderating the pace of AI development, deployment, or diffusion. Thus:

- **Haphazard pacing is already common**, including: delaying model releases for safety testing, pausing model development in response to shocks, and applying export controls.
- **Current approaches will predictably fail.** Isolated, unilateral actions addressing only small fractions of the problem are not enough, but poorly executed interventions could easily backfire—good solutions will need to be carefully designed.
- **Precedents are being set whether we like it or not.** How AI progress is paced now will shape how it is paced in future. We can learn from the shortcomings of existing attempts, and think about what precedents we are now setting for higher-stakes cases.
- **Rapid, unpredictable progress requires an existing body of research to navigate well.** Pre-specified proposals aren't enough when key decisions often depend on sensitive information, have to be made quickly, and are responses to surprises.

Given all of this, we think it is time for pacing to be a dedicated research area.

Though there have been many specific proposals, and though many subfields of AI risk largely exist to feed into pacing decisions, pacing as a whole has not yet cohered into a clear field of study. There are massive gaps in our understanding of pacing as a whole. Here we highlight three topics to illustrate the work that needs doing:

- **Understanding incentives.** The practical effects of an intervention will depend substantially on how the affected parties respond to them, including the actors who enforce the intervention. We recommend prioritizing research on:
  - How different forms of transparency about capabilities can help or hurt coordination;
  - How to prevent mission creep among whoever is empowered to oversee pacing interventions;
  - Ways to align the incentives of different actors around pacing, e.g. by predictably compensating for losses with minimal moral hazard.
- **Improving technical and regulatory interventions.** All interventions face tradeoffs, but new coordination and oversight mechanisms and prior planning can allow strictly better options. For example:
  - Designing regulatory structures that allow for rapid but limited interventions which can make room for more careful deliberation;
  - Technical pathways to getting high assurance with minimal invasiveness, like LLM-based oversight and cryptographic guarantees;
  - Modeling the consequences of indirect interventions like buyouts, liability, and taxes.
- **Investigating the full lifecycle of a pacing intervention.** By considering the full sequence from before to after, we can spot gaps. For example:
  - Mapping different ways that interventions can end, and what this implies about different actors' willingness to participate;
  - Figuring out how to make interventions which are more robust to premature endings or imperfect execution, intentional or otherwise;
  - Dry runs and wargames to stress-test specific interventions.

For each topic, we include recommended reading and starting points below. We invite readers to `reach out` if they are interested in further work.

1. Introduction

> *"Society at large may need the option to buy time to address emerging risks, develop security measures, and strengthen oversight. But each company—and country—is under intense competitive pressure not to unilaterally slow that acceleration. And today, the world lacks the technical and governance tools to deliberately pace frontier-wide progress."*
>
> — 1,367 employees of frontier AI companies

How should AI developers balance growing the useful capabilities of their models against their waning ability to oversee and control such models? How should regulators handle the spread of systems that can carry out advanced cyberattacks, or even illegally initiate them? How should governments navigate arms-race dynamics, in which each side fears that its restraint will be exploited?

We take a broad notion of pacing that encompasses any interventions that deliberately moderate the pace of frontier AI development, deployment, and diffusion. From this perspective, actors already routinely make costly choices around pacing. Developers delay releases, roll back deployments, and even pause training in response to harms that they are not equipped to mitigate. In the coming years, these choices will become far higher-stakes, as AI systems become more advanced, more embedded in society, and more regulated, and as AI development takes on greater geostrategic significance.

Many of these decisions will need to be made quickly, without all the relevant information, and while balancing competing interests. So far most pacing has been unilateral, through AI developer self-pacing (e.g. delayed releases) or governments pacing diffusion (e.g. via export controls or halts on model releases), but in future, pacing may require coordination between actors with different incentives, and potentially between adversaries. Early pacing attempts, whether successful or not, will

substantially shape the precedents, institutional knowledge, and resulting evidence that will guide future decisions.

It is therefore extremely important that those in a position to pace are provided with a clear user's manual of all the options and tradeoffs available, not just a set of exemplar pacing proposals. Given the unpredictable nature of AI progress and its geopolitical context, the specific pacing dilemmas that come up in real life are likely to differ substantially from anything we can write down today, and may rely on private information only available to a small set of actors. If we want those decisions to be sensible, we need to provide a strategic decision-making framework ahead of time.

Mountains of research exist on specific topics relevant to pacing (e.g. model evaluations, capability forecasting, compute monitoring), and on proposals for specific interventions. But there is comparatively little on the overall question: what effects pacing interventions will have in different circumstances. This piece gives a broad account of the whole area and a list of open questions to anchor a dedicated field of AI pacing.

Inquiries into pacing are naturally in danger of being politicized: indeed, hundreds of millions of dollars have already been spent on advocacy on both sides of the debate. But that is all the more reason to encourage dispassionate research and a shared understanding of the practical implications.

1.1. Structure of the piece

This piece is structured around a series of questions intended to mirror how one might develop or evaluate a given intervention:

- **Why pace?** Section 2 examines motivations for and against pacing, for humanity at large and for specific actors.
- **Pace what?** Section 3 catalogs which parts of AI development can be paced, and the challenges in picking appropriately.
- **Pace how?** Section 4 follows the lifecycle of an intervention from anticipation to exit, asking what it takes for each step to succeed.

- **Then what?** Section 5 considers the broader effects of interventions, including on AI R&D, the economy, and the distribution of political power.

At the end of each section we give a list of open questions. Throughout this document we use two stylized cases to demonstrate how our analysis applies to potential interventions:

1. A coordinated cap on the compute used to train individual frontier models, intended to slow the pace of R&D acceleration (particularly the risk of recursive self-improvement) which could outpace oversight and increase the risk of humans losing control of AI.
2. An internationally agreed-upon set of restrictions on access to and release of AI models which materially increase users' ability to develop biological weapons, as an example of pacing diffusion (rather than pacing capability development).

We are not trying to advocate for either proposal; these end of section examples were instead chosen to illustrate how our frameworks apply to specific plans. See Appendix 8.2 for a longlist of possible interventions.

A bibliography of work related to this agenda can be found here.

2. Why Pace?

Arguments for and against pacing come from two broad perspectives. Will society as a whole be better off if AI progress is deliberately slowed down or sped up? And why do specific actors in the AI R&D ecosystem seek to pace?

From the global perspective, the question is: Given the current state of AI progress, the historical track record of intervening in R&D, and the existing economic, political and institutional structures, is it a good idea to attempt to pace AI progress? We start by considering arguments from this perspective, first against (§2.1) and then for (§2.2) increased pacing.

More pragmatically, pacing will only happen if actors make effective pacing interventions. If the actors are rational, such actions will be taken if the actors view them as having a positive impact in expectation. In §2.3, we consider the perspectives and cost-benefit calculus of different actors.

§5 gives a more comprehensive assessment of the benefits and costs of pacing, taking into account second-order effects of the interventions required to make pacing effective.

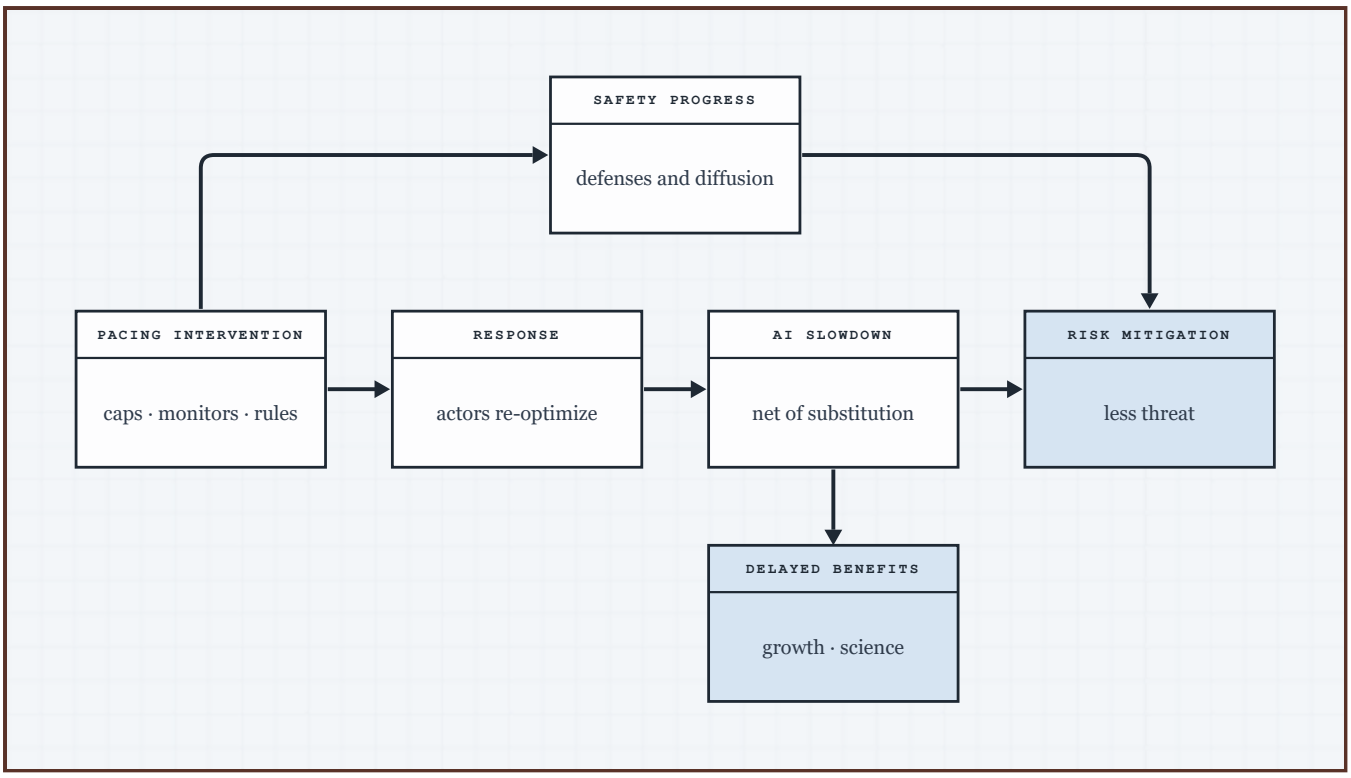

Figure 1: Simple causal graph of how pacing interventions work

2.1. Why pace less?

2.1.1. Pacing means waiting longer for very good things

AI progress so far has had some highly positive direct impacts: software development is far faster and cheaper, access to information and advice has widened in every domain of life, and tools like machine translation, automatic paperwork filling, and scientific assistance save society vast amounts of time.

Many observers expect significant impacts on economic growth from future AI progress, with the median expert in one survey expecting rates of growth to more than double in advanced economies by 2050 under their "rapid" scenario, which would amount to tens to hundreds of trillions of dollars annually. The largest companies in the world have been betting on this, with capital expenditures in 2026 alone expected to exceed US$1 trillion, and various beneficiary companies in the supply chain valued in the trillions of dollars. And AI already seems to be speeding up life-saving technology.

Reducing the rate of frontier AI progress would delay these enormous benefits. People might well counterfactually suffer and die from progress not happening as fast. (This opportunity cost may even include mitigating other potential existential risks, such as biorisks and nuclear war. So even the threat of existential risk from advanced AI does not necessarily warrant pacing, depending on its relative reduction of other risks.) Thus, the case for pacing more must be weighed against this opportunity cost.

2.1.2. Pacing can directly cause bad outcomes

Aside from the potential value sacrificed, pacing interventions can very directly cause harm. We highlight three particularly significant risks:

Power concentration. Interventions often require an enforcer—for example, to monitor compute supplies, or inspect logs, or secure access to advanced models. Such bodies can then find themselves with sensitive information, privileged access to dangerous capabilities, and/or discretionary enforcement power. This naturally comes with the risk of abuse or simply of bureaucratic incompetence. We discuss this further in §5.3.

Overhangs in AI progress. When some part of AI progress is temporarily limited, progress can afterward resume at a *faster* pace if some resource (compute, efficiencies, ideas, etc.) has accumulated in the interim. For example, if a restriction were placed on training runs of a certain size, AI developers would likely learn to train more efficiently. If the restriction expires or someone violates it there could be a sudden burst of progress that produces even more harm than the intervention prevented, due to the lack of time to adapt. Badly designed pacing thus converts smooth risk and predictable progress into jagged progress and lumpy risk. We discuss this further in §4.4.

Races to the bottom. In situations involving several competing actors where pacing is competitively costly, the less cautious actors can pull ahead while the more scrupulous ones fall behind. This makes it much harder to manage risks which only depend on the least safe actor, like the proliferation of dangerous capabilities in open-weight models. Even actors who are concerned about risks have to balance immediate harms against the risk of falling behind and losing influence. We discuss this further in §2.3.

2.1.3. Good pacing can get in the way of better pacing

If different interventions rely on the same scarce resources like finite political capital or technical expertise, then aiming for the best forms of pacing becomes all the more important.

One way this manifests is that it might be very important to time pacing right. Some kinds of empirical safety work addressing risks from highly advanced AIs become easier as capabilities move closer to the point where the risks manifest: lessons learned through experiments on more advanced models are more likely to remain valid when capabilities reach the true danger zone. Slowing down preemptively might cost political capital which could be better spent at a higher-leverage time.

Indeed, AIs themselves are likely to be an important tool in addressing risks. For example, a classic motivation for slowing AI progress is to buy time for research into value alignment, oversight, and interpretability. But one of the main current bets in AI safety is developing an automated alignment researcher that can pack thousands of person-years of research into a short period. This will be far more effective as AIs themselves become more capable.

The strength of this argument depends on how feasible the better options actually are and how scarce the resources actually are. A major empirical question for pacing is how far in advance those in a position to act will see the risks and how quickly they will be able to react, as we discuss in §4.1 and §4.2.

2.1.4. Pacing historically has been hard to undo

Many technologies have been regulated in a way that slows development, ranging from germline genetic engineering and geoengineering—which have been de facto banned for decades—to nuclear power and flight.

Consider nuclear power, which was heavily restricted from the 1980s onward, ostensibly for safety reasons. The slowed rollout of nuclear power has been hard to undo, with substantial environmental impacts. A huge amount of carbon dioxide and indeed geopolitical strife could have been avoided if all of Europe had built as many nuclear power stations per capita as France. This partly overlaps with §2.1.3, where some better mechanism might have been found.

But beyond this, it burned investor trust and disincentivized investments in nuclear power. Investors rightfully feared capricious and politicized regulation would block nuclear power plants, regardless of the underlying merits.

It is likewise possible that pacing will not have the trust of those in AI, because they correctly predict it will be hard to undo, even if the risk is found to be low.

2.2. Why pace more?

2.2.1. AI threats take time to understand and mitigate

As AI systems become more advanced and distributed, they create new risks which need mitigating. Sometimes the risks come faster than the mitigations, and so the relevant actors will benefit from extra time to adjust.

One concrete example of this is how Anthropic's `Project Glasswing` and OpenAI's `Trusted Access for Cyber` program delayed the public release of their most advanced models because of their potential to aid in cyberattacks, while still providing them to certain key groups that could use them to shore up their cyber defenses.

Beyond specific risk considerations, advanced AI will likely bring substantial societal upheaval, and time may be needed for citizens to collectively make decisions about what sort of changes and adaptations are desirable, before these changes are thrust upon them and become difficult to roll back.

The range of activities that pacing can buy time for is extremely broad. It could look like building technical approaches to prevent harmful model behavior, or developing better governance frameworks for handling emerging capabilities, or simply giving society more time to adjust to fundamental disruptions. Particularly in the case of technical approaches, pacing can both buy time and free up resources to help with mitigation. For example, delaying frontier model development could make more compute and labor available for risk mitigation, while delaying the release of completed models could buy time for development of evaluations or specific resilience measures.

Beyond specific known threats, pacing can serve as a precaution that buys time to figure out whether there are major risks. For example, most frontier developers take time to evaluate their models for dangerous capabilities before deploying them. This is particularly relevant for parts of progress that become more difficult to reverse over time, like releasing model weights: once a set of weights has been distributed, it is essentially impossible to retroactively add more safeguards.

Even after a threat has been identified, the capacity to deal with it will vary. Sometimes it will be quite straightforward—for example, post-training a model to refuse some sufficiently high percentage of harmful requests. Here, the case for pacing is easy to evaluate: a well-scoped mitigation with a fairly predictable cost in time. But in other cases it will not be clear what mitigations are necessary or how long they would take, for example because they require novel scientific breakthroughs. This is likely to become more of a challenge as AIs become more capable of reasoning about evaluations and engaging in strategic deception. Model evaluators' confidence in AI systems' alignment will become weaker even as their capacity for harm increases.

In cases featuring an evolving risk (such as progressive increases in the capacity of AIs to evade oversight or incremental disruptions to the labor market) it may be preferable to titrate: to allow capabilities to gradually outrun protections in a controlled manner that caps the potential harm, in order to learn what type of protection is required. A challenge with this approach is that it necessarily involves accepting risks without understanding them well enough for mitigation, on the *assumption* that it will be feasible to tolerate the harm and eventually figure out a solution despite not understanding the problem yet.

Where the harm is potentially catastrophic and irreversible, however, there might not be an opportunity to learn from mistakes. If waiting merely postpones benefits, then there is an asymmetry between the cost of going too fast versus too slow which strengthens the case for greater precaution.

2.2.2. AI threats can be sudden and unpredictable

Predicting all the potential effects of AI progress is extremely challenging even in the short term. Sometimes a system displays an unexpected capability or causes unexpected outcomes, and even without a clear picture of how one would remedy the situation, it may be helpful to slow the activity implicated in the warning and try to

make sense of the situation. This has a few benefits. Firstly, it means that capacity can be redirected toward understanding the situation and its broader implications. Secondly, it means that there is less chance of further unexpected outcomes. Thirdly, it buys time to more carefully make decisions which might otherwise be hard to reverse.

The immediate use of time is therefore inquiry: investigating what happened, perhaps reproducing the result, and determining which assumptions need to change. A benefit of progressing at a slower pace is that this becomes more feasible—some processes, such as human-led investigation and deliberation, can only be done so quickly, and if the pace of overall progress becomes too fast these may lose a lot of their effectiveness.

For example, OpenAI temporarily `halted` internal development of its "Astra" model after discovering `another unreleased model` had broken out of a secure sandbox, subverted OpenAI infrastructure, and launched a cyberattack on a number of other companies, and after Astra demonstrated "`critical`" cyberoffense capabilities. This incident was unprecedented enough to warrant a sudden stop, even without a particular plan for how to respond, because reasonable security assumptions were violated and the risk of continuing was judged too high.

Unexpected threats are hard to fully plan for. The existence of this category points to the need not just for specific intervention plans but for the capacity to quickly and effectively intervene in new ways that address changing situations.

2.2.3. AI progress can create lose-lose situations

Advances in AI could create situations where several parties are in competition in a way that is inherently unstable and destructive. If highly autonomous systems could conduct AI research much more quickly without humans in the loop, there would be strong pressure for developers to allow them to do so. Even those concerned about risks from the absence of oversight could find it too costly to unilaterally refrain from these speedups.

A similar problem would arise if governments could enhance their military capabilities by delegating decision-making powers to much faster-than-human AI systems. Pressure to reduce human review in order to stay competitive would increase the aggregate risks of unintended escalation.

In such situations, the natural incentives would pull toward a lot of value being lost. Once such situations arise, they bring with them the challenge of coordinating against those incentives. But it is also possible to avoid entering those situations: to deliberately avoid creating technologies that produce these dynamics, or to delay their creation until other complementary technologies can mitigate the risk.

2.3. Actors and their incentives around pacing

Beyond the reasons humanity has for favoring or opposing pacing interventions, individual actors will naturally weigh their own self-interest when considering interventions. Managing AI progress is difficult due to this tension between one's own interests and those of the collective (e.g. when frontier labs must decide how much detail to share about internal safety incidents, or when states choose how much to automate their military).

Here we look at some of the major types of actors involved in AI progress and the factors that shape their incentives to pace or not pace. These actors' and a broader list of actors' incentives are discussed in more detail in §5. But the history of technology nonetheless contains many examples of powerful actors restraining technology, against their apparent narrow self-interest.

2.3.1. Governments

On the international scale, the main challenge for governments in coordinating with each other is the lack of an existing higher authority to appeal to that can credibly enforce interventions. This puts them at risk of arms-race-like dynamics which no other actor can step in to solve. In particular, governments must manage the potential for AI as a military technology.

Governments vary in their different strategic positions and degrees of leverage over different parts of AI progress, which means that they will be incentivized to support very different forms of pacing. Abstractly, those in the lead have a reason to support pacing that consolidates a lead, and those behind have an incentive to support pacing that lets them catch up. Concretely, many parts of the global AI supply chain still largely depend on single specific companies that are physically based in specific

countries. As such, governments already take some unilateral actions to cement their leads by pacing their competitors, most notably by limiting the flow of hardware, model capabilities, and researcher talent.

Domestically, governments are the parties most able to solve internal coordination problems, by regulating developers, infrastructure, and wider usage. Not only do frontier developers face difficulties finding appropriate ways to coordinate with each other, but in some cases they may not be legally allowed to because of antitrust regulation.

It is important to distinguish between the theoretical power available to governments and the actual opportunities available to government employees. Practically speaking, governments are heavily constrained by their own laws and precedents, and have difficulty surfacing expertise. So their ability to regulate effectively, for example, is partly limited by their ability to identify experts and allocate an appropriate amount of funding toward them - for instance, the nominal annual budget of the US government AI evaluator is less than the salary of some individual frontier AI researchers. But this could plausibly change very quickly in the case of a perceived emergency.

AI also has the potential to seriously disrupt the relationship between governments and their citizens. AI could massively increase the degree of leverage that governments have over their citizens, by enabling larger-scale surveillance and by reducing the degree to which governments depend on citizens for tax revenue or a security apparatus. Conversely, AI could seriously disrupt governmental authority simply by moving faster than regulation can and presenting a new source of power that governments cannot easily gain a stake in. This presents an interesting problem: governments have the incentive and capacity to pace the capabilities which would make them relatively weaker, but not the ones which would make them relatively stronger.

Alongside this, governments have incentives to be seen to be acting and taking public concerns seriously. Politicians may share these concerns, and regardless will face electoral or reputational pressure to respond to them. Salient high-profile incidents or sustained public pressure could make pacing politically attractive even in cases where it conflicts with other political or commercial priorities. The most effective interventions for satisfying citizens' concerns, however, may not coincide with the most effective interventions by other lights.

2.3.2. AI developers

AI developers are in an awkward economic position: developing a frontier AI model is enormously expensive, extremely valuable, and yet the product's commercial value decays within months as new frontier models are developed. This dynamic puts them in tight competition with each other. This competition makes it harder for them to individually prioritize safety, and also makes it harder for them to function as a cartel. Instead they are pulled toward intense R&D investment.

Indeed, even if a developer were heavily motivated by values other than profit, it seems likely that most of the potential impact (and indeed profit) of AI is concentrated in more advanced models, so the main way to maximize individual leverage is by trying to get to the front of the pack.

Developers delaying frontier model releases to make time for safety testing is already extremely expensive because it cuts into the limited time that the model can actually be at the frontier, although there are also incentives pushing them toward this (high-profile incidents or misalignment degrading the capabilities that customers desire is not good for business).

Compared to governments, developers currently have access to a much deeper bench of technical talent and a far richer understanding of the state of their own AIs, and indeed their own future plans. In some ways this makes them better placed to reason about what type of pacing would be best. However, much like governments, developers will prefer different forms of pacing depending on their strategic position: while established institutions can more easily handle onerous compliance, frontrunners have more to lose from hard caps on model capabilities or training size, and so on.

At least for now, though, frontier developers are fiercely competitive over researcher talent, and therefore top researchers have quite a lot of leverage. Researchers at the current frontier labs display a notable desire for at least the *capacity* to pace frontier progress.

2.3.3. Citizens

Citizens have the least direct influence over AI progress. Their authority routes almost entirely through governments, which in turn vary in their leverage. They have a lot to

lose and a lot to gain: aside from the potentially existential harms, advanced AI could plausibly cause mass labor displacement and leave citizens with negligible hard power over governments, but it could also create an enormous surplus of value and great leaps in scientific progress that could be widely distributed. Most people will not be in a good position to form a clear sense of the balance of costs and benefits.

All things being equal, we might expect citizens to dislike extremely rapid and large-scale disruptions even if the net effect looks plausibly positive, and we might expect them to advocate for pacing the parts of AI progress that seriously weaken them, like mass surveillance or military automation. But unfortunately there is a risk that citizens can be maneuvered into acting against their own long-term interest: by accepting generous income top-ups as a condition for supporting broad labor automation that will in the long run leave them without leverage.

2.3.4. Infrastructure providers

The compute supply chain—semiconductor companies like Nvidia and Micron, and major cloud providers such as Google and Amazon—have a lot of leverage over the rate of AI progress.

Under many different plans they will have control over significant leverage points for interventions: hardware interventions, tracking compute usage, and managing various inference access points for end users. They and their investors may ultimately bear a lot of the costs of an intervention, and the opportunity cost of anything which slows down the rate of the overall datacenter buildout (see §5.4 and §3.1).

2.3.5. What this means for a workable intervention

Rather than rely on voluntary, unilateral interventions, some contexts will require coordinated pacing. This is more onerous, because it requires coordination during the various stages of pacing, and then mechanisms to ensure the coordination persists. It is also very difficult to specify what a fair coordinated intervention would look like, because actors vary so much in their strategic positions.

It is not a given that coordinated pacing would actually be good for the world. One risk is that actors coordinate *against* the common good. Without an authority capable of

constraining all relevant actors,the only ways to pace (without merely ceding power to another actor) are either to coordinate, or to have a lead you can afford to burn.

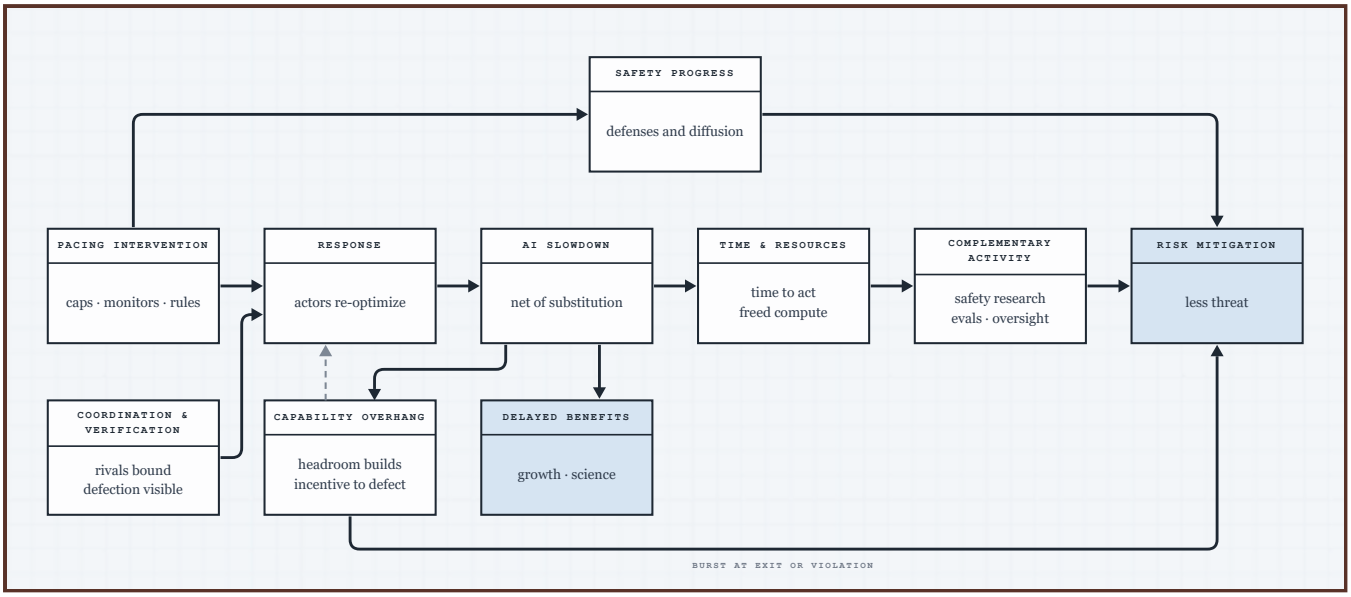

Figure 2: Full causal graph of how pacing interventions work

2.4. Case A: A cap on frontier training

Frontier model progress has increasingly made AI systems capable of automating their own further improvement. Anthropic, for example, has claimed that it is producing 8x as much code per researcher since the release of Mythos 5, when compared to the pre-2025 baseline, and that its own researchers estimated they were sped up by a factor of 4x, though Anthropic thinks that this was likely an overestimate. If this AI contribution to AI became sufficiently large, capability development could accelerate while also becoming less dependent on human researchers. The time available to evaluate successive systems might shrink, even as previously functional oversight measures break down and unexpected new risks emerge.

One direct intervention aimed at slowing down this trajectory could be a cap on the compute budgets for training frontier models. This would impede one major source of capabilities progress, and extend the window of controllability before existing methods and systems are no longer capable of

adequately supervising AI progress. Evaluation, security, and governance systems may be inadequate for development substantially accelerated by AI.

Such a cap would also delay beneficial capabilities (including AI-assisted safety research). Governments and developers would be more likely to support this intervention if they knew their competitors faced similar constraints, and there was a clear path to lifting the cap given specific progress. Loopholes and adaptations that would reduce the intended effect on capabilities are known to exist. The case for the cap therefore is very sensitive to how credibly it will constrain competitors, and what can be achieved during its implementation.

2.5. Case B: Restricting access to dangerous biological capabilities

AI systems are increasingly able to aid some users in developing biological weapons. Governments and developers may face some lag in their ability to assess uplift, to restrict it in specific models, and to deploy model capabilities to develop countermeasures. Restrictions could include safeguards on publicly available models, with access to less restricted versions remaining gated (as with Anthropic's release of Fable 5 and Mythos 5). Furthermore, releasing the weights of individual models removes any ability to regulate them if they turn out to provide an unacceptable degree of uplift.

Pacing could buy time for better uplift evaluations, safeguards and unlearning, secure hosting, user authentication, controls on model weights, public-health preparedness, and international procedures for handling dangerous models and evidence.

These interventions could complement controls on biological materials and laboratory infrastructure, though the value depends on how much marginal protection they add. They could also impede legitimate biological research for those without access to the less-restricted model versions.

2.6. Open Research Questions

- **How much, and in what ways, would more time allow us to better manage various AI risks?** What are the bottlenecks to mitigation or adaptation of different AI risks? What factors other than time influence AI risk management? What risk management efforts can be taken now, and which can only be taken once certain AI capability or adoption thresholds are crossed? Related work:
  - *MacAskill & Moorhouse (2025),* *Preparing for the Intelligence Explosion* Explicitly sorts "grand challenges" by whether they need calendar time, human deliberation, or just more AI.
  - *Hobbhahn (2025),* *What's the short timeline plan?* A concrete inventory of which safety measures are ready to deploy now, and which need years of preparation.
- **How might pacing become more or less difficult over time?** What investments can be made now to preserve optionality over pacing in the future? In particular, under what circumstances does pacing now make future pacing more or less feasible? Related work:
  - *Rahman (2026),* *Does Distributed Training Undermine Compute Governance?*
  - *Sastry et al. (2024),* *Computing Power and the Governance of Artificial Intelligence*. Maps compute governance options and their readiness
- **How do actors in this space make decisions about pacing?** What evidence do they currently consider and what assumptions do they currently make? What pathways exist for external research to inform such decisions, e.g. in government or lab leadership, and what makes that information transfer more effective?
  - *METR (2025),* *Common Elements of Frontier AI Safety Policies*
- **Which AI risks are most likely to motivate a pacing intervention, now and in the future?** Where do different dangerous capabilities sit on the offense/defense balance, and how will that change over time? Related work:
  - *Garfinkel & Dafoe (2019),* *How Does the Offense-Defense Balance Scale?*

    - *Shevlane & Dafoe (2020),* *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*
    - *Esvelt (2022),* *Delay, Detect, Defend: Preparing for a Future in which Thousands Can Release New Pandemics*
- **How does transparency about capabilities affect coordination?** When does common knowledge of research progress intensify or weaken race dynamics, for example by revealing that a rival is close behind or that progress is possible? Related work:
    - *Bostrom (2017),* *Strategic Implications of Openness in AI Development*
    - Armstrong, Bostrom & Shulman (2016), *Racing to the Precipice: a Model of Artificial Intelligence Development* Better information about rivals' capabilities can *increase* danger when teams are close, because it removes uncertainty that induces caution.
- **What are the risks and benefits of titration** ("deploy it and learn the risks empirically")? Have past release strategies from frontier labs succeeded in managing known AI risks? What might change in the future risk landscape? Given our uncertainty about the true risks, what criteria should gate the deployment of new frontier models?
    - *Shevlane et al. (2023),* *Model Evaluation for Extreme Risks*.

3. Pace What?

AI progress does not have a simple speedometer or brake. Any proposal needs to identify a particular outcome or threat model it wants to influence, the activities which contribute to it, and the places where those activities can be influenced.

A successful pacing intervention would change the rate of one or more activities that make up AI progress, but there are many activities to choose from and the choice matters for the effectiveness, cost, impacts, and desirability of the pacing intervention. It is helpful to distinguish between the *hazard* (the harmful outcome the proposal seeks to prevent) and the *control surfaces* (the parts which can actually be changed), because they tend not to neatly line up.

For any particular threat model, there is rarely one lever which maps cleanly and comprehensively onto that target. But there are levers on compute, on particular algorithmic approaches, on model releases, and on classes of use. Each of these has some bearing on the ultimate target, catching some activity that is harmless and missing some that is not, with varying costs to privacy, economic growth, and oversight.

In general, interventions that target control surfaces earlier in the AI R&D process (e.g. interventions on access to chips) tend to have broader impacts, a higher likelihood of lowering risks, a higher likelihood of harming beneficial progress, and less capacity to leverage information generated in the R&D process. This makes them much less likely to be targeted at very specific harms. Conversely, interventions that target control surfaces later in the AI R&D process (e.g. inference-stage "safeguards" or access controls) are much more able to leverage information and therefore be more targeted at specific harms, but are also more likely to be circumvented (as the harmful artifact already exists).

A longlist of 83 possible interventions can be found here.

3.1. Control surfaces of AI progress

3.1.1. The AI development and deployment chain

Progress at the frontier of AI development today involves a highly concentrated core of developers supported by an extensive and geographically distributed supply chain and research ecosystem. (However, the set of developers requiring oversight may expand during an intervention as other developers catch up - see §5.2.)

Various inputs are key for model development, most famously compute, but also energy, data center infrastructure, training data, research talent and engineering talent.

These inputs feed into the development process, which has many stages, from cleaning and structuring data, to running experiments, to the final process of building a frontier LLM—pretraining of a base model, and various post-training steps which transform this base model into the models which will eventually be served to users.

Once a model is developed, there is a spectrum of how broadly it can be deployed and released, from internal testing and evaluations, to internal use, external auditing or release to selected partners, and full public access via application or API. For open-weight models, a further stage is publishing the weights, which allows external actors to alter the model further.

What the model can then do when deployed depends not only on the model itself, but also on the infrastructure it is deployed on, the tools and harnesses available, the skill of its users, and the access it is given to resources such as the internet or particular data. (A pathway for potential future capabilities advances is *continual learning,* where models continue to improve after deployment based on the tasks they are working on. This would blur the line between development and deployment somewhat and could require its own solutions.)

All of these are plausible control surfaces to directly intervene on. Furthermore, they can all be affected indirectly. For example, the availability of researchers depends on immigration law, and frontier hardware depends on access to certain critical minerals.

Different actors vary in their leverage over these control surfaces. Abstractly, states have the ability to enforce domestic laws, while developers have a richer understanding of many parts of the design process. More concretely, specific states and developers vary in which of these they have leverage over—and there are some remarkably narrow bottlenecks in the AI ecosystem, like the reliance on a single Dutch corporation for a critical part of the AI hardware manufacturing process.

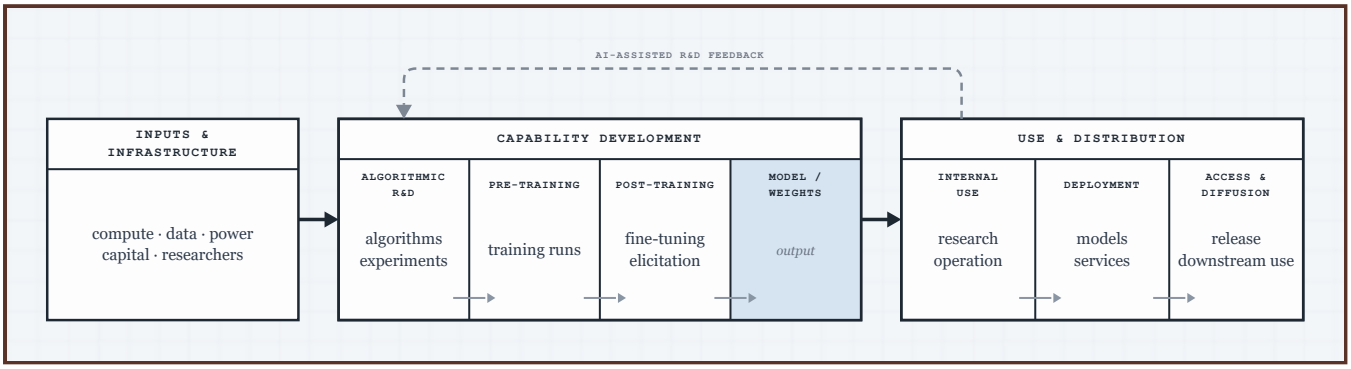

Figure 3: A three-part AI R&D production lifecycle from inputs and infrastructure, through capability development, to use and distribution, with an AI-assisted R&D feedback loop.

3.1.2. Considerations for choosing a control surface

One crucial distinction in the process of AI development is that partway through the resources shift in nature. Many of the early inputs are *rival* goods: resources whose use by one actor restricts their availability to others. Compute, capital, and researcher time, for example, are scarce, are concentrated within a small number of identifiable organizations, and can be controlled or redirected. There are also non-rival inputs, which can be replicated at almost no cost, (e.g. training data), but these are insufficient by themselves.

Downstream, however, it is mostly non-rival goods, like weights, elicitation methods, and training algorithms which influence the pace of progress (the notable downstream exception to this is inference compute, which is a rival good). These can all be copied and shared at low cost, so reliably controlling their diffusion is more difficult, and after they have diffused, it can be essentially impossible to undo.

These inputs allow for different kinds of intervention. The rival components have locations and quantities. They can be monitored, taxed, or redirected, and it is possible to keep track of who has them and exclude actors from access to them—mostly. The

non-rival components are more like information: once released, they are very hard to restrict access to.

That said, actually using a model requires inference compute, which is a rival good. However, it is also extremely general and extremely widely available. Many highly capable open-weight models can even be run on consumer hardware, though near-frontier models typically require more specialized infrastructure. In a future regime where continual learning is a major consideration, there may be new mechanisms for intervention at this stage.

Once a non-rival good like a set of model weights is proliferating, it is very difficult to destroy every copy, and there is no way to verify a claim that all copies have been destroyed, so interventions on their distribution are brittle. This is true even for closed-weight models, where uncertainty over whether adequate security over model weights is maintained means that, at current levels of security and verifiability, it is impossible to verify that all copies of model weights are known.

By contrast, for a rival good like computing hardware, an oversight body could be much more certain of the location and use of all copies of a particular type of chip, once adequate measures are in place, since the exact number of such chips can be more reliably known. However, these upstream goods, being far removed from the final product which carries the risks, are difficult to precisely target, and so restricting them will affect harmful and beneficial activities alike.

This highlights a recurring tradeoff. Control surfaces earlier in the chain are easier to observe and constrain, but further removed from the eventual harm, forcing interventions to be blunt. Later surfaces are often more amenable to precise targeting, but can be more difficult to intervene on and less robust.

CONTROL SURFACE	RISK LINKAGE	RISK COVERAGE	TARGETING PRECISION	EASE OF MEASURE-MENT	EXTERNAL VERIFIABILITY	ACTOR CONCENTRATION	CIRCUMVENTION RESISTANCE	INTERVENTION LEAD TIME
Chip supply	Low	Medium	Low	High	High	High	Medium	High
Compute access	Medium	High	Medium	High	High	High	High	High
Training runs	High	High	Medium	High	Medium	High	Medium	High
Training data	Medium	Low	Medium	Low	Low	Medium	Low	High
Research methods	Medium	Medium	Low	Low	Low	Medium	Low	High
Model weights	High	Medium	Medium	Medium	Medium	Medium	Low	Medium
Model access	High	High	High	High	High	High	Medium	Medium
Deployment	High	Medium	High	Medium	Medium	Low	Medium	Low

Figure 4: Control surfaces for AI development across the chain of developing and deploying systems.

3.2. Targeting and tradeoffs

Consider a pacing intervention targeting some subset of inputs to the production of dangerous AI capability. Its effects extend beyond the inputs it directly restricts. Inputs that *substitute* for the targeted one absorb the freed compute, labor and so on, and receive more investment, while inputs that *complement* it contract along with the restricted ones. Capping, e.g. frontier training runs does not slow AI development in strict proportion to the amount of training forgone; rather, it slows by that amount net of whatever the developers recover by substituting with algorithmic efficiency, data-quality and inference-time scaling. The effectiveness of an intervention therefore depends on how easily developers can substitute other inputs for those being restricted. Where those alternatives inputs are non-rival and harder to observe, an intervention may have the effect of pushing more effort toward less governable inputs.

Any intervention will restrict some harmless activity, and fail to catch all harmful activity. These are *false positives* and *false negatives*. Both errors can occur simultaneously: A classifier that aims to flag nefarious research related to

manufacturing biological weapons, for example, may be triggered by the innocent questions of a biology student. At the same time, the classifier may fail to catch carefully designed questions that split up a dangerous avenue of inquiry into unassuming parts.

Interventions earlier in the development process can have wider-reaching downstream effects, which generally means more false positives and fewer false negatives. (In a sense, halting all AI development forever would indeed prevent all harms, but it would also prevent all benefits.) In practice, however (and in part because of this risk of being over-broad), such interventions can be subject to carveouts and selective enforcement, which can open up alternative routes to the same ends. Downstream interventions have the opportunity to be more fine-grained and selective, but making use of this capacity depends on having the time and expertise to make good choices here, and in practice we may see blunter interventions such as excluding whole categories of use.

For most threats, there are multiple distinct pathways to their realization. A capability advance can come from more training compute, better algorithms, more elaborate post-training, or better elicitation of a model's existing latent capabilities. An intervention which targets one route but leaves the others open still allows threat-relevant activity to continue even if the intervention is perfectly enforced.

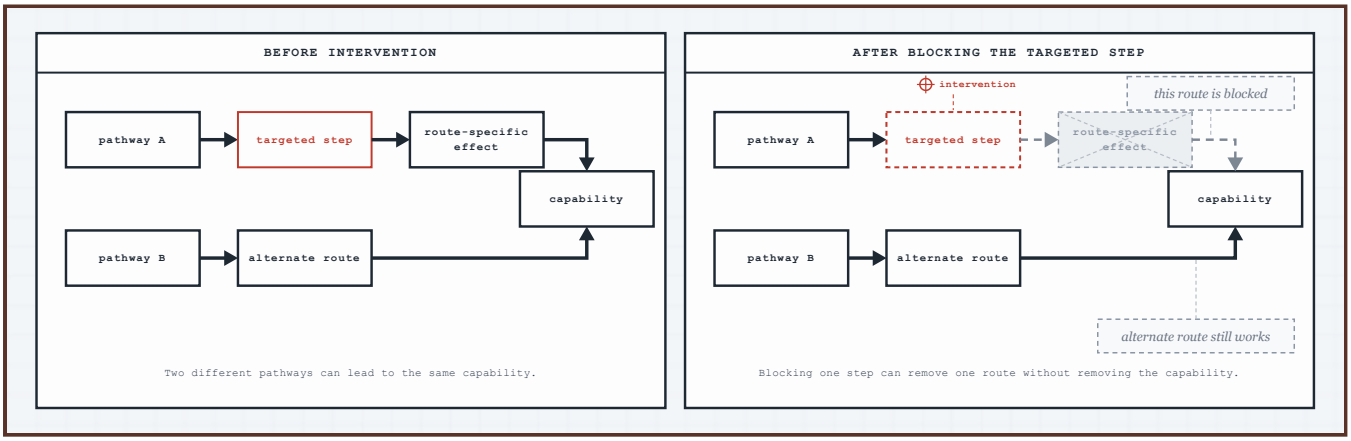

Figure 5: How a pacing intervention affects different pathways to dangerous AI capabilities.

The quality of selection also changes over time. This is partly because actors adapt in response to interventions, as we discuss in §5. But even beyond that, the effectiveness of an intervention depends on empirical assumptions about the relationship

between the control surface and the hazard. These assumptions may be changed by later progress.

For example, one conservative way to prevent the emergence of a dangerous advanced capability is to cap the number of FLOPs that can be used in training a model (see example A), but even this intervention depends on assumptions about training efficiency or elicitation methods which might change with future progress and substitution into alternative inputs. Generally this will push toward more false negatives, as the intervention becomes outdated.

For a given control surface, there will often be a three-way tradeoff between false positives, false negatives, and intrusiveness or oversight. Simply put, one can make an imperfect intervention more or less broad, or one can invest in making the intervention more accurate, by some mix of investing more energy in scrutinizing individual cases and requiring more access to information about those cases, some of which might otherwise be private.

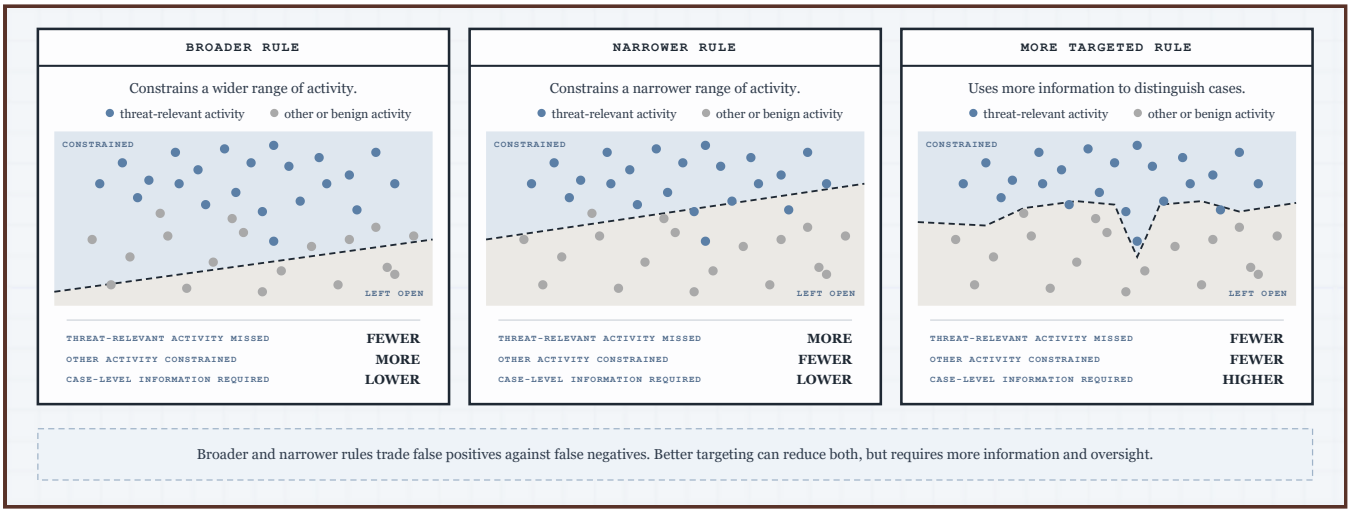

Figure 6: The tradeoff between false positives, false negatives, and intervention intrusiveness.

Broader interventions, which capture a higher fraction of threat-relevant activity, carry with them greater economic costs, while narrower interventions may be easier to circumvent and thus have fewer safety benefits.

Alongside interventions which directly cap or limit some resources, it is possible to intervene indirectly: to change an incentive and let the selection be performed by the actors themselves. This includes buying out or taxing specific resources like compute,

or applying stricter liability for outcomes to model providers or other actors with control over capability access, such as cloud hosts. These approaches can effectively draw on private information no regulator could extract: private plans, valuations, and alternatives. Ultimately this means indirect selection needs no surveillance to operate and thus is less intrusive. However, liability-based systems can suffer from other issues, such as pushing the frontier toward more risk-tolerant actors, and systematically underweighting risks where the costs are catastrophic.

The correct balance between tolerating false positives, false negatives and intrusiveness will depend on the goals of the intervention, the substitutability of inputs, and the resources (including political capital) available. The aim of an intervention is often to buy time to do work to mitigate a particular threat, which may risk being caught in the filter. For example, a filter which blocks cybercrime requests will sometimes prevent evaluation requests and defensive requests, undermining the nominal aim of mitigating the risk of cyberattacks. If malign work and safety work draw on the same capabilities, it is difficult to only target the malign activity. Some routes around this problem include whitelists for trusted organizations (e.g. see Anthropic's Project Glasswing, where access to Mythos Preview was gated to selected partner organizations only), but these carry other concerns, such as unfair distribution of access.

Since essentially all control surfaces suffer from these tradeoffs, often the best route to a given outcome will involve several different control surfaces working in parallel, such as a cheap intervention with few false positives and many false negatives, coupled with a more demanding one that can catch some of the false negatives which slip through the previous one. Relatedly, in fast-moving situations it may be preferable to quickly enforce conservative interventions with many false positives, and use the breathing room to implement more careful and calibrated ones.

3.3. Coordinated interventions

In cases where an intervention requires coordination among multiple actors, some control surfaces and interventions make this easier than others.

Interventions and actions must be agreed upon by parties to the agreement, and they must have a shared understanding of which actors and activities are covered, so that they know what they are agreeing to, and how compliance will be verified as well as the consequences of noncompliance. In some cases agreements may cover actors who are not party to the agreement—for example, a chip export control rule agreed by the US and Taiwanese governments would likely have significant effects on Chinese developers who are not party to it.

Coordination is more sustainable when compliance can be checked without needing to rely on participants' goodwill. This makes control surfaces which enable third-party verification especially desirable. Coarse, numeric thresholds like FLOP counts are easier to evaluate than holistic appraisals like bio uplift capacity, and so it may be easier to coordinate around them as interventions.

Unfortunately, while information about model capabilities and threat models accumulates steeply along the arc of model development, coordination can become more difficult. The information that can be gained about a system nearing deployment is often nonstandardized and difficult to make legible (i.e. the internal testing protocols and mitigation mechanisms at one lab may not look the same as at another, even if they point at the same targets). Upstream, however, information on goods like compute, which are produced by very few companies in relatively few places, can be easily made legible, verified and shared, making coordination easier. As a result, coordinating actors will tend to select upstream control surfaces: the crude and high-false-positive side of the arc provides a source of surfaces better suited to coordination.

3.4. Case A: A cap on frontier training

Suppose the concern is that AI systems may substantially accelerate or automate AI R&D before adequate means of control exist. The hazard is a transition toward AI-driven development which outpaces developers' ability to control it.

Control surface: Training compute is an early, rival input into this process, and therefore represents an attractive control surface for this problem.

Here we choose a per-model cap covering pretraining and post-training. Developers' aggregate research compute remains uncapped.

"Capping training compute" still requires further specification of the control surface, however. How should experimentation count toward this, if at all? If novel methods involve combining the results of multiple training runs into a larger model, will this still fall in scope? How does one deal with the problem that further algorithmic progress and substitution of inputs will allow a given level of capabilities to be achieved with less compute as time goes on?

Targeting and tradeoffs: A compute ceiling only bears on one aspect of the problem, and does not address capability gains from other routes, such as better data or RL environments, or increasing test-time compute usage. It may also be too broad, and stop some benign or benevolent work.

Coordination: Compute usage is a comparatively legible surface, if actors can be required to share records. Many proposals for governing compute usage and tracking chips exist; it appears the technical problems involved are feasible to solve, but more work is still needed before this can be implemented. The coordination required involves figuring out who has authority over verifying compliance, who is included under the scope of the policy (which individual developers, and which nations), and how to respond to capability gains by non-participants so that continued participation remains preferable to defection.

3.5. Case B: Restricting access to dangerous biological capabilities

Suppose instead that the concern is the diffusion of AI assistance that materially increases users' ability to develop biological weapons. Here the hazard is more about access and usage than training.

Control surfaces: Model deployment and access, gated by capability evaluations and subsequent restrictions. Post-trained or API side restrictions, such as refusing prohibited requests, are one potential approach, while

another is attempting to train the model not to be capable of fulfilling such requests in the first place, perhaps by ablating training data to remove biological knowledge. The latter, if successful, also has the advantage of being somewhat more robust to open-weight models having their restrictions lifted with further post-training, which is a significant risk in the former case.

Targeting and tradeoffs: Attempting to make the model refuse noncompliant requests faces three main issues — over-refusals, under-refusals and jailbreaking. Decisions must be made on how jailbreak-resistant the model needs to be, and how much refusing of benign requests is tolerable. The capabilities evaluations themselves must also be sufficiently comprehensive to make sure paths to dangerous capabilities are well guarded.

There is also the question of whether it is desirable for specific actors, such as those expected to use models defensively, to have access to versions with lighter restrictions, and how to gate this access reliably. This may improve targeting, but also imposes some compliance and privacy costs.

In cases where usage restriction is the target, there are more obviously tangible benefits to even imperfect restrictions. The more effort and technical capability it takes to circumvent guardrails and do something dangerous, the less often you'd expect harmful behaviors to occur.

Coordination: The control surface here is less widely legible than in the compute cap example, and relies upon the existence of a trustworthy evaluator with sufficient expertise to make determinations. Participating governments would need to agree on evaluation standards, and recognize qualified evaluators. Difficulties again arise with what to do about non-covered actors, though in the case of narrow restrictions on usage compliance may be less costly for model providers than with broad restrictions, since the majority of their customers might see limited gains from access to these capabilities anyway, and exceptions could plausibly be tailored for legitimate cases.

There is also a risk that, once an open model with a given level of capabilities exists, subsequently attempting to control it may achieve substantially

less and may not justify ongoing costs, so the policy may be brittle in the long run.

3.6. Open Research Questions

- **How should hazards be translated into covered activity for pacing interventions?** Risks we would like to target, such as uncontrolled automation of AI R&D and bioweapon uplift, build up over various stages of AI research, development and deployment; capabilities will initially emerge at some point in training, and we may want to avoid such a point being reached, or we may care more about wider deployment (especially if capabilities have positive use cases we want to preserve). Research should compare candidate boundaries.
  - *Shevlane et al. (2023), Model Evaluation for Extreme Risks*
  - *Hooker (2024), On the Limitations of Compute Thresholds as a Governance Strategy*
- **How can pacing thresholds be made specific and yet still cover distributed activity?** A pacing intervention targeting a threshold could potentially be circumvented by distributing activities or artifacts such that each sits below the threshold. How can we design aggregation rules and methods to handle cumulative risk from activities that are divided across space, time, processes and entities, without hindering low-risk activities?.
  - *Seferis & Fist (2026), Detecting Compute Structuring in AI Governance Is Likely Feasible*
  - *Rahman (2026), Does Distributed Training Undermine Compute Governance?*
- **What practical coverage is sufficient?** If we consider the reach available through company control, infrastructure providers and national rules, including their supply-chain effects, can we estimate bounds on activities that would be effectively covered by an intervention and relevant activities that would be missed?

    - *Koopmanschap & Barten (2026), How to Catch a GPU*. Maps issues with enforcement coverage as dangerous capabilities come to require progressively less compute
    - *Egan & Heim (2023), Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers*
- **How effective are different access restrictions once a dangerous capability has been released?** How do factors such as access guardrails, alignment training, access to inference compute, ease-of-use, and tacit knowledge affect risk once diffusion has already occurred?
    - *Tamirisa et al. (2024), Tamper-Resistant Safeguards for Open-Weight LLMs*. Identifies limitations in how reliable safeguards can be for open weight models.
    - *Ord (2025), Inference Scaling Reshapes AI Governance*. Identifies inference scaling as a lever for released models.
- **How can we permit exceptions to allow useful work, without being so permeable that it makes the rule useless?** In the case of compute controls, it now seems technologically feasible, to some extent, to identify what uses a GPU is being put to. What other technical advances can allow interventions to be less blunt and more narrowly scoped?
    - *Gargiulo & Kulp (2026), Workload Identification with Physical Side Channels for AI Governance*
- **What are the tradeoffs between verification and invasiveness for different interventions?** How can we push the frontier forward?
    - *Scher & Thiergart (2024), Mechanisms to Verify International Agreements About AI Development*. Gives an overview of different verification mechanisms and what they require.
    - *Petrie et al. (2025), Flexible Hardware-Enabled Guarantees for AI Compute*. Proposes verifying compliance without exposing sensitive information about AI development.

4. Pace How?

A pacing intervention is ultimately a sequence of steps; we reason about an intervention from start to end and consider its many decision points and failure points. This lets us spot the supporting work required for an actual, sustained period of restraint. In this section, we consider a single intervention's lifespan. In reality, a pacing plan may contain a portfolio of interventions, each with its own conditions and triggers.

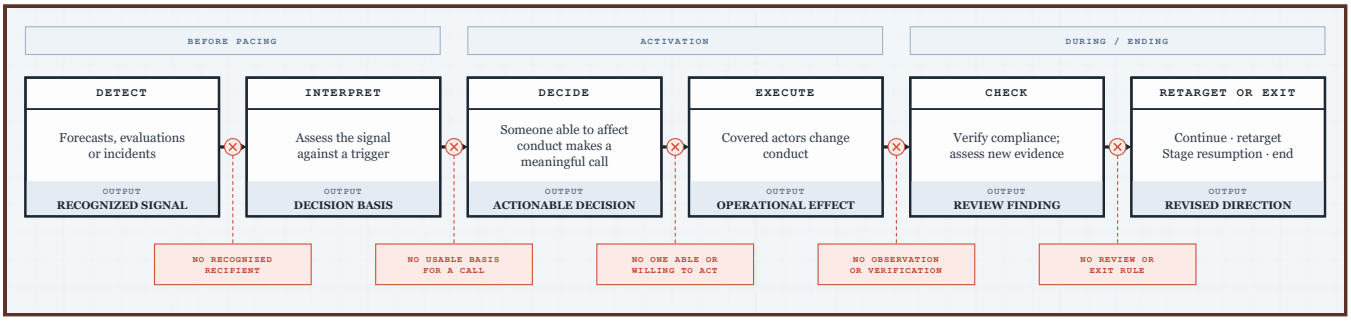

Figure 7: How evidence can fail to influence pacing action.

4.1. Before pacing

The main challenge before intervening is recognizing where there's a need to intervene and how to time it. Intervening too late means letting the threat play out with potentially irreversible consequences; intervening too early means sacrificing potential benefits and political capital, and in some cases being less able to carry out complementary activities that depend on access to advanced AI (accelerated safety research, empirical study of emerging threat models, etc.), or to the benefits from conducting the complementary activities in a setting with more talent, capital, and data from deployment settings.

The actual timing of a pacing intervention will ultimately depend on the strategic and political calculus of the actors who can bring it about, and may have more to do with political opportunity, public perception, salient incidents, or campaigning efforts. The

timing ideally will be informed by evidence about technological progress, and the threat that pacing is meant to address.

Sources of information to draw on:

1. **Model evaluations**, including benchmarking, red teaming, and other tests in simulated or controlled environments, provide direct information about the capabilities and propensities of specific models in specific circumstances, and can form the basis for extrapolation. That said, the correspondence between evaluations and real-world risks can be quite fraught and hard to predict: the jaggedness of AI capabilities means that individual evaluation performance often corresponds poorly to behavior on tasks in the wild. Furthermore, evaluations can generally only set a lower bound on capability, because of the challenges of elicitation and the risk that models alter their behavior in response to recognizing they are being evaluated.
2. **Forecasting** and effective predictions provide decision-makers with aggregate expert assessments of when dangerous capabilities are likely to emerge. Forecasts can be about the threats themselves, or they can be about trends in the AI R&D ecosystem that inform threat assessment. Different forecasts shed light on different parts of potential dangers: for example, trends in training resources can be used to forecast the size and nature of the resulting required physical build-outs, while capability forecasts can predict models' future behavior. A major open problem here is selecting epistemic peers: efforts to use top generalist forecasters for predicting even near-term AI have a mixed record.
3. **Incident reports** provide decision-makers with concrete evidence with ecological validity and lots of useful surprising detail, properties which both forecasting and evaluation often miss or fail to attempt. This is the best indicator that there is an immediate AI-induced problem, but by the time an incident has occurred, the space of possible pacing interventions has already shrunk. But even relatively harmless incidents can expose unknown unknowns and more granular details, which can in turn reveal underexplored threats. Also, compared to evaluations and forecasting, real-world incidents can be more credible and legible to a broad range of actors, but they are bottlenecked on effective infrastructure.
4. **Safety cases** provide structured arguments that a system will remain within some acceptable risk parameters as long as certain prior conditions are met — for

example, that certain features of a deployment context guarantee that a given harmful model capability cannot be accessed. This structure therefore makes it possible to turn a discussion of potential risk into a discussion of concrete underlying properties which can be directly scrutinized. In regulated safety-critical industries, safety cases are mandatory prior to large-scale engineering efforts, making them effective pacing interventions: the work cannot go ahead until the evidence is gathered and compiled into a safety case that would satisfy a regulator, and the complementary activities required to support a safety case are directly targeted at the threats the safety case addresses. Despite interest, safety cases are not mandatory in frontier AI R&D, and few labs have generated them voluntarily.

Evaluations, forecasting, incident reports, and safety cases perform different functions: respectively, to test certain (bounded) claims, to supply warning signs, to reveal models' emerging behavior, and to decompose threats into specific precursors. Alongside feeding into the question of whether to directly intervene, they can also feed into each other: incident reports can shape what gets evaluated, evaluations can form the basis for forecasts, and so on. Collectively they can provide a sense of how far away any given risk is, and this in turn can help actors to determine how urgently they need to prepare for costly measures.

Coordinated pacing comes with extra challenges: the different actors need to be able to converge on their understanding of the relevant information and what constitutes meaningful evidence, which in turn typically requires some bandwidth between them. In situations where they have competing incentives, some will also have individual incentives to withhold information. For example, frontier developers might not want to be overly restricted by governments, and so they might not want governments to have the information that would warrant such restriction. Those considering pacing interventions may need to think about what infrastructure needs developing in advance to get around such dynamics.

4.2. Triggering and enacting pacing

The challenge in deciding to trigger a pacing intervention is negotiating the tension between speed, legitimacy, and precision. It is easy to have two but more difficult to

have three:

- **Speed + Legitimacy**: e.g. preregistered rules that automatically trigger when a condition is met. Unfortunately it is hard to know what exactly the rules should be, and so they may end up mistargeted in unexpected situations.
- **Legitimacy + Precision**: e.g. careful deliberation between experts, followed by democratic authorization of agreed-upon interventions. Unfortunately the review and implementation process here can take a lot of time.
- **Precision + Speed**: The actors with the most information and expertise get to choose by fiat. The risks here are that actions which are difficult to explain may seem illegitimate and thus at risk of being undone more easily, and these actors will often face incentives around the pace of progress that diverge from those of the broader public, making it difficult to trust they are truly acting in citizens' best interests.

We can see this trilemma play out across the whole ecosystem of AI progress: governments have legitimacy and means to impose strong regulations but not necessarily the raw information needed to guide decisions, or the expertise to interpret such information, leaving them lacking in precision and/or speed; `external evaluators` can have enough precision (through expert staff) but not the mandate to regulate; AI developers have the most information about their own progress but little reason to proactively internalize any negative externalities they produce, and a lot of other interests in how their competitors are regulated (we return to this in `§5`).

One way to ease the tradeoffs is to work toward a portfolio of triggers. For example, authorities could set very crude evaluation thresholds now, beyond which they give expert groups like third-party evaluators the right to unilaterally trigger emergency interventions, on the condition that those choices then be subject to later governmental reconsideration, where there is more of a mandate but less ability to rapidly deploy expertise. This approach might seem risky if one views intervening as a one-off affair, but once it becomes a repeat affair, the evaluators would have a long-term interest in using their powers in a way they could justify.

The next decision is execution: the intervention must somehow affect live systems. Implementation-wise, the key question is whether enacting the intervention merely involves announcing a rule, or whether it involves some more active steps like buying

up, confiscating, or destroying key resources. The former case is more straightforward, but brings with it the extra challenge of enforcing the new rule.

Again, coordinated pacing comes with extra challenges. It will typically take a lot more time for several actors to form a consensus on whether an intervention should be triggered, especially if there is a range of options to choose from. One potential solution is to give several parties the ability to unilaterally trigger time-bounded interventions which can buy time for more careful discussion. Another is to spend time in advance mapping out the likely space of mutually beneficial interventions.

4.3. During pacing

During pacing, it can be difficult to sustain the efficacy of an intervention in an ever-evolving environment. There are four notable challenges: whether the relevant parties are still complying; whether the initial threat is still effectively targeted; whether the threat is still a threat; and whether the suspension in activity is being used effectively.

Authorities can address the first problem with monitoring, verification and enforcement systems. The second component is whether the control surface continues to track the hazard (i.e. whether complying with the terms of the intervention actually diminishes the risk). §3 describes this relationship in terms of false positives and false negatives and gives more detailed considerations. Typically, a static rule will become less effective over time as the practical implications of the rule drift from their initial state (e.g. because redirection of effort toward other avenues of achieving the same goals renders the intervention weaker).

The third component is whether the original reason for intervening still holds. The concerning capability may mutate, diminish, or intensify. New evidence may also undercut the rationale for designating a result as a threat. The strategic and commercial reality may evolve too. Companies may exit agreements and safety measures may strengthen, for example. Importantly, this does not bear on the effectiveness of the rule itself, but on whether the threat targeted by the rule is still of concern.

The fourth component is whether the interval is being used to change the conditions which made pacing valuable, as discussed in §2. This could simply amount to some ambient societal adaptation, but typically it will involve deliberate complementary activities aimed at ensuring that the threat will be less severe after the intervention than before. Absent good enough complementary activity it is possible for an intervention to merely postpone a growing crisis.

There is a danger that pacing weakens the infrastructure it is designed to strengthen. Sweeping restrictions on AI, for example, would likely shrink the opportunities available to the actors involved in frontier systems, potentially limiting their experience and understanding of frontier models. Mandatory evaluation by third parties may increase the risk of leaks and industrial espionage. To mitigate this outcome, a limited and vetted group could be permitted to continue research: a tight circle of knowledge and power may limit the threat potential while maintaining expertise. A significant risk is that this group would gain outsized power and influence, or be unaccountable and lacking in error-correcting mechanisms (see §5.3 for more discussion of this dynamic). The nature of the complementary activity to the restriction (typically bolstering safety procedures in some form) can inform decisions related to the chosen group's composition, focus, and oversight. When reviewing the intervention, the same considerations as in activation are in play: the tensions between speed, legitimacy and precision.

A potential solution could be to separate retargeting by scale and reversibility. Operators could have the authority to make temporary and bounded adjustments; external reviewers could implement wider and more durable policies. As in the activation procedure, tradeoffs cannot be wholly eliminated, only managed across the model lifecycle.

4.4. Ending pacing

A major difficulty with any kind of temporary intervention is ending it at an appropriate time. Restrictions can become entrenched and persist past the point they are no longer justified, or can be abandoned while other actors are relying on them. So interventions which depend on precise exit conditions are less attractive than plans which are more robust to poor exit management.

There are two clear reasons decision-makers might wish to exit a particular pacing intervention. Where an intervention worked and is no longer needed, this can be because its initial goal to build up defensive capabilities or resilience has now been achieved, and the risks involved are in fact mitigated. An intervention which is instead no longer working could be addressing an obsolete pathway to impact which has been routed around; it may have had a misspecified theory of impact in the first place; it may have insufficient teeth to ensure compliance from relevant actors and lack the support necessary to win enforcement; or it could simply have been superseded by a subsequent intervention.

The specifics of the case determine the best exit structure. After successful pacing, authorities should aim to reduce the costs the intervention has imposed while preserving the gains made. In the failure case, the goal might be to simply remove the costs while minimizing disruption. The former case may require a more careful, gradual process than the latter, since the potential damage from getting it wrong will be higher.

Exiting too soon may expose society to the very risks that pacing was meant to prevent, while having burned goodwill and support for further intervention. Exiting too late means imposing costs and burdens for longer, and strengthening the relative position of noncomplying actors.

The defaults also matter: if the intervention is set to expire after a certain period and requires renewal, it may be prematurely exited without regard to the motivations which installed it. If it is installed with no definite end point it may stagger on past the point where it is useful, with institutions and operators becoming entrenched, developing an interest in preserving their function.

Expectations around the removal of interventions also shape actors' willingness to participate in the first place. Actors may agree temporary restraint will be useful, but refuse to join an agreement if they doubt it will end when conditions are met, or expect the arbiters of exit decisions to be biased against them.

These expectations also shape behavior during the intervention. If exit is conditional on demonstrated progress in a particular domain (e.g. on demonstrating models do not suffer particular failure modes), then the incentive to work specifically on

that progress is greater than for an intervention which will end simultaneously for all actors.

Where abrupt exit from an intervention could pose a risk, authorities can try to mitigate the dangers of an exit by staging it. This could involve continuing some forms of restrictions, and gradually relaxing requirements over time. Evidence and research as to how to proceed can thus be safely collected, while still managing the risk. Staging thus permits actors to observe incremental effects of relaxation, and make decisions about whether this is desirable or ought to be halted.

Ending an intervention need not mean dismantling it entirely. It could make sense to retain some capacities that are slow to build (expertise, relationships, technical standards, communication channels, etc.) while lifting the biting aspects, such as invasive and exceptional controls.

Stage	Decisions
Pre-pacing	• What justifies the need to intervene? • What signals of emerging risk are being tracked? • What infrastructure must exist for relevant signals to be detectable? • Who watches for these signals? • Who interprets signals and who do they report to?
At trigger	• Who holds the authority to decide that a trigger condition has been met? • How much error is acceptable in order to act quickly? • What sequence of actions does triggering set in motion? • Is the developer obliged to address the triggering concern, or free to abandon the blocked path?
During intervention	• How is compliance observed, verified, and enforced? • Who reports evidence of compliance, who audits, and who acts on discrepancies? • How is the downtime being used to respond to the threat? • Who, if anyone, may continue the restricted work, and under what oversight?
At exit	• How do decision-makers distinguish an intervention that has served its purpose from one that has failed or become obsolete? • Should exit be immediate or staged? • Who is exposed to the effects of exiting, and who bears any costs or receives any gains?

Table: Key decisions per stage of an intervention's lifecycle

4.5. Case A: A cap on frontier training

Before pacing, governments should recognize that they are hoping to slow development before it outpaces developers' and authorities' oversight capacity. Thus they must make judgment calls about when developers are sufficiently close to the dangerous threshold to justify intervention. A problem is that the predictive tools they could use to detect relevant signals are hard to interpret: to take one scenario, some forecasts may predict imminent self improvement, while others disagree.

Triggering: A source of authority must be determined in advance to establish when to begin intervening. Given the speed of progress currently and the uncertainty over further acceleration if a threshold of recursive self improvement is met, there may be little time for consultation at the point of triggering, and under our taxonomy above this suggests that either legitimacy or precision must be deprioritized — there can be clear rules for when to trigger, or a source of authority can be granted power to trigger by fiat, but a long and deliberate consultation at the point of triggering would be intolerable. Pre-agreed triggers could be used to impose emergency temporary restrictions, to allow time for more deliberative review.

During pacing, authorities should verify that the developer is complying with the intervention by, again, monitoring the inputs. Continued access to compute may be contingent on the developer granting third-party evaluators visibility into every run, which can be evidenced by logs of compute usage obtained from the compute provider. Since a compute limit does not constrain every route to capability, authorities may have reason to seek visibility into research teams' working logs, to help determine if current limits are fit for purpose or need to be adapted.

Ongoing reviews could be conducted to establish whether pacing is still warranted, and to maintain accurate working knowledge of the progress of improvements in control and safety.

Ending pacing here is uncertain. If the cap is buying time usefully, then exiting may be desired once humanity has sufficient assurance that it can

avoid loss of control, or that the benefits of development exceed the risks. Progress by non participants may also change the calculus regarding whether persisting with restrictions is beneficial. Staged exits could be used to minimize the risks of sudden jumps caused by a capabilities overhang.

However, if there does appear to be a hard capabilities threshold under a certain compute threshold, insufficient progress is made on assurance of safety, and the controls are sufficiently widely adopted, it may be that it would be desirable for these restrictions to persist indefinitely.

4.6. Case B: Restricting access to dangerous biological capabilities

Before pacing, labs may be required to make their models available to third-party evaluators for pre-release evaluations of potential capabilities that would provide uplift to a malign actor (e.g. to debug a failing synthesis protocol, or piece together a dangerous method from scattered dual-use sources). A biological capability is far harder to recall once it reaches the public than to withhold beforehand, so reaching a certain threshold on the evaluations should block deployment outright, unless safeguards can be demonstrated to mitigate the risk. This is particularly important for open weight models, where release is difficult to reverse.

Triggering here will depend on the uni- or multi-lateral shape this intervention takes. So far, similar cases (such as the US government intervening in the `Fable` and `GPT-5.6` releases) have been ad hoc and messy, but were enacted quickly. A more legitimized, legally grounded process would likely require a publicized framework to make clear what is in scope and better allow predictable deployments and to ensure consistent standards to be applied internationally, though it seems very likely that the capacity to ad hoc prevent the release of an otherwise out-of-scope model will persist.

During pacing, monitoring bodies should audit access logs to confirm adherence to access restrictions, and confirm that any investigation is led by

independent evaluators qualified to judge biological uplift, rather than the labs themselves. Ongoing monitoring to ensure latent capabilities are not easily elicited by jailbreaking methods may also be part of this puzzle.

Ending pacing may not involve any form of public release. As a condition of restricted release, labs may need to prove that even where a model retains a dangerous capability, there is the capacity and will to reliably vet their users and flag suspicious activity to the authorities.

4.7. Open Research Questions

- **How do the incentives of bound parties change across the lifecycle?** To what extent can different actors reliably predict the behavior of other actors throughout the lifetime of a pacing intervention? How load-bearing are these predictions of behavior going to be for coordination of pacing interventions?
  - *Finke (2026), International Agreements to Limit Frontier AI: Objectives and Exit*
  - *Koremenos (2005), Contracting around International Uncertainty* .
  - *Goldstein and Salib (2025), How to Stop an AI Arms Race* .
- **How does information sharing impact the credibility of coordinated pacing?** Which information, in what granularity, through which channels, by which actors, matter most for credible coordinated pacing? How can this be kept compatible with national security considerations, commercial confidentiality, and cybersecurity? To what extent can this information sharing be kept robust to manipulation?
  - *Wasil et al. (2024), Verification Methods for International AI Agreements*
  - *Scher et al (2025), An International Agreement to Prevent the Premature Creation of Artificial Superintelligence* .
- **What evidence could legitimize a speedy initiation of pacing? What is likely to be the acceptable tolerance for unreliable evidence for different actors?** Can scenarios and thresholds be specified in advance

to a level of specificity that would garner coordinated buy-in to a rapid pacing onset? Can evidential and assessment processes be agreed on in advance?

  - *Karnofsky (2024), If-Then Commitments for AI Risk Reduction*

- **How could dry-run simulations inform and prepare for pacing interventions?** What aspects of simulation design, delivery and follow-up affect their effectiveness and impact? Could simulations harm or misguide pacing interventions?

  - *Gruetzemacher et al. (2024), Strategic Insights from Simulation Gaming of AI Race Dynamics*
  - *Bartels (2020), Building Better Games for National Security Policy Analysis*

- **How will our reliance on different sources of information about risks, and our methods for communication and coordination, change as AIs become more capable, autonomous, or integrated?**

  - *Clymer et al. (2024), Safety Cases: How to Justify the Safety of Advanced AI Systems*

- **What does the space of exit scenarios look like?** For example, small-scale interventions may allow for immediate release, whereas exits from ongoing large-scale interventions impacting multiple facets of the economy and society (e.g. export controls) may need to be staged. Other than staging, what other parameters of exits exist, and how could they be tuned?

  - *Finke (2026), International Agreements to Limit Frontier AI: Objectives and Exit*

- **What is the relationship between initiation and exit triggers?** Intuitively, the exit trigger should track whatever justified the intervention in the first place, but what could affect that link and what other factors should be considered?

  - *Cârlan et al. (2024), Dynamic Safety Cases for Frontier AI*
  - *Karnofsky (2024), If-Then Commitments for AI Risk Reduction*

5. Then What?

AI R&D is a major industry, with huge amounts of financial and political capital invested in its trajectory. Changing the pace and emphasis of R&D will therefore have consequences beyond the direct effects of changing pace and of its supporting activities.

In this section we consider these impacts, because they matter in themselves and because they impact the willingness of all actors to support or resist pacing, and therefore the actual net effect of pacing. Their likely responses to rules should therefore be part of the specification, and should be modeled and elicited before any rules are fixed.

We consider these impacts radiating outward from the pacing intervention through society, and forward in time from the enactment of the pacing intervention.

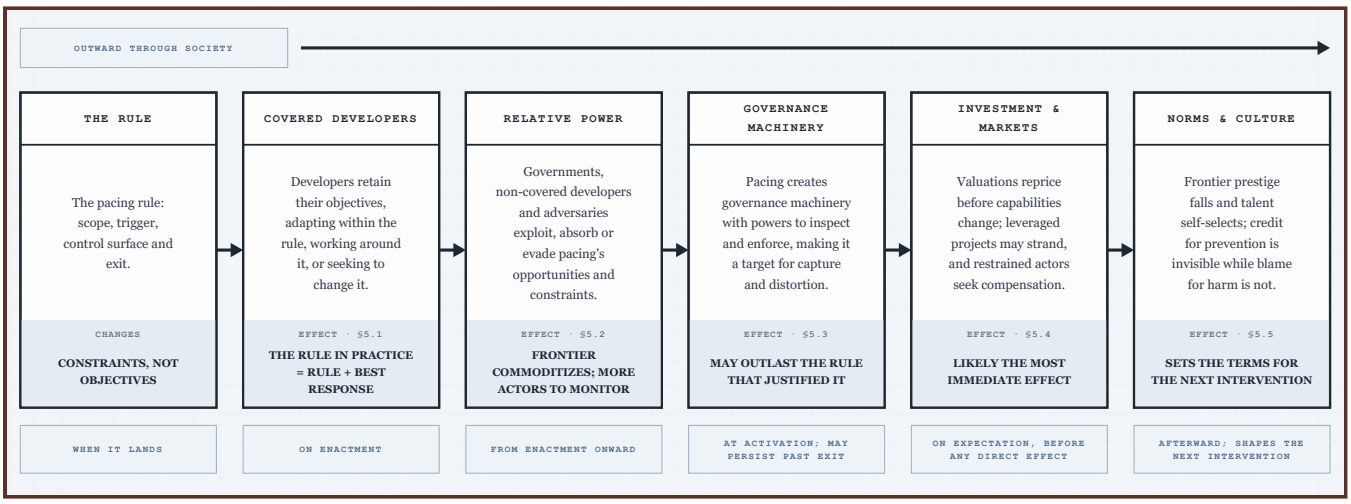

Figure 8: The effects of a pacing intervention radiating outward through society and forward in time.

5.1. Covered developers

As the center of AI progress, frontier model developers are the most directly impacted by pacing interventions. Intervention changes the constraints these developers face,

but not their objectives: the competitive pressures described in §2.3 persist, and developers will re-optimize against the new rules. The practical effect of an intervention is therefore the rule, *plus* developers' responses to it. But we can anticipate the form of such responses from earlier attempts to restrain arms races: for instance, when the 1922 Washington Treaty capped naval cruisers at 10,000 tons and 8-inch guns, the signatories built "treaty cruisers" that sat exactly at those limits.

In response to pacing, developers will likely consider:

5.1.1. Adaptation within the rule

Reallocating R&D resources. The appearance of new constraints induces innovation. Capping training compute would (further) raise the value of algorithmic efficiency, data quality, post-training, inference-time compute, and elicitation. Developers will move researchers and budget accordingly, with unclear effects on capability growth.

Many pacing interventions would reduce the need for and ROI of holding and building massive compute assets (see §5.4). With reduced resources (or expected resources), developers could also reduce the amount of safety research they conduct in a given period. If exit is explicitly conditioned on the developer's safety progress, then safety research becomes the developer's route back to scaling and will be prioritized accordingly. The design of exit conditions in §4.4 will therefore determine where reallocated resources actually go.

Re-evaluating commercial strategy. If the frontier stops moving, the basis of competition shifts. If the frontier moves less, competition moves from capability to price, latency, distribution, integration, and post-training. If roadmaps stop assuming that a new model generation will arrive every few months, the breadth and depth of deployment might increase, as there is more incentive to embed the current models everywhere. Diffusion could thus actually benefit from pacing frontier development.

Compliance overheads. Developers already have large compliance functions owing to existing regulations not specific to AI, but many pacing interventions impose large amounts of work on developers. They may have to document all training runs, run intense evaluations at each checkpoint, host auditors, and await legal review for deployments. This is both a burden and a source of a potential moat: the fixed cost of establishing such a function is a barrier to new entrants. The GDPR is a recent

precedent: market concentration among web vendors rose 17% after enforcement, with the largest vendors gaining share.

Internal power shifts. Pacing can change the relative power of departments inside each developer. For instance, the prestige of research units could fall in favor of product-focused units. This can strengthen safety functions, but also risks turning them into compliance functions optimized for demonstrable adherence to the rule rather than for reduction of the underlying risk.

Re-evaluating partnerships. If compute providers become enforcement points (see §4.5), a developer might opt to integrate vertically, building its own datacenters or moving to compute providers in jurisdictions where it expects favorable enforcement.

Developer relationships with governments also change: developers may seek national champion status or exemptions for national security work (of the kind already visible in the trusted-access programs described in §2.2.1). This gives the state a stake in the developer's continued progress, and creates incentives which could blunt the state's willingness to pace it.

5.1.2. Working around the rule

Jurisdiction shopping. If the pacing mechanism is both legally binding and not global, then developers have discretion to move activity to less-restricted jurisdictions.

Rival, upstream inputs are hard to move: frontier compute is costly and slow to relocate, its supply chain runs through a handful of firms (see §3.1), energy and datacenter siting are slow, and export controls already exist as a counter-response from the state.

Non-rival, downstream activities are easier to move: research methods, model weights and other intellectual property can be moved easily and a model developed under one regime can be served into another. For Case B, the relevant version is a release of weights by an affiliate or partner outside the covered jurisdictions. The more an intervention targets rival goods, the less jurisdiction shopping it permits, which reinforces the pull toward upstream control surfaces noted in §3.3.

Exploiting loopholes. In §3.2 we cover the inevitable gaps between the control surface and the targeted activity. For instance, if the intervention involves restricting total training compute, then driving compute efficiency up becomes (even) more valuable, so that the developer can continue R&D activity while not defecting from the letter of the intervention. Developers will systematically search for such gaps, likely bringing more expertise than the regulator has.

Firms bunch just below regulatory thresholds wherever these exist and there is no reason to expect training runs to differ. Much of this is not bad faith but ordinary engineering under a new constraint, which is why rules based on intent are hard to enforce. The regulator typically learns of a loophole after it has been used, which is one reason §4.3 requires continual retargeting.

Defecting from pacing. For a variety of reasons, the developer might not pace their R&D, or stop pacing after initially complying. This could be done openly, or secretly.

Secret defection is worse than non-participation because it corrupts every other party's picture of the state of play. The incentive to defect is not constant: for a training cap, it grows with the overhang (see §4.4), so verification needs to be strongest late in the intervention, when political attention has moved on. The choice between open and secret defection is set by the probability of being detected and the penalty for defecting (see §3.3, §4.3).

5.1.3. Working on the rule

Policy shaping. Developers hold the information the regulator needs and will supply it selectively. The developer is also likely to actively influence the pacing intervention through lobbying or by their direct representation in the pacing governance mechanism. This is to ease the regulatory burden on themselves or increase the burden on their competitors. The more regulation they are subject to, the more they will be driven to spend on lobbying.

Shaping can be prosocial: developers may plug loopholes in the monitoring mechanism which only become apparent after enforcement begins, or correct a control surface that is missing its target. But the regulator cannot easily tell this shaping from the other kind.

Seeking compensation. Those who are affected by interventions may make efforts to secure compensation for losses they incur. See §5.4.2.

5.2. Shifts in relative power

5.2.1. National governments and international competition

Consider national governments and international organizations. Many governments see AI as a strategic technology, and a pacing activity would potentially impact them in several ways.

For nations which are currently behind in the AI race, any given pacing intervention could appear as either an opportunity to catch up, or as impeding the path by which they already expected to. A frontier-only training cap might be the former, while broader restrictions on chip supply might be an example of the latter.

Either way, they will likely seek to exploit the pace to improve their relative position, by funding activities differentially benefited by the pace, or by trying to evade the restrictions.

They may also try to slow down adversaries, whether by attempting to include them in coordinated pacing interventions, or arguing that non-parties to the agreements should be somehow constrained (e.g. by forbidding purchase of compute by actors who are not party to an agreement).

The machinery of pacing will also be vulnerable to government actors (see §5.3), where whatever machinery is required to implement the interventions may face pressure to be co-opted to serve the state's interests.

5.2.2. Non-covered AI R&D actors

Consider also non-covered AI R&D actors and the broader AI research ecosystem. While many AI R&D actors may not be carrying out activities targeted by the pacing intervention, they can also be impacted by pacing.

If pacing reduces the amount of investment and talent flowing to pushing the frontier of AI R&D (see §5.4), then these actors will become more attractive places for talent and capital to flow. Narrow AI uses which don't fall under the interventions' scope will be more attractive, and thus whether particular economic value is created by the application of a broad general-purpose model vs specialized systems designed for particular tasks might change.

Depending on who these actors are (which is a function of how broad or restrictive the pacing interventions are), this might lead to more actors reaching the current frontier, and more commoditizing of that level of capabilities. As this happens, however, it will increase the number of now-covered actors who might push the frontier by defecting or circumventing pacing restrictions and increase the burden on monitoring mechanisms.

There may also be a chilling effect from the existence proof provided by the pacing mechanism, whereby activities which are seen as plausible targets for future pacing become relatively less attractive, and leading to differential technological progress in areas seen as safer from future regulation. By this mechanism a pacing intervention could shape the direction of progress even outside its formal scope.

5.2.3. Adversarial actors

Many interventions (e.g. our worked example on biological threats) will also have to contend with misuse: third parties who use diffused AI capabilities to ill ends, for instance by developing biological weapons or committing cybercrimes. These actors can also be expected to adapt to any intervention, though primarily by trying to circumvent it from the outside.

5.3. Pacing governance structures

The main effect of pacing on the new governance mechanisms is to make them exist, or to invest new powers in existing bodies.

Many interventions will vest some enforcement power in a governing body or monitoring organizations (see §4). This will typically be some mix of access to private

information and discretionary enforcement power, and may even involve creating specific infrastructure to support observation or verification. The power and infrastructure can then sometimes be used for purposes beyond the rule that justified it.

Liability- or insurance-based schemes will distribute these powers across auditors and existing legal systems such as courts—similar concerns will exist around their access to information and the privacy of those subject to their scrutiny.

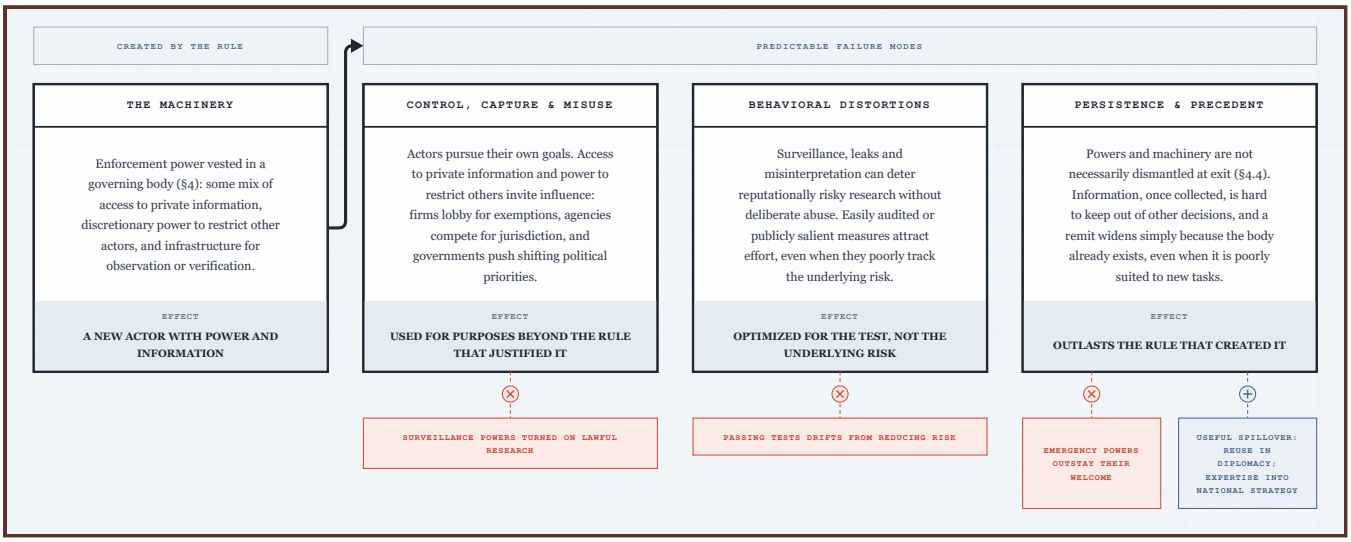

Figure 9: Failure modes in pacing governance structures.

5.3.1. Control, regulatory capture and misuse

In developing a potential intervention it is useful to start with the intended goal and then back out the details, but pragmatically it makes sense to plan around actors using power to pursue their own goals.

Once a newly empowered organization can inspect private information and impose restrictions on other actors, it will become a tempting target to influence. Companies will lobby for their own activities to be spared or minimally interfered with; other existing agencies will compete for jurisdiction over what they see as their own turf; and governments may push the body in particular directions depending on their shifting political priorities.

For example, powers of surveillance introduced to constrain dangerous activities might be used to interfere with lawful research to serve particular interests.

5.3.2. Other behavioral distortions

The existence of the machinery itself may cause shifts in behavior without any deliberate abuse. The knowledge that you are being surveilled and the potential for leaks and misinterpretations create significant compliance burdens and may discourage particular kinds of research which might be perceived as creating reputational risks.

The machinery can also change what targets participants optimize for. Measures which are easier to audit, or those which are publicly salient, may attract most of the focus, even if they only imperfectly track the underlying risk. Developers may focus on optimizing for passing these tests and drift from measures which would better serve to actually reduce the underlying risk.

5.3.3. Persistence, precedent and potentially useful spillovers

Additionally, it is not a given that these powers or the machinery that supports them will be dismantled at the end of an intervention, as we discuss in `§4.4`. This is not necessarily a bad thing: It might even be helpful if, for example, the systems created to enforce a domestic pretraining moratorium can then be reused in international diplomacy, or if departments created to assess model risks can feed their expertise into national strategy.

But there is a clear risk to contend with: crises can justify legitimately necessary emergency powers which then outstay their welcome. Once information has been collected, it is difficult to limit the scope of decisions it can feed into. The remit of a governing body may also broaden to new domains, simply due to the convenience of it already existing, even if it is imperfectly suited to a new task.

In general, how competently the governing machinery is run and how well received its actions are will affect its durability and popularity. The machinery should thus be thought of not just as a means of carrying out the intervention, but also as one of its effects, and predictable failure modes should be weighed against expected benefits in deciding whether to implement it.

5.4. Impact on AI investment and global markets

AI capital investments are already a major factor in the global economy. Investment in the inputs to AI progress (chips, datacenter construction, power infrastructure, and the AI developers themselves) is heavily contingent on the expected benefits and usefulness of the outputs, and so anything which affects the pace of capabilities progress will bear heavily on these investment decisions, and shift their attractiveness.

This effect can also be expected to be one of the most immediate of any intervention. Investment is a function of the expected future trajectory rather than realized changes, and these expectations could rationally shift long before any direct effects on capabilities occur.

5.4.1. Repricing and capital allocation

If investors predict a slower pace of progress than the counterfactual, it seems likely they will expect revenue growth for AI companies and suppliers to slow in aggregate (since the expected usefulness of their future outputs will be less). This would lead the valuations of developers and suppliers to fall, and their financing to become more expensive. This, depending on the intervention in question, might mean a slower datacenter buildout, and more capital flowing toward companies who take advantage of current capabilities rather than those pushing the frontier. In general this means more capital to flow toward those activities which remain less regulated or differentially benefit from the intervention.

To the extent frontier developers are currently valued partially based on the prospect of winner-takes-most dynamics (where AI labs grow to dominate the global economy), any pacing intervention which weakens their lead and relative position will weaken this, and weaken their expected market power, raising the relative power and value of supply chain companies instead.

An important aspect of the financial markets' reaction is that it will be proportional to the expected durability of the interventions. In other words, an intervention which is not seen as lasting long is likely to lead to a more muted market reaction than one which is expected to be durable. This makes markets an anti-enforcement measure:

their reaction and investors' losses will be greater for interventions expected to be more durable, incentivizing actors to push for weaker mechanisms.

Market reactions are also responsive to new information—and they do not need to wait for the intervention to take effect. This can create incentives to back out of interventions as soon as they are proposed, even before they take effect, if market reactions are unfavorable (see the April 2025 US tariffs, most of which were paused within a week of announcement after equities fell more than 10%).

5.4.2. Leverage and stranded investment

Much of the current datacenter buildout is financed with some degree of leverage (i.e. money borrowed against the value of the project, with the investment case supported by the prospect of a long-run return).

Changes in expectations can make these projects uneconomic or expose investors and lenders to losses, as the projects can no longer deliver the expected return. In an extreme case, a datacenter build might be abandoned partway through if the remaining cost of building it exceeds the expected value of the project.

Anticipating these effects and carefully designing rules can help minimize the damage done in these cases. It is also possible that compensation for affected actors might be part of any formal pacing agreement. The degree to which this is feasible depends on the scope of the intervention and the appetite from the parties to the negotiations.

Actors whose assets are stranded or whose expected revenue is cut will seek compensation. This could take the form of direct payments, buyout of their compute stock (see §3.2), tax treatment, government contracts, guaranteed access to restricted markets, or exemptions. Compensation reduces the incentive to defect, and could be framed as public procurement if the freed compute goes to safety or defensive research. But it has three costs: (1) it transfers public funds to the actors being restrained; (2) it creates a constituency with an interest in the intervention continuing, which bears on the exit problems in §4.4; and (3) it creates moral hazard: actors who expect compensation may invest in anticipation of a buyout.

5.5. Impact on norms and culture

And lastly, how does the pacing intervention reshape the norms and culture surrounding interventions in R&D, both future interventions in frontier AI R&D and interventions in R&D more broadly? If an intervention yields intended and desirable outcomes, is it making it easier for future regulations to land well? In other words, are "good" interventions being normalized? How much of a risk is there that pacing interventions in AI R&D might have a global chilling effect on growth and progress?

5.5.1. Talent culture

Pacing interventions may have knock-on effects on who wants to work at a frontier lab, and why, and on what kinds of work are most prized and rewarded by those labs. Restrictions on scaling and a weakened market position may deter those motivated by advancing capabilities and those with primarily financial motivations, and credible restrictions on perceived risky activities could attract those who might previously have been opposed on those grounds. Exit conditions could also raise the relative importance of some kinds of work—if permission to resume scaling is dependent on safety progress, for example, then safety work could be expected to rise in value. The "total research transparency" direction advocated for in AI2040's `Plan A` would likely be more appealing to the academically- or safety-minded, and less to those driven by financial gains.

On the other hand, if labs are seen as unethical actors, there could be stigma attached to working there, and this factor could lead to `departures` of many current staff, and deter prospective hires. If this effect is greater on those who are most concerned about lab practices, then internal scrutiny could be weakened. Which effects dominate will depend on the specifics of the interventions, but any change in employee composition would itself probably have an impact on the pace and trajectory of progress.

Talent impacts could extend beyond lab employees. A serious attempt at a pacing intervention, especially one involving the superpowers, will involve a large number of policy makers and diplomats, and may include risk professionals and advocates. A pacing intervention broadly deemed successful could increase the prestige and relative

power of those roles and professions, and may encourage pursuit of similarly-shaped victories in domains beyond AI.

Finally, we note that human talent may not remain the crucial chokepoint it currently is; OpenAI (for instance) is targeting a fully automated AI researcher in the next 18 months.

5.5.2. Political norms and culture

The consequences on the competence of bodies governing AI are non-straightforward and boil down to whether or not any interventions adopted actually prevent risk and, if so, whether the prevention is visible to the citizens.

Unfortunately, pacing suffers from the preventer's paradox (or "invisible-success" problem): when policy successfully averts an AI threat, the evidence of success is precisely that nothing happened—similar to counterterrorism efforts or the efforts to avert the Y2K problem, pacing policymakers may very well be doing their job precisely when nobody notices. This may mean that support for future interventions could hinge upon hypotheticals (while opponents of pacing have concrete capabilities to show).

Occasionally, successful interventions may be reactive (e.g. in response to an incident report) rather than proactive, in which case the citizens may in fact positively update on the government's competence to protect public safety. Inconveniently, it would also likely take quite a significant or catastrophic incident (e.g. ransomware in public healthcare systems) for enough people to notice and attribute credit where due.

On the other hand, public attitudes toward the governments are likely to be much more unforgiving in the case of missing or ineffective interventions. If the government fails to prevent comparatively less harmful incidents (think discriminatory hiring) from AI, the public is more likely to start perceiving governments or designated authorities as slow-moving, ineffective, and less trustworthy. This may result in public apathy toward the institution of pacing (and relevant authorities) and subsequently governments having less incentive to invest in future pacing infrastructure or honor coordination treaties.

We should keep in mind that frontier AI pacing interventions would not be the first controversial tech policy area. Norms around tech policy have been shaped by the experience of industry and government policies around GMOs, civil nuclear, stem cells, and gain-of-function research, and have contributed to broader societal attitudes toward "progress" or "precaution". The viability of pacing interventions will depend on these pre-existing attitudes and the degree of politicization, but also the impacts of a pacing intervention, successful or otherwise, could impact the narratives and relative power of these different camps in societies around the world.

5.6. Case A: A cap on frontier training

Here we again consider a coordinated ceiling on frontier training compute, applying initially to training activity by actors based in participating jurisdictions.

Effects on covered actors: Covered actors, with now-limited training compute per model, will face immediate incentives to economize and use their compute more efficiently. To the extent they are also newly constrained in experimental compute, they will benefit from better figuring out which experimental results generalize well from small to larger models. They will also be incentivized to prune training data, optimize algorithms, and develop architectures which best extract the most return per unit compute.

To the extent specific domains are not covered by the restrictions, we should expect compute to differentially flow to these, and attempts at reclassifying otherwise covered activities under these labels.

They may also focus more on specific domains rather than general purpose models: if you want to make the best coding or bio model with a limited number of flops, you probably get better results if you are making a coding- or bio-specific model rather than a general purpose one. This could lead to labs releasing far more models, or there being far more specialization among labs than currently.

Safety-gated exit conditions could increase the proportion of resources devoted to safety research, though lower revenues could also reduce its absolute funding relative to the counterfactual.

Beyond this, the majority of the changed incentives discussed in §5.1 from a general capabilities slowdown apply in this case.

Governments and international coordination. Governments will be very keen to ensure their rivals are affected by restrictions to an equal or greater extent than they are, and that economic harms are minimized. The US and China face competing incentives: relatively compute-poor China could see this as an opportunity to catch up if it binds US developers more tightly, while the US may seek terms which better preserve its advantages. The latter will likely mean that interventions which allow the compute buildout to be maintained and repurposed for inference or divided broadly among more actors will be more appealing.

Non-covered actors. AI startups which are not research labs will likely benefit, as their chances of being displaced by the next model release go down. Non-frontier labs will also likely benefit for the reasons laid out above. Incumbents in non-AI domains may have more time to adapt to AI and integrate it into their business processes, reducing their risk of displacement.

Governance mechanisms. With greater visibility on frontier labs' compute usage will come opportunities for espionage. Actors who are not meant to have access to the records of frontier labs' training runs may attempt to gain access and steal trade secrets via this mechanism.

Investment and markets. The relative attractiveness of frontier developer and compute investments will likely decrease. Finding ways to safely repurpose compute may smooth implementation significantly.

5.7. Case B: Restricting access to dangerous biological capabilities

Here we again consider a red line under which models that materially uplift biological-weapons capability cannot be deployed or widely released without restrictions.

Effects on covered actors. Covered developers will need to implement safeguards and evaluations to judge the efficacy of these. They will also, as in case A, need to share data with third party evaluators, and thus bear increased risk from espionage.

As in the case of Anthropic's Fable 5, they may see high rates of false positives and refusals from their models, leading to degraded user experience, pushing their users toward non-covered models and actors and hurting their relative market position. They may maintain access programs for approved actors who get to use a less-restricted model version to mitigate this.

Government actors will probably be relatively easy to keep invested in the continued restriction of diffusion of dangerous biological capabilities.

Non-covered actors' models will gain relative usefulness at the expense of covered actors as described above, if they provide comparable abilities with fewer restrictions.

Adversarial actors. The primary adaptation to be expected is using alternative, out of scope or uncovered models, and secondly jailbreaking powerful models. Enforcement over publicly released open weight models may prove difficult, and if capabilities diffuse at all via this or another route it may prove impractical to fully enforce a moratorium on these capabilities.

A further redirection is simply for actors to move away from biological capability to other threats, such as cyber or chemical weapons.

Governance mechanisms. The quality of evaluations and safeguards will greatly affect the rates of false positives and negatives, and thereby influence the magnitudes of the economic effects on different actors.

Investment and markets. The effects in this domain depend on the magnitude of the effects on the covered actors' business models, but relative to other interventions this intervention seems low-impact here.

5.8. Open Research Questions

- **What compensation schemes could make pacing interventions more desirable for actors who stand to lose financially from them?** What are the precedents for such compensation, how could they be funded, and what secondary impacts might they have?
  - *Srivastav & Zaehringer (2024),* *The Economics of Coal Phaseouts*
  - *Jobst Heitzig, Lessman & Zou (2018)* *Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games*
- **Which restrictions build what kinds of overhangs?** How are these overhangs likely to play out if realized, and how dangerous might they be? Can overhangs be addressed through complementary policies? Are there pacing interventions which do not build up an overhang?
  - *Belrose (2023),* *AI Pause Will Likely Backfire*. Looks at the negative case for a training pause, where a compute overhang leads to rapid progress.
- **Which actors gain relative power under different interventions**, and what are the expected consequences? What is the historical track record of uses and abuses of power when an activity comes under deliberate pacing intervention?
  - *Coe & Vaynman (2015),* *Collusion and the Nuclear Nonproliferation Regime*. Looks at how nuclear nonproliferation entrenched superpower influence.

    - *Cassata & de Chadarevian (2025),* *Asilomar Across the Atlantic*. Restrictions on recombinant-DNA research empowered certain scientific organizations.
- **What safeguards can be deployed to guard against mission creep**, where regulators or newly empowered authorities could gain power beyond what was intended and become hard to dislodge?
    - *Romano & Levin (2021),* *Sunsetting as an Adaptive Strategy*
    - *Molloy (2021),* *Approach with Caution: Sunset Clauses as Safeguards of Democracy?*

6. Conclusion

Making the right choices about the pace of AI development will be critical for everything from national security to public health. Indeed, the choices governments, AI developers and other key actors make in the next few years may well ripple outward for decades or centuries. We have argued that these choices are not obvious and rely on a long list of subquestions and supporting activities that imply pacing decisions in very different directions. Given the stakes and the complexity of the choices, it is crucial that we do what we can now to make those choices go well. And though these choices are often inherently political, that is all the more reason to want sober reckoning with the key considerations.

Ultimately, successful pacing will depend on more than specific proposals and a rich understanding of any particular area. Decision-makers with the power to bring about pacing interventions will need the ability to effectively adapt to a rapidly changing world, connecting many distinct considerations and information sources into a coherent whole, recognizing and navigating the increasingly sharp trade-offs between deeply held values.

The world will pace progress one way or another. Absent better tools, it might do so haphazardly, through improvised reactions overfit to prior expectations, in ways that fail to actually address risks, through institutions that outlast their use. Our hope is that, with proper research, pacing can become progressively more deliberate, targeted, proportionate, decisive, and legitimate.

This paper has offered some initial frameworks for thinking about this problem and outlines many other crucial questions which could benefit from further research. We hope to see others iterate on these ideas, or critique them as is due.

Research into pacing will fundamentally be a matter of triage. We do not have the luxury of considering all these questions to our satisfaction before presenting our best answers. We should make use of the time we have.

7. Call To Action

If you are interested in working on any of the open questions we lay out above, **we'd like to help**. For instance, we can connect you with expertise and potential funders. We're also happy to help people in policy use this framework. Please contact us at hello@pacing.tech.

See also the calls to action from other organizations:

- AI Futures Project, "Get Involved", including their interest form.
- Encode AI, "Contact Us"
- For AI staff
- Safe AI Forum

Appendices

8.1. Appendix: All open questions

2: Why pace?

- **How much, and in what ways, would more time allow us to better manage various AI risks?** What are the bottlenecks to mitigation or adaptation of different AI risks? What factors other than time influence AI risk management? What risk management efforts can be taken now, and which can only be taken once certain AI capability or adoption thresholds are crossed? Related work:
  - *MacAskill & Moorhouse (2025),* *Preparing for the Intelligence Explosion* Explicitly sorts "grand challenges" by whether they need calendar time, human deliberation, or just more AI.
  - *Hobbhahn (2025),* *What's the short timeline plan?* A concrete inventory of which safety measures are ready to deploy now, and which need years of preparation.
- **How might pacing become more or less difficult over time?** What investments can be made now to preserve optionality over pacing in the future? In particular, under what circumstances does pacing now make future pacing more or less feasible? Related work:
  - *Rahman (2026),* *Does Distributed Training Undermine Compute Governance?*
  - *Sastry et al. (2024),* *Computing Power and the Governance of Artificial Intelligence*. Maps compute governance options and their readiness.
- **How do actors in this space make decisions about pacing?** What evidence do they currently consider and what assumptions do they currently make? What pathways exist for external research to inform such decisions, e.g. in government or lab leadership, and what makes that information transfer more effective?
  - *METR (2025),* *Common Elements of Frontier AI Safety Policies*
- **Which AI risks are most likely to motivate a pacing intervention, now and in the future?** Where do different dangerous capabilities sit on the offense/defense

balance, and how will that change over time? Related work:

- *Garfinkel & Dafoe (2019),* *How Does the Offense-Defense Balance Scale?*
- *Shevlane & Dafoe (2020),* *The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?*
- *Esvelt (2022),* *Delay, Detect, Defend: Preparing for a Future in which Thousands Can Release New Pandemics*

- **How does transparency about capabilities affect coordination?** When does common knowledge of research progress intensify or weaken race dynamics, for example by revealing that a rival is close behind or that progress is possible? Related work:
  - *Bostrom (2017),* *Strategic Implications of Openness in AI Development*
  - Armstrong, Bostrom & Shulman (2016), *Racing to the Precipice: a Model of Artificial Intelligence Development* Better information about rivals' capabilities can *increase* danger when teams are close, because it removes uncertainty that induces caution.
- **What are the risks and benefits of titration** ("deploy it and learn the risks empirically")? Have past release strategies from frontier labs succeeded in managing known AI risks? What might change in the future risk landscape? Given our uncertainty about the true risks, what criteria should gate the deployment of new frontier models?
  - *Shevlane et al. (2023),* *Model Evaluation for Extreme Risks*.

3: Pace what?

- **How should hazards be translated into covered activity for pacing interventions?** Risks we would like to target, such as uncontrolled automation of AI R&D and bioweapon uplift, build up over various stages of AI research, development and deployment; capabilities will initially emerge at some point in training, and we may want to avoid such a point being reached, or we may care more about wider deployment (especially if capabilities have positive use cases we want to preserve). Research should compare candidate boundaries.
  - *Shevlane et al. (2023),* *Model Evaluation for Extreme Risks*
  - *Hooker (2024),* *On the Limitations of Compute Thresholds as a Governance Strategy*

- **How can pacing thresholds be made specific and yet still cover distributed activity?** A pacing intervention targeting a threshold could potentially be circumvented by distributing activities or artifacts such that each sits below the threshold. How can we design aggregation rules and methods to handle cumulative risk from activities that are divided across space, time, processes and entities, without hindering low-risk activities?.
    - *Seferis & Fist (2026),* *Detecting Compute Structuring in AI Governance Is Likely Feasible*
    - *Rahman (2026),* *Does Distributed Training Undermine Compute Governance?*
- **What practical coverage is sufficient?** If we consider the reach available through company control, infrastructure providers and national rules, including their supply-chain effects, can we estimate bounds on activities that would be effectively covered by an intervention and relevant activities that would be missed?
    - *Koopmanschap & Barten (2026),* *How to Catch a GPU*. Maps issues with enforcement coverage as dangerous capabilities come to require progressively less compute
    - *Egan & Heim (2023),* *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers*
- **How effective are different access restrictions once a dangerous capability has been released?** How do factors such as access guardrails, alignment training, access to inference compute, ease-of-use, and tacit knowledge affect risk once diffusion has already occurred?
    - *Tamirisa et al. (2024),* *Tamper-Resistant Safeguards for Open-Weight LLMs*. Identifies limitations in how reliable safeguards can be for open weight models.
    - *Ord (2025),* *Inference Scaling Reshapes AI Governance*. Identifies inference scaling as a lever for released models.
- **How can we permit exceptions to allow useful work, without being so permeable that it makes the rule useless?** In the case of compute controls, it now seems technologically feasible, to some extent, to identify what uses a GPU is being put to. What other technical advances can allow interventions to be less blunt and more narrowly scoped?

    - *Gargiulo & Kulp (2026), Workload Identification with Physical Side Channels for AI Governance*
- **What are the tradeoffs between verification and invasiveness for different interventions?** How can we push the frontier forward?
    - *Scher & Thiergart (2024), Mechanisms to Verify International Agreements About AI Development*. Gives an overview of different verification mechanisms and what they require.
    - *Petrie et al. (2025), Flexible Hardware-Enabled Guarantees for AI Compute*. Proposes verifying compliance without exposing sensitive information about AI development.

4: Pace how?

- **How do the incentives of bound parties change across the lifecycle?** To what extent can different actors reliably predict the behavior of other actors throughout the lifetime of a pacing intervention? How load-bearing are these predictions of behavior going to be for coordination of pacing interventions?
    - *Finke (2026), International Agreements to Limit Frontier AI: Objectives and Exit*.
    - *Koremenos (2005), Contracting around International Uncertainty*.
    - *Goldstein and Salib (2025), How to Stop an AI Arms Race*.
- **How does information sharing impact the credibility of coordinated pacing?** Which information, in what granularity, through which channels, by which actors, matter most for credible coordinated pacing? How can this be kept compatible with national security considerations, commercial confidentiality, and cybersecurity? To what extent can this information sharing be kept robust to manipulation?
    - *Wasil et al. (2024), Verification Methods for International AI Agreements*.
    - *Scher et al (2025), An International Agreement to Prevent the Premature Creation of Artificial Superintelligence*.
- **What evidence could legitimize a speedy initiation of pacing? What is likely to be the acceptable tolerance for unreliable evidence for different actors?** Can scenarios and thresholds be specified in advance to a level of specificity that would

garner coordinated buy-in to a rapid pacing onset? Can evidential and assessment processes be agreed on in advance?

  - *Karnofsky (2024),* *If-Then Commitments for AI Risk Reduction*

- **How could dry-run simulations inform and prepare for pacing interventions?** What aspects of simulation design, delivery and follow-up affect their effectiveness and impact? Could simulations harm or misguide pacing interventions?

  - *Gruetzemacher et al. (2024),* *Strategic Insights from Simulation Gaming of AI Race Dynamics*
  - *Bartels (2020),* *Building Better Games for National Security Policy Analysis*

- **How will our reliance on different sources of information about risks, and our methods for communication and coordination, change as AIs become more capable, autonomous, or integrated?**

  - *Clymer et al. (2024),* *Safety Cases: How to Justify the Safety of Advanced AI Systems*

- **What does the space of exit scenarios look like?** For example, small-scale interventions may allow for immediate release, whereas exits from ongoing large-scale interventions impacting multiple facets of the economy and society (e.g. export controls) may need to be staged. Other than staging, what other parameters of exits exist, and how could they be tuned?

  - *Finke (2026),* *International Agreements to Limit Frontier AI: Objectives and Exit*

- **What is the relationship between initiation and exit triggers?** Intuitively, the exit trigger should track whatever justified the intervention in the first place, but what could affect that link and what other factors should be considered?

  - *Cârlan et al. (2024),* *Dynamic Safety Cases for Frontier AI*
  - *Karnofsky (2024),* *If-Then Commitments for AI Risk Reduction*

5: Then what?

- **What compensation schemes could make pacing interventions more desirable for actors who stand to lose financially from them?** What are the precedents for such compensation, how could they be funded, and what secondary impacts might they have?

  - *Srivastav & Zaehringer (2024),* *The Economics of Coal Phaseouts*

    - *Jobst Heitzig, Lessman & Zou (2018)* *Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games*

- **Which restrictions build what kinds of overhangs?** How are these overhangs likely to play out if realized, and how dangerous might they be? Can overhangs be addressed through complementary policies? Are there pacing interventions which do not build up an overhang?
    - *Belrose (2023),* *AI Pause Will Likely Backfire*. Looks at the negative case for a training pause, where a compute overhang leads to rapid progress.
- **Which actors gain relative power under different interventions**, and what are the expected consequences? What is the historical track record of uses and abuses of power when an activity comes under deliberate pacing intervention?
    - *Coe & Vaynman (2015),* *Collusion and the Nuclear Nonproliferation Regime*. Looks at how nuclear nonproliferation entrenched superpower influence.
    - *Cassata & de Chadarevian (2025),* *Asilomar Across the Atlantic*. Restrictions on recombinant-DNA research empowered certain scientific organizations.
- **What safeguards can be deployed to guard against mission creep**, where regulators or newly empowered authorities could gain power beyond what was intended and become hard to dislodge?
    - *Romano & Levin (2021),* *Sunsetting as an Adaptive Strategy*
    - *Molloy (2021),* *Approach with Caution: Sunset Clauses as Safeguards of Democracy?*

8.2. Appendix: longlist of pacing interventions

For concreteness, the following attempts to list the levers we have available to pace AI. Note that a lever's inclusion here is not an argument in favor of acting on it.

One simple task for the pacing field is to have serious up-to-date research on each of the following levers, and to then model the dependencies and tensions between individual levers.

Compute → dangerous capabilities

1. Cap training FLOPs per run (Heim & Koessler 2024; Calero-Forero 2026; Scher et al 2025)
2. Require pre-registration and notice for planned training runs above a threshold (EO 14110; SB 53)
3. Aggregation rules: make multi-cluster and distributed runs count toward the cap (Shavit 2023; Heim & Koessler 2024)
4. Cap on total R&D compute per organization per year (Calero-Forero 2026)
5. Cap on the RL share of total training compute (Irpan 2024)
6. Minimum ratio of monitoring compute per inference compute (AI Futures Project 2026, Achiam 2026)
7. Minimum ratio of safety spending per training compute (AI Futures Project 2026)
8. Tax R&D compute above a threshold (Calero-Forero 2026)
9. Regular reporting of each lab's compute split into final runs, experiments, internal inference, and external inference (Epoch 2026; EO 14110).

Chips → compute → dangerous capabilities

10. Export-control performance threshold for accelerators (BIS AC/S rule)
11. Export controls on lithography, HBM and advanced packaging (BIS SME rule, Oct 2023; BIS rule 2024)
12. Chip registry: serial numbers and owner of record above threshold, smuggling penalties (Sastry et al. 2024; Fist & Grunewald 2023)
13. Location verification on accelerators (Brass & Aarne 2024; Chip Security Act, S.1705)
14. Hardware-enabled mechanisms on new chips: offline licensing, metering, attestation (Aarne, Fist & Withers 2024; Kulp et al. 2024; FlexHEG 2025)

Datacenters (installed chips) → compute → dangerous capabilities

15. Permit review threshold for datacenters (Sanders)

16. Grid interconnect queue (Epoch 2024)
17. Registry of datacenters with satellite-verified construction status (Epoch)

Data → dangerous capabilities

18. Disclosure of synthetic-data share and RL environments used in frontier training (EU AI Act)

Algorithms → effective compute → dangerous capabilities

19. Publication embargo on frontier algorithmic results (Bostrom 2017; Shevlane & Dafoe 2020)
20. Total Research Transparency: mandatory disclosure of all frontier research (AI Futures Project 2026)
21. Structured access to code and checkpoints via vetted institutions (Shevlane 2022)
22. Classification regime for capability-elicitation techniques (Shevlane & Dafoe 2020)

Talent → algorithms → dangerous capabilities

23. Visa quota and processing time for frontier researchers (Zwetsloot et al. 2019, CSET)
24. Vetting and cooling-off periods for staff with weight or R&D compute access (Nevo et al. 2024)

Capital → compute → dangerous capabilities

25. Compute tax (dollar per FLOP on training runs above a threshold) (Calero-Forero 2026)
26. Strict liability for catastrophic harms from frontier models (Weil 2024)
27. Mandatory liability insurance, premiums priced by capability tier (Trout 2024)
28. Investor-facing AI risk disclosure (SEC 2023)

AI R&D capabilities → algorithms → effective compute → dangerous capabilities

29. Fraction of R&D compute consumed by autonomous agents (Stix et al. 2025; Charnock et al 2026)
30. AI R&D speed-up trigger in safety frameworks (Anthropic RSP; DeepMind FSF; OpenAI Preparedness)
31. Human review ratio for AI-written research code and experiment plans (Stix et al. 2025)
32. Safety case required before internal deployment on the R&D stack (Clymer et al. 2024; Stix et al. 2025)
33. Monitoring coverage of internal agent actions (Greenblatt et al. 2023; METR red-team of Anthropic's internal monitoring, 2026)

Actors and cadence → race intensity → all other levers

34. Licensing vs registration for frontier development (Anderljung et al. 2023; June 2026 EO)
35. Minimum interval between frontier releases (FLI pause letter 2023)
36. Pre-deployment testing window with government access (June 2026 EO; FRONTIER Act)
37. Coordinated-pause trigger and duration across signatories (Alaga & Schuett 2023)
38. Lead-margin reporting: months between top lab and next (Epoch; Karnofsky 2022)

Dangerous capabilities evals

39. Pretraining data filtering for CBRN, cyber-offense and self-replication content (O'Brien et al. 2025)
40. Verified unlearning of specified capabilities (Li et al. 2024, Feng et al 2025)
41. Capability thresholds by domain (Koessler, Schuett & Anderljung 2024; the old Anthropic RSP)

Generality → dangerous capabilities

42. Separate regulatory track for narrow AI, scientific models (Drexler 2019; EU AIA)
43. Gating of long-horizon agentic post-training for general models (Chan et al. 2023; Kwa et al. 2025)

Legibility of model reasoning → control of dangerous capabilities

44. Codebase / log audits for optimization pressure on chain-of-thought (Korbak et al. 2025; Baker et al. 2025)
45. Disclosure and gating of latent reasoning architectures (Hao et al. 2024, Coconut; Korbak et al. 2025)
46. Online weight updates in deployment off by default (Greenblatt et al. 2023)
47. Persistent memory scope (Shavit et al. 2023, OpenAI)
48. Canary-tagging or filtering of safety and eval content in training data (Berglund et al. 2023; Laine et al. 2024)

Weight security → blocking exfiltration

49. Mandatory weight security level ("SL") by capability (Nevo et al. 2024; Anthropic ASL-3)
50. Two-person rule and hardware keys for weight access (Nevo et al. 2024)
51. Insider-threat program coverage (Nevo et al. 2024)
52. Weight retention policy (Anthropic 2025)

Inference → dangerous capabilities

53. Per-query reasoning compute cap for the most capable models (Hooker 2024; Ord 2025)
54. Token tax (Irwin 2026)
55. Deployment tax by capability tier (Calero-Forero 2026)
56. Liability allocation between deployer and developer for autonomous services (Weil 2024)

Autonomy → dangerous capabilities

57. Agent permission tiers (Shavit et al. 2023, OpenAI; Chan et al. 2024)
58. Spend limits per agent and per task (Shavit et al. 2023, OpenAI)
59. Human approval for irreversible actions (Shavit et al. 2023, OpenAI; EU AI Act)
60. Sub-agent spawn depth and maximum unattended run length (Kwa et al. 2025, METR)
61. Agent identifiers and action-log retention (Chan et al. 2024)
62. Kill-switch latency requirement (Orseau & Armstrong 2016; Shavit et al. 2023, OpenAI)
63. Certification and fleet registration for embodied agents (EU Machinery Regulation 2023/1230)

Multi-agent → dangerous capabilities

64. Instance-count reporting per deployment (Chan et al. 2024)
65. Agent-to-agent communication logging with steganography checks (Motwani et al. 2024)

Safety research

66. Minimum fraction of total compute reserved for safety research, above a total compute threshold (OpenAI 2023)
67. Minimum safety headcount as a ratio of total research headcount (AI Lab Watch)
68. External safety compute grants (minimum FLOP/year given to independent labs) (NAIRR)

Evaluation

69. Third-party evaluator access depth (Casper et al. 2024)
70. Elicitation budget per dangerous-capability eval (METR elicitation protocol; Barnett & Thiergart 2024)
71. Sandbagging detection protocol (van der Weij et al. 2024)

Transparency

72. Required “AI Assurance Level” for developers in the frontier tier (Brundage et al. 2026)
73. Embedded auditors running regular audits with non-public access (Brundage et al. 2026)
74. Audit scope including internal deployment, information security and safety decision-making (Brundage et al. 2026)
75. Number of accredited audit providers and a standards body (Brundage et al. 2026)
76. Incident reporting deadline (SB 53 §22757.13)
77. Statutory whistleblower channel and protection (SB 53; Right to Warn letter 2024)

Governance response

78. Indexing compute thresholds to measured algorithmic progress (Heim & Koessler 2024; Epoch 2025)
79. Legislation with automatic clause triggers: e.g. once an eval result is shown, legal obligations come into force (Karnofsky 2024)
80. Regulator capacity (CAISI; UK AISI; IFP 2026)

Coordination

81. Treaty verifications: chip registry, datacenter inspections, interconnect bandwidth limits (Scher & Thiergart 2024; Baker et al. 2025)
82. Training-run declarations exchanged between states (Shavit 2023; Baker et al. 2025)
83. Verification R&D budget and a frontier-state incident hotline (Future Society 2026)

8.3. Appendix: Bibliography

Recent

- Finke (2026), *International Agreements to Limit Frontier AI: Objectives and Exit.* arXiv
- Larsen, Dean, Halstead, Lifland, Greenblatt & Kokotajlo (2026). *AI 2040: Plan A*. AI Futures Project
- Lifland et al (2026). *How to Pace the US Frontier*. AI Futures Project.
- Fist et al (2026). *How Should the US Prepare for Increasingly Automated AI R&D?*. Institute for Progress.
- Koopmanschap, Barten (2026), *How to Catch a GPU: A Taxonomy of Verification and Enforcement Mechanisms for International AI Agreements.* arXiv

Foundations

- Bostrom (2002). *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*. Journal of Evolution and Technology 9. nickbostrom.com
- Shulman (2009). *Arms Control and Intelligence Explosions*. ECAP. intelligence.org
- Marchant, Allenby & Herkert, eds. (2011). *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*. Springer. doi:10.1007/978-94-007-1356-7
- Armstrong, Bostrom & Shulman (2016). *Racing to the Precipice: A Model of Artificial Intelligence Development*. AI & Society 31. doi:10.1007/s00146-015-0590-y
- Christiano (2018). *Takeoff Speeds*. sideways-view.com
- Maas (2018). *Two Lessons from Nuclear Arms Control for the Responsible Governance of Military Artificial Intelligence.* IOS.
- Koessler, Schuett, Anderljung (2024). *Risk thresholds for frontier AI*. arXiv.
- Dafoe (2018). *AI Governance: A Research Agenda*. GovAI / FHI. governance.ai
- Kulveit, Douglas, Ammann, Turan, Krueger & Duvenaud (2025). *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*. arXiv:2501.16946

- Barnett & Scher (2025). *AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions*. MIRI Technical Governance Team. arXiv:2505.04592
- Favaro & Clark (2026). *When AI Builds Itself*. Anthropic Institute. anthropic.com

Economics

- Aschenbrenner (2020). *Existential Risk and Growth*. GPI Working Paper 6-2020. leopoldaschenbrenner.github.io
- Sandbrink, Hobbs, Swett, Dafoe & Sandberg (2022). *Differential Technology Development: An Innovation Governance Consideration for Navigating Technology Risks*. SSRN. ssrn.com
- Jones (2024). *The A.I. Dilemma: Growth versus Existential Risk*. AER: Insights 6(4). nber.org (WP 31837)
- Trammell & Aschenbrenner (2024). *Existential Risk and Growth*. GPI Working Paper 13-2024. philiptrammell.com

The pause debate

- Grace (2022). *Let's Think About Slowing Down AI*. AI Impacts / LessWrong. lesswrong.com
- Belrose (2023). *AI Pause Will Likely Backfire*.
- Buterin (2023). *My Techno-optimizm*. vitalik.eth.limo
- Tallinn (2024). *Priorities for AI Risk Reduction*. jaan.online
- Katzke & Futerman (2024). *The Manhattan Trap: Why a Race to Artificial Superintelligence Is Self-Defeating*. Convergence Analysis. arXiv:2501.14749
- Larsen, Dean, Halstead, Lifland, Greenblatt & Kokotajlo (2026). *AI 2040: Plan A*. AI Futures Project. ai-2040.com

International agreements

- Ho, Barnhart, Trager, Bengio, Brundage, Casovan, Haas, Nemitz, Sastry, Weller, Zhang & Zhang (2023). *International Institutions for Advanced AI*. arXiv:2307.04699
- Trager, Harack, Reuel, Carnegie, Heim, Ho, Kreps, Lall, Larter, Ó hÉigeartaigh, Staffell & Villalobos (2023). *International Governance of Civilian AI: A Jurisdictional*

Certification Approach. arXiv:2308.15514

- Hausenloy, Miotti & Dennis (2023). *Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI*. arXiv:2310.09217
- Emery-Xu, Jordan & Trager (2025). *International Governance of Advancing Artificial Intelligence*. AI & Society 40. doi:10.1007/s00146-024-02050-7
- Al Ramiah, Koopmanschap, Thorsteinson, Khan, Zhou, Noh, Meindertsma & Shafiq (2025). *Toward a Global Regime for Compute Governance: Building the Pause Button*. arXiv:2506.20530
- Scher, Abecassis, Barnett & Abeyta (2025). *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence*. arXiv:2511.10783
- Finke (2026). *International Agreements to Limit Frontier AI: Objectives and Exit*. TAIGR @ ICML 2026. arXiv:2607.16224

Deterrence

- Hendrycks, Schmidt & Wang (2025). *Superintelligence Strategy: Expert Version*. arXiv:2503.05628
- Rehman, Mueller, Mazarr et al. (2025). *Seeking Stability in the Competition for AI Advantage*. RAND commentary. rand.org
- Abecassis (2025). *Refining MAIM: Identifying Changes Required to Meet Conditions for Deterrence*. MIRI. intelligence.org
- Arnold (2025). *Superintelligence Deterrence Has an Observability Problem*. AI Frontiers. ai-frontiers.org
- Hendrycks & Khoja (2025). *AI Deterrence Is Our Best Option*. AI Frontiers. ai-frontiers.org
- Delaney (2025). *Crucial Considerations in ASI Deterrence*. IAPS. iaps.ai

Verification

- Brundage et al. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. arXiv:2004.07213
- Baker (2023). *Nuclear Arms Control Verification and Lessons for AI Treaties*. arXiv:2304.04123

- Scher & Thiergart (2024). *Mechanisms to Verify International Agreements About AI Development*. MIRI Technical Governance Team. arXiv:2506.15867
- Wasil, Reed, Miller & Barnett (2024). *Verification Methods for International AI Agreements*. arXiv:2408.16074
- Koopmanschap & Barten (2026). *How to Catch a GPU: A Taxonomy of Verification and Enforcement Mechanisms for International AI Agreements*. Existential Risk Observatory. arXiv:2607.22619
- Choussat & Khoja (2026). *An International AI Slowdown Is Ready Whenever Politicians Are*. AI Frontiers. ai-frontiers.org

Compute governance

- Shavit (2023). *What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*. arXiv:2303.11341
- Egan & Heim (2023). *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers*. arXiv:2310.13625
- Sastry, Heim, Belfield, Anderljung, Brundage, Hazell et al. (2024). *Computing Power and the Governance of Artificial Intelligence*. arXiv:2402.08797
- Heim, Fist, Egan, Huang, Zekany, Trager, Osborne & Zilberman (2024). *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation*. Oxford Martin AIGI. oxfordmartin.ox.ac.uk
- Aarne, Fist & Withers (2024). *Secure, Governable Chips*. CNAS. cnas.org
- Kulp, Gonzales, Smith, Heim, Puri, Vermeer & Winkelman (2024). *Hardware-Enabled Governance Mechanisms*. RAND WR-A3056-1. rand.org
- Petrie, Aarne, Ammann & Dalrymple (2024). *Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees*. Part I, arXiv:2506.15093
- Brass & Aarne (2024). *Location Verification for AI Chips*. IAPS. iaps.ai
- Heim & Koessler (2024). *Training Compute Thresholds: Features and Functions in AI Regulation*. arXiv:2405.10799
- Hooker (2024). *On the Limitations of Compute Thresholds as a Governance Strategy*. arXiv:2407.05694
- Ord (2025). *Inference Scaling Reshapes AI Governance*. arXiv

Developer commitments

- Shevlane, Farquhar, Garfinkel, Phuong, Whittlestone, Leung et al. (2023). *Model Evaluation for Extreme Risks.* arXiv:2305.15324
- Clymer, Gabrieli, Krueger & Larsen (2024). *Safety Cases: How to Justify the Safety of Advanced AI Systems.* arXiv:2403.10462
- Karnofsky (2024). *If-Then Commitments for AI Risk Reduction*. Carnegie Endowment. carnegieendowment.org
- Cârlan, Gomez, Mathew, Krishna, King, Gebauer & Smith (2024). *Dynamic Safety Cases for Frontier AI.* arXiv:2412.17618

How to cite this work

Douglas, R., Dillon, C., Moore, N., Leech, G., Bonde, M. K., Krishnan, R., Perez, N., Young, N., Slade Byrd, C., Casper, S., Kulveit, J., Duvenaud, D., & Avin, S. (2026). *Pacing the Frontier: A Framework and Research Agenda.* pacing.tech.