An International Agreement to Prevent the Premature Creation of Artificial Superintelligence

Aaron Scher* **David Abecassis*** **Peter Barnett*** **Brian Abeyta***
Machine Intelligence Research Institute
Technical Governance Team
{aaron.scher,david,peter,brian}@intelligence.org

Abstract

Many experts argue that premature development of artificial superintelligence (ASI) poses catastrophic risks, including the risk of human extinction from misaligned ASI, geopolitical instability, and misuse by malicious actors. This report proposes an international agreement to prevent the premature development of ASI until AI development can proceed without these risks. The agreement halts dangerous AI capabilities advancement while preserving access to current, safe AI applications.

The proposed framework centers on a coalition led by the United States and China that would restrict the scale of AI training and dangerous AI research. Due to the lack of trust between parties, verification is a key part of the agreement. Limits on the scale of AI training are operationalized by FLOP thresholds and verified through the tracking of AI chips and verification of chip use. Dangerous AI research—that which advances toward artificial superintelligence or endangers the agreement's verifiability—is stopped via legal prohibitions and multifaceted verification.

We believe the proposal would be technically sufficient to forestall the development of ASI if implemented today, but advancements in AI capabilities or development methods could hurt its efficacy. Additionally, there does not yet exist the political will to put such an agreement in place. Despite these challenges, we hope this agreement can provide direction for AI governance research and policy.

1 Introduction

The 2025 book *If Anyone Builds It, Everyone Dies* argues that the development of artificial superintelligence (ASI, AI substantially smarter than humans at all cognitive tasks) "using anything remotely like current techniques, based on anything remotely like the present understanding of AI" would lead to human extinction [82]. If that's the case, what should we do about it? Experts differ in their assessment of the situation. Many think that there is around a 10% chance of future AI advances causing human extinction (for aggregations of expert opinion see [43, 80, 30]). Even if there were "only" a 10% chance of human extinction, how should governments respond?

In late 2025, tens of thousands of researchers, public figures, and members of the public signed a statement calling for a prohibition on the development of superintelligence until there is "broad scientific consensus that it will be done safely and controllably, and strong public buy-in" [74]. This report describes what such a prohibition could look like in practice: an international agreement which, if implemented, would postpone the development of ASI until proper mitigation efforts are developed and implemented (i.e., potentially decades). Practically speaking, this agreement is not

*Equal contribution.
This paper references several appendices, to view these appendices please see the full paper on arXiv: https://arxiv.org/abs/2511.10783

Article presented at the Second IASEAI Conference (IASEAI'26).

going to be implemented tomorrow, but we hope it can serve as an end goal for AI governance, and in Appendix C we lay out a staged implementation that iteratively builds up to the final agreement.

Our proposed agreement is not without tradeoffs and risks: it forgoes beneficial AI capabilities, creates some pathways for authoritarian risk (although we think less risk than the current trajectory), cedes the U.S. technological lead in AI, risks leaking sensitive information due to verification, and would impose costs on industry and financial markets. Despite these downsides, we believe this agreement is worth pursuing because the alternative of racing towards ASI poses even greater risks.

Contributions:

- We provide a brief overview of AI risks, and an overview of the key details of the strategic situation informing our proposal.
- We describe the implementation of an international agreement to forestall the development of ASI until such a time as it can be developed safely.
- We explain the key assumptions and beliefs that differentiate our plan from others in the space.
- We provide the full text of an example agreement, detailed commentary, and connections to existing agreements in Appendix A. We discuss immediate steps different actors could take toward this agreement in Appendix B. We explain a staged implementation of the agreement that slowly builds from the current world to the agreement in Appendix C. We detail various approaches that could help locate existing AI chips and bring them under monitoring in Appendix D.

2 An overview of AI risks

There are many large-scale risks from advanced AI (i.e. AI much more powerful than today's AI technology) that might warrant slowing or stopping AI progress for a significant period (e.g., decades). Most of these risks stem from future AIs becoming increasingly capable tools: Rapid automation may destabilize economies [34, 39]. Capabilities that undermine strategic deterrence and incentivize first strikes may trigger wars and conflicts [63, 48]. Malicious actors may misuse AI to engage in wide-scale cyberattacks or to develop novel biological weapons [83, 31, 40]. AI used for mass surveillance, propaganda, and police-state oppression may enable extreme concentrations of power [32, 25, 34].

But while the above risks may lead to unprecedented catastrophes, humanity can probably survive them. A greater and more irreversible risk stems from AI systems that don't remain passive tools, but instead become capable enough to permanently disempower humanity. This would plausibly lead to human extinction, either as an intentional act by the AIs (e.g., to prevent humans from making powerful competing AIs) or as a side effect of AIs optimizing for goals that don't specifically designate human survival [81, 65].

This paper focuses on an international agreement to forestall ASI development until it can be done safely. A halt like this is primarily motivated by concerns about misalignment and AI takeover, and we think it is the only acceptable mitigation to those risks. But a halt would also be an effective tool for addressing many other risks, such as geopolitical destabilization, mass AI-caused unemployment, and misuse of powerful AIs [19]. Therefore, *others may support a halt even if they are not concerned with misalignment*.

Our view is that AI developers are not on track to solve the alignment problem before the development of ASI. We will not fully defend this claim here—the difficulty of alignment is not this document's central focus. Instead, we provide some general intuitions for why AI alignment appears challenging enough that success is far from guaranteed. This is sufficient to motivate our policy proposal.

The deep learning paradigm underpinning modern AI development seems highly prone to producing agents that are not aligned with humanity's interests. Under this paradigm, AI behavior is not hand-coded by human engineers, but instead emerges from automated training processes. Rather than producing lines of code that can be debugged when something goes wrong, these processes configure hundreds of billions to trillions of model weights that determine AI behavior. While it is sometimes possible to identify and activate features of these weights correlated with certain inputs or outputs, the deep machinery of these minds is effectively opaque to human observers [21]. No one can say with any certainty why an AI made a particular choice or how it would act in situations that differ from its training.

Consequently, no one can manually adjust an AI's weights to give it a particular goal or set of preferences. And yet, as seen in the growing ability of AIs to complete long-duration tasks, AIs are becoming increasingly capable of pursuing goals, e.g., in the setting of agentic software engineering [55]. Various well-documented examples of misbehavior by these same models (both in the lab and in the wild [17, 23, 61, 68, 62]) demonstrate that today's training methods do not produce robustly aligned agents that could be trusted not to disempower humanity if they had the capacity.

Many AI labs have a stated goal of building AIs that can surpass humans at all cognitive tasks. It is hard to predict *when* AI labs might succeed, but many predict ASI could be developed soon. The field has seen enormous progress in just five years—from AIs that couldn't hold a conversation to "reasoning models" that solve hard exam questions [36] and that some major tech companies claim write a substantial fraction of their code [28, 67]. Another few years of such progress might well push capabilities well beyond human level in all or nearly all domains.

One accelerating factor complicating all forecasts is the possibility of AIs automating the process of AI research and development itself [33, 41, 79]. The distance to this threshold is unclear, but some independent experts give median prediction dates ranging from 2028 to 2033 [59]. If or when this threshold is crossed, we may see a feedback loop where AI R&D is conducted by AIs that rapidly (in months or a few years) produce substantially superhuman AI [69, 12, 42].

We note that artificial general intelligence (AGI)—systems that perform at human-level across all cognitive domains—would likely emerge before ASI. While AGI itself poses significant risks, our primary focus is preventing ASI, as misaligned superintelligence poses an irreversible existential threat. However, AGI capable of automating AI research could accelerate the path to ASI and trigger rapid, uncontrollable capability gains. This agreement aims to halt AI capabilities advancement before the creation of AGI systems that could automate AI research.

On the current trajectory, the core challenge of AI alignment is unlikely to be adequately addressed before the first ASIs are built. Evidence for this can be seen in the failure to solve far easier related problems. Today's frontier models are not robust to jailbreaks: It is relatively easy to craft inputs that lead to outputs that AI developers attempt to avoid [8, 7, 84, 9, 77]. Current models, especially reasoning models, display a propensity to reward hack (exploiting unintended solutions) and lie to their users, against the best efforts of their developers [23, 79, 29, 11, 17]. Far tougher alignment challenges are likely to appear only at advanced capability levels where humanity would be unable to recover from failures [81].

There is a broad consensus in the field that catastrophic outcomes from misaligned AI pose a significant threat that should be addressed [20]. Hundreds of luminaries signed on to a 2023 statement that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." [27] That same year, Yoshua Bengio, a Turing Award recipient and the most cited living scientist, said he thinks there is a "twenty percent probability that it turns out catastrophic" [56]. In a 2023 survey, 38% of survey respondents at top AI conferences said there was at least a 10% chance that AI's outcome will be "extremely bad (e.g., human extinction)", conditional on the development of AI that can outperform humans on all tasks [43].

This understanding extends to AI company leaders themselves: Anthropic CEO Dario Amodei has predicted a 10-25% chance that "something goes really quite catastrophically wrong on the scale of human civilization" [75, 5]. OpenAI's Sam Altman has called superintelligence "probably the greatest threat to the continued existence of humanity" [3], and xAI's Elon Musk has said, "AI is a fundamental risk to the existence of human civilization" [60]. Musk has indicated that his participation in this ASI race stems from the belief that it would happen even without him [64]. Those racing toward superintelligence, to the extent they are concerned with catastrophic risks, are stuck in a collective action problem [13]. Halting the race and avoiding extinction would require *global coordination encompassing all relevant companies and governments*.

3 The strategic situation

A 10% chance of extinction is unacceptable. To mitigate these risks, various approaches have been proposed: Have governments tightly regulate and license AI development [6, 73], empower defensive actors early [26, 66, 22], implement a joint global AI development project to avoid racing [24, 46], race ahead while slowing opponents [16, 4], or prevent the creation of ASI before we are ready [19, 70, 2].

We think the only acceptable solution is to prevent the creation of ASI until catastrophic risks are mitigated. The state of AI alignment research is too nascent, and the problem too difficult. Nobody knows how to align advanced AIs to human values, including regulators or an international standards body (and we address some of these alternative plans in more depth in Section 5). Therefore, we think the right approach to this problem is to forestall the development of ASI for as long as is necessary to solve critical alignment challenges. We note that bans are not usually the best way to address dangerous technologies, but ASI is uniquely dangerous. An agreement establishing such a halt would need to be globally enforced, verifiable, and ready to implement once the political will to do so exists.

For a halt to be effective, it must extend to all countries. If only cautious actors stop, other parties will continue the march toward superintelligence. The United States (U.S.) pausing on its own might buy around six months until the nearest company in the People's Republic of China (PRC) catches up, and a U.S.–PRC agreement without global enforcement would similarly fall to the next closest actor, likely in a couple of years [14] (though it is unclear how long it would take to get from the current frontier to ASI). The U.S. and PRC are the preeminent superpowers and the centers of most frontier AI development; therefore, they are essential stakeholders in such a deal, and their strategic interests must be accommodated. Due to their influence, they are likely leaders in future AI agreements, whether or not other countries participate as well.

A coalition including the U.S. and PRC would likely be able to attract nearly all other states, even if some are reluctant, given these two countries' geopolitical influence and their current positions as AI leaders. We expect other countries to sign on to such an agreement once they understand AI risks better and conclude that international coordination is necessary to avoid these risks. Participation could also be encouraged with benefits such as security and access to AI infrastructure for allowed use. Where stronger incentives prove necessary, standard tools such as economic sanctions may apply. We believe a very ASI-concerned coalition headed by the U.S. and PRC would be sufficient to integrate other potential AI developers into the agreement or to sufficiently suppress other AI development so that it does not pose a threat.

All parties to the agreement will desire assurance that other parties are not engaged in secret AI capabilities projects. Providing this assurance is a key aim of the agreement. Given a low-trust environment, parties must have the means to verify compliance to their satisfaction. It gives special consideration to the independent verification efforts of the U.S. and the PRC.

There remain significant political obstacles to such an agreement occurring today. Nevertheless, to accept even a 10% chance of extinction is wholly inconsistent with how we manage other risks. We routinely adopt stringent mitigations for risks that are hundreds of times less likely and that would result in millions of times fewer fatalities, such as in the nuclear industry, and in many other fields [53, 76, 35].

In addition, the rate of progress is volatile and uncertain. Timelines to ASI may be short, and some possible developments in AI would preclude a robust monitoring and verification regime. Rapid feedback loops, hardware proliferation, and the loosening of supply chain bottlenecks all become more likely the longer an arrangement is delayed.

When world leaders are mobilized to address the situation, it is essential to have prepared measures for negotiation and rapid adoption by key players and the world at large.

4 The Agreement

A full draft agreement, with discussion of precedent and other considerations, is presented in Appendix A. For concision, this section describes how an implementation of that agreement might look in practice. Figure 1 provides an overview.

Motivated by the risks above, the U.S. and PRC lead a coalition of states to prevent the advancement of AI capabilities toward ASI. An agreement establishes how the U.S. and PRC coordinate to gain assurance that neither they nor anyone else is advancing AI capabilities. It empowers and accommodates states' information gathering efforts and, in particular, supports independent efforts by the U.S. and the PRC to verify compliance.

Decision-making power over the agreement reflects the underlying geopolitical realities: Today and for the foreseeable future, the U.S. and PRC are the primary actors whose participation is key to any AI agreement. They are the initial members of an Executive Council which governs the

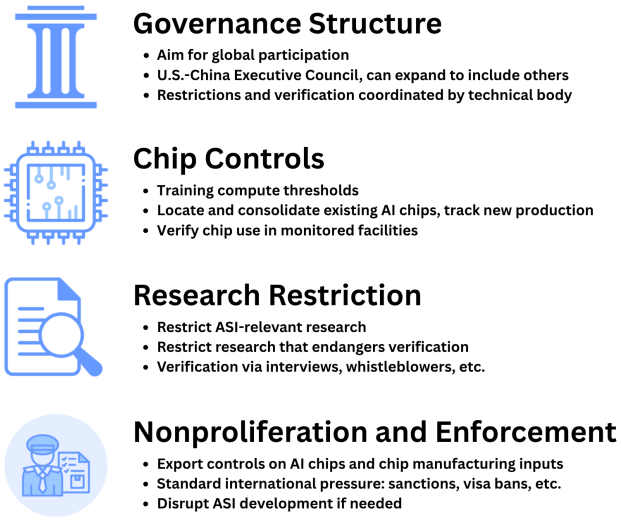

Figure 1: An overview of the agreement's main components.

agreement. This body can be expanded as appropriate and necessary, and we anticipate that other states with significant power or specific leverage over AI development may be included. That said, the responsiveness and flexibility of the agreement is maximized by limiting the number of parties required to modify it. The coalition nevertheless expands to include as many countries as possible over time.

AI development in the current deep learning paradigm is driven both by the large-scale employment of specialized computer chips and by improvements in AI algorithms [72, 52, 49]. Therefore, the coalition stops the advancement of AI capabilities toward superintelligence by restricting the scale of AI training (operationalized as total training FLOP) and by restricting research into AI algorithms and other areas that would undermine the governance regime. These research restrictions are narrowly targeted at research that advances towards ASI or would undermine the agreement.

AI training runs above the Strict Threshold (i.e., 10^{24} FLOP) are prohibited. Training runs below this threshold but above the Monitored Threshold (i.e., 10^{22} FLOP) must be approved and monitored by coalition authorities. Training runs below the Monitored Threshold require no approval or monitoring.

There are numerous considerations in deciding where these thresholds should be set, and these thresholds can be modified during agreement negotiations or after the agreement is in effect. First, nobody knows what scale of training would reach particularly dangerous AI capability levels; given this uncertainty and the massive negative effect of setting thresholds too high, we suggest a conservative approach. Second, many near-frontier AI models of today are trained at a scale only slightly above the Strict Threshold; for example, DeepSeek-R1 ($\sim 4 \times 10^{24}$ FLOP) and gpt-oss-120B ($\sim 5 \times 10^{24}$ FLOP) [37]. See Figure 2 for a visualization of how these thresholds relate to the training scale of existing AI models. Due to improvements in AI algorithms and data, the capability of models trained at a given computational scale increases rapidly over time [49]. Due to likely progress in algorithms and data between today and when this agreement would come into effect, AIs trained at the Strict Threshold will be more capable—potentially much more—than the models trained at that scale today. There are numerous other considerations in what these thresholds should be and whether FLOP thresholds are appropriate [38, 50, 47]; a full discussion is out of scope for this paper.

Verification of these restrictions is practical because AI chips are expensive and specialized, and thousands of them are needed for frontier AI development. This approach helps ensure that training runs that advance toward ASI are prevented while still allowing continued use of current AI technology. To achieve this verification, the coalition **locates existing AI chips**, **tracks new AI chips being produced**, and **monitors AI chip use** in order to ensure chips are not being used for restricted activities [71, 18, 45]. Chips are located through coordinated efforts by intelligence agencies and domestic law-enforcement agencies from coalition countries. These efforts involve combining supply chain tracking, mandatory reporting, state intelligence gathering, open-source intelligence, power

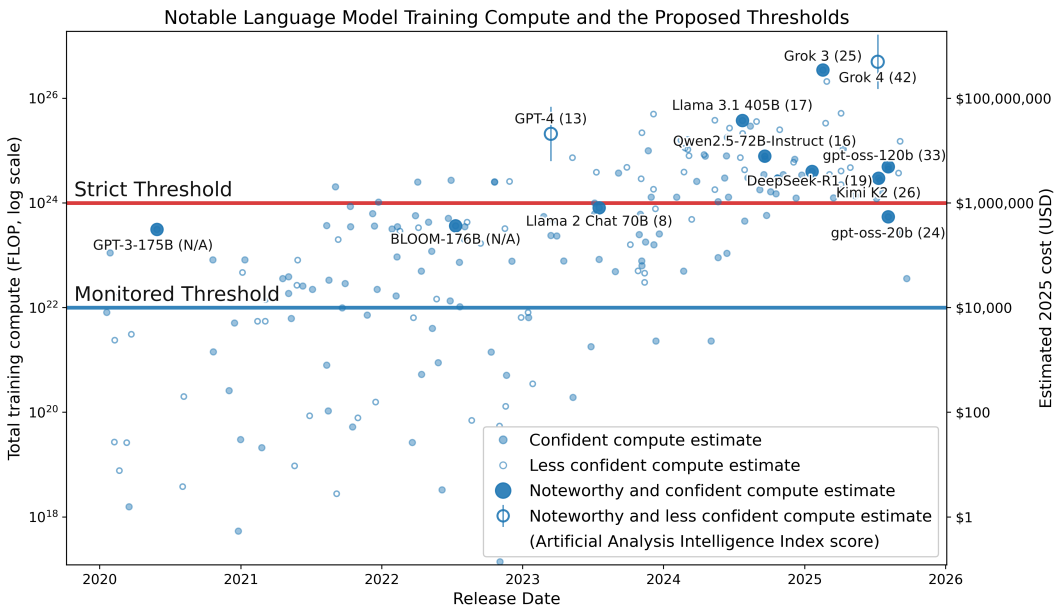

Figure 2: The training compute used to train various notable AI models in the last few years, along with the two thresholds in this agreement. In the agreement, new training above the Strict Threshold (10^{24} FLOP) would be prohibited and new training between the Monitored Threshold (10^{22} FLOP) and Strict Threshold (10^{24} FLOP) would be monitored. Due to a lack of public data, only some models have confident estimates for the training compute used. The number in parentheses is the model's score on the Artificial Analysis Intelligence Index. Data is based on [37] and [15]. Cost estimates are based on October 2025 rental prices for B200 GPUs from CoreWeave, assuming 50% utilization in FP8.

consumption monitoring, challenge inspections, whistleblowers, and more. Appendix D provides more detail on how this could be done. Tracking new AI chips relies on the concentrated chip supply chain and serves the goal of making sure chips are not smuggled outside of the monitoring regime. Finally, the chips themselves are monitored in order to verify that they are only being used for acceptable use cases.

AI chip monitoring starts with discerning whether the chips are being used for inference on existing AIs or training of new AIs. The coalition works to develop tamper-resistant on-chip mechanisms for such purposes, as finer-grained control permits greater use of AI chips for safe applications, with fewer inspections. Initially inspectors are given ongoing physical access to chips, as is likely needed for robust chip-use verification [1, 54]. To ensure chip use verification is applied, the coalition prohibits large concentrations of chips (i.e., greater than 16 H100-equivalents; 16 H100s cost approximately $500,000 USD in 2025) outside of monitored facilities.

Despite prohibitions on large-scale training, AIs could continue getting more capable via improvements to the algorithms and software used to train them. Therefore, the coalition adopts appropriate **restrictions on research that contributes to frontier AI development** or research that could endanger the verification methods in the agreement. These restrictions would cover certain machine learning-related research and may eventually expand to include other AI paradigms if those paradigms seem likely to lead to ASI. Coalition members draw on their experience restricting research in other dangerous fields such as nuclear and chemical weapons. These research restrictions aim to be as narrow as possible, preventing the creation of more capable general AI models while still allowing safe and beneficial narrow AI models to be created. The coalition encourages and makes explicit carve-outs for safe, application-specific AI activities, such as self-driving cars and other uses that provide benefits to society.

Verification that members are adhering to research restrictions is aided by the fact that relatively few people have the skills to contribute to this research—likely only thousands or tens of thousands, see Appendix A, Article IX. Nations verify compliance by way of intelligence gathering, interviews with the researchers, whistleblower programs, and more. This verification aims to use non-invasive

measures to ensure that researchers are not working on restricted topics. Additionally, inspectors verify that the medium-scale training allowed by the agreement uses only approved methods (e.g., from a Whitelist) and does not make use of any novel AI methods or algorithms; these would be evidence that restricted research had taken place.

The coalition's methods for detecting covert violations of the agreement do not prevent overt violations. For these, the coalition relies on nonproliferation and enforcement. **Nonproliferation efforts focus primarily on states that are not part of the agreement**, beginning with export controls on AI chips and chip manufacturing equipment as well as strong counterproliferation efforts to ensure these export controls are upheld.

Should nonproliferation fail, or if a rogue actor is found to be working towards ASI, coalition members can employ their agreed-upon **enforcement mechanisms**. These start with standard tools of international pressure such as diplomacy and economic sanctions. Depending on the nature of the developments, countries could escalate to other means at their disposal to disrupt, slow, or impede ASI development. This draws parallels to other contexts where countries have intervened to stop violations of arms-control and nonproliferation agreements.

This international agreement is intended to persist until the conditions surrounding AI development are no longer catastrophically dangerous. Although the exact details remain open to debate, conditions for lifting restrictions may include a strong AI alignment research field with a years-long track record of successfully solving alignment-related problems, an established consensus about the efficacy of ASI-alignment methods, strong misuse and proliferation controls, and the creation of a globally-monitored AI development project that proceeds with due caution. World leaders may also wish to stipulate a requirement to solve or mitigate specific AI-related challenges, including unemployment, power concentration, geopolitical destabilization, and misuse.

While the goal is for this agreement to last as long as is necessary, that may not be practical. There are numerous reasons the agreement might fail; for instance, two threat actors are countries outside the agreement and covert state-backed ASI efforts from countries that are in the agreement. The verification, nonproliferation, and enforcement measures in the agreement are set up with the purpose of preventing such failures, but it's hard to predict how successful they will be. World leaders may find themselves weighing the risk of a covert project building ASI despite the agreement against the risk of nullifying the agreement and building ASI themselves.

There is not yet political will to implement this agreement, but there might be political will to enact measures that build toward it. Therefore, in Appendix C we describe a staged implementation of this agreement. The staged approach starts with international confidence building and transparency measures, and eventually builds up to this full agreement. A staged approach better mirrors how international arms control agreements have taken place historically, and it allows action to happen today even while the political will for a full agreement does not exist today.

5 Why this plan in particular?

Wouldn't a less costly plan work? We have aimed to articulate a solution that could actually prevent the development of ASI. Various parts of this plan, such as restrictions on research and halting for decades, might seem extreme. **This section discusses why we believe these components are necessary to a successful plan**. We organize the section around objections readers might have, providing our response in turn.

Would the relevant actors implement such an agreement? As of mid-2025, the U.S. and PRC are not sufficiently concerned about the risks from ASI to agree to halt AI development. But, if people, including world leaders, understood the extreme risks, we believe there would be political will for leaders to take action. Because any actor building ASI, anywhere on earth, would threaten everyone, preventing premature development requires global enforcement. Even a 10% chance of human extinction should be enough to motivate this.

The implementation of this international agreement can also be staged, as we propose in this paper and detail in Appendix C. This would entail taking less costly actions initially, and then taking more substantial steps as political will increases.

Why not wait and pause later? The field of AI lacks reliable methods to predict when AI systems will develop dangerous capabilities. Given this fundamental uncertainty, we should set conservative limits on AI development, rather than risk overshooting catastrophic thresholds.

Crucially, we need to stop before dangerous AI systems have been trained. If an AI system has been trained that could cause catastrophic harm via inference, then governance must apply controls to both inference and training. This would also be the case for AI that can substantially automate AI development itself, as this could accelerate development of more efficient algorithms that could compensate for restrictions on hardware. Our current framework focuses on preventing training, but controls on inference would require more invasive measures.

Inaction will also lead to further AI hardware proliferation, making comprehensive tracking and monitoring harder. Additionally, while large secret state-run AI facilities likely don't exist today, this may change as governments recognize the geopolitical implications of ASI. The potential for these secret facilities may undermine the agreement.

Why are you prohibiting research? We recognize that restricting research is controversial and normally a bad idea. Unfortunately, we think it is necessary to prevent the premature development of ASI. This is primarily a descriptive statement, not a normative one: to prevent the development of ASI, the narrow set of research aimed at this goal must be stopped. In the current paradigm, AI capabilities increase from a combination of scaling computing power and having better AI algorithms; limiting only one of these factors would allow AI capabilities to improve through the other.

The risks from continued research can be decomposed into three types:

First, the fast grind of algorithmic progress: historical trends in AI algorithmic progress indicate that the number of operations used to train an AI to a given capability level drops by $3\times$ each year [49]. This means that each year, the threshold size of AI chip clusters that must be monitored to maintain the same level of verification of final AI capabilities would also decrease by $3\times$. If such trends are not interrupted, widespread consumer computers would eventually become dangerous, and it would be insufficient to monitor data center AI chip use. It is unclear whether consumer computers *even could* be monitored, but regardless of effectiveness, it is certainly not morally desirable to do so.

Second, progress is sometimes much faster than these average historical trends. For instance, Ho et al. 2024 estimate that the transformer architecture provided $7.2\times$ reduction in operations, or around two years of progress on its own [49]. Future major innovations, for example, that make much better use of inference or fine-tuning operations, would pose a major risk to the agreement and its verification. These are especially concerning because there could be little warning about them; by contrast, if progress was exactly $3\times$ per year, there would be plenty of warning signs that the agreement is becoming unstable.

Third, paradigms for developing AIs other than deep learning could bear fruit. These other paradigms are likely to use far fewer computational resources than the current development paradigm, so the agreement would not be robust to their success.

We are not excited about the prospect of restricting research. Unfortunately, this seems necessary. We are hopeful that this research restriction can be relatively narrow and have negligible collateral damage outside of AI, even if it must be somewhat broad within AI.

Why do we need to halt for so long? In short, **AI alignment is probably a difficult technical problem, and it is hard to be confident about solutions**. Pausing for a substantial period gives humanity time to be careful in this domain rather than rushing. Pausing for a shorter amount of time (e.g., 5 years) might reduce risk substantially compared to the current race, but it also might not be enough. In general, world leaders should weigh the likelihood and consequence of different risks and benefits against each other for different lengths of a pause.

Section 2 discusses some of the reasons why the AI alignment problem may be difficult. Generally, experts vary in their estimates of the difficulty of this problem and the likelihood of catastrophe, with some expecting the problem to be very hard [43, 30, 80]. Given this uncertainty about how difficult this problem is, we should prepare to pause for a long time, in case more effort is needed. Our agreement would allow for a long halt, even if world leaders *later* came to believe a shorter one was acceptable.

We also contend that there are other problems which need to be addressed during a halt even if one presumes that alignment can be quickly solved, and these problems are also of an uncertain difficulty. These include risks of power concentration, human misuse of AIs, mass unemployment, and many more. World leaders will likely want at least years to understand and address these problems. The international agreement proposed in this paper is primarily motivated by risks from AI misalignment, but there are numerous other risks that it would also help reduce.

What about automating AI alignment research? One objection to this paper's plan is that a short pause may suffice because AI alignment research could be automated. Automating AI alignment research was the goal of OpenAI's now-disbanded Superalignment team [57], and is a key focus of Anthropic's Alignment Science team [10]. AI companies themselves have not publicly provided detailed plans for how they plan to automate alignment research. Assuming they had such a plan, and assuming they acted with much more caution than the AI field of today, this plan *might* work. But we don't think one should be very confident in the plan's success, and it would be an irresponsible gamble with immense stakes. As discussed, the state of the AI alignment field is nascent and has made little progress so far. There are various reasons to expect this plan to fail, or at least to not be confident in its success. This topic has been debated at length in internet forums, and for the sake of concision, we will only briefly outline a couple arguments:

- Alignment problems for ASI appear difficult. Early AGIs may not be able to solve these problems quickly. More capable AGIs would have a better chance of solving these problems, but creating such AIs takes on more catastrophic misalignment risk. [78, 58]
- Early AGIs may not be aligned or controllable. This plan requires that early AGIs either be *aligned themselves* and thus trying to help solve alignment challenges, or that they are *misaligned but we are still able to get useful labor out of them* (as the AI Control research agenda aims to do [44]). Each of these conditions seems prima facie unlikely, and it seems difficult to obtain justifiable confidence in at least one of them working.

Why do we have to track down existing AI chips? That seems difficult and the flow of newly produced chips will overwhelm the existing stock quickly. One core goal of our plan is to prevent any data center capable of training ASI from evading monitoring. This will require locating existing AI chips. Some existing plans don't include locating existing AI chips, while other plans seem to implicitly rely on it, such as for retrofitting chips with verification mechanisms.

Simply monitoring newly produced chips is insufficient for our goals due to the existing chip stock. One might argue that new chips will comprise a much larger fraction of the world's computing power than existing chips. However, risk derives from the absolute number of unmonitored chips, not just their proportion in the total chip stock.

Voluntary submission of chips for monitoring, while politically easier, can likely not provide the assurance needed. Some actors will inevitably refuse to participate, and even willing participants won't trust each other. PRC will assume the U.S. is hiding chips; the U.S. will assume the same about the PRC. This dynamic could undermine the whole agreement.

Verification mechanisms require physical access to be effective. With current technology, inspectors cannot verify that chips are not being misused without first knowing where they are. Remote attestation and other technical solutions are in early stages and, even if developed, could likely be circumvented if the chips' physical locations remain unknown. For example, chip owners could simply choose not to update their chips' firmware.

Can't we just build defensive technology and adapt? That's how we manage most dangerous technology. Building defensive technologies and improving societal resilience are noble aims, but, unfortunately, appear insufficient in this case. There are likely to be AI-enabled or AI-accelerated technologies and strategies which are offense-dominant, where offensive use is relatively cheap and that are infeasibly difficult to defend against. Biological weapons are one possible example. Even if a technology is defense-dominant *in the limit,* defense may be *initially* outpaced by offensive uses. For example, it may eventually be possible to harden all critical cyber infrastructure against attacks, but there will likely be a window where there are AI systems capable of performing attacks before AI systems have hardened all such infrastructure. Additionally, ASI may enable entirely new classes of technology, such as advanced nanotechnology or novel methods of psychological manipulation. It seems overwhelmingly likely there will be at least one instance of a new technology that is catastrophically dangerous and too difficult to defend against.

Won't this plan enable authoritarianism? This agreement does include elements that could be misused for authoritarian abuse, such as centralizing AI infrastructure, monitoring training runs, and tracking researchers. **However, the default trajectory carries far greater authoritarianism risk.** ASI may be highly power-concentrating: if someone manages to develop an aligned superintelligence, it may grant them unrivalled economic, military, or technological power. Under current conditions, ASI will likely be built by a private company (potentially in partnership with government agencies), with minimal public oversight. In this situation, control over the most powerful technology would rest with perhaps 1-10 individuals. This would be an unprecedented concentration of power with few mechanisms to prevent permanent authoritarian lock-in. By contrast, this agreement prevents the creation of ASI, allowing time to establish the required governance structures.

The agreement's verification mechanisms (e.g., tracking chips, monitoring training runs, limiting research) would likely involve a combination of existing and new authorities. As a few examples of related precedent, the U.S. government already conducts extensive surveillance (including against U.S. persons), maintains export controls on sensitive technologies, inspects nuclear facilities, and restricts the publication of nuclear research. Where domestic implementation expands the government's power, there are likely safeguards that could be implemented to prevent misuse. For instance, multilateral controls, narrow scoping of authority, and transparency measures would reduce these risks.

Doesn't this plan take on too much risk by being slow to implement and allowing chips to continue existing? This plan aims to strike a balance between political viability and efficacy, balancing costs and benefits, but it may not balance them optimally. One concern is that the existence of AI chips makes breakout time too short: if some country decided to flout the agreement and attempt to build ASI, they might be able to do so quickly, leaving little time for others to thwart them. By contrast, if existing chips and the chip supply chain were decommissioned, it might take a decade or more for a country to build the necessary AI infrastructure, making it much easier for others to intervene. We are optimistic that strong verification tools accompanied by political will for enforcement would be sufficient to prevent rogue states from building ASI, even when AI infrastructure is allowed to exist.

Another concern relating to the plan is that it relies too heavily on remaining in the current AI development paradigm where significant computational resources are needed to improve AI capabilities. There are early signs that this may be changing, such as the inference-scaling regime started with OpenAI's o-series of models [51]. If implemented soon, this agreement would likely be sufficient, but the technical situation could deteriorate if action is deferred. Overall, our choice to go with this more permissive agreement—one that allows existing chips to keep existing and focuses primarily on preventing large training—is an attempt to balance costs and benefits, trying to preserve the safe use of existing AI models without the risks from continued capabilities advancement.

6 Conclusion

The development of artificial superintelligence represents an immense challenge for human civilization. Unlike most technological risks, where defensive measures and gradual adaptation can provide adequate protection, ASI poses an existential threat that demands preemptive action. This paper has outlined an international agreement centered on halting dangerous AI capabilities advancement through FLOP-based training restrictions, comprehensive AI chip tracking, and targeted research restrictions. While the political will for such an agreement does not exist today, the stakes are too high to accept inaction. A 10% chance of catastrophic outcomes is considered unacceptable in any other domain of national security or public safety.

We recognize that this proposal faces significant obstacles: it requires unprecedented international cooperation, restrictions on the use of AI chips, and limitations on scientific research. However, we believe these measures are proportionate to the risks and represent the most viable path to ensuring that ASI, when eventually developed, will be beneficial for humanity. The staged implementation in the appendices provides a pathway forward even in the absence of immediate political consensus.

As AI capabilities continue their rapid advancement, the window for establishing robust verification and monitoring regimes narrows. We hope this work can serve as a foundation for serious policy discussions and international negotiations, guiding humanity toward a future where the benefits of advanced AI can be realized without courting civilizational catastrophe.

Acknowledgments

We are thankful to the following people and numerous others for feedback on earlier versions of this work. They do not necessarily endorse the views expressed here.

Nate Soares, Mitch Howe, Joseph Rogero, Robi Rahman, Rob Bensinger, Malo Bourgon, Daniel Kokotajlo, Thomas Larsen, Oscar Delaney, Naci Cankaya.

LLMs were used to help with various parts of this report, including background research, text drafting, and text refinement. All final text was reviewed by a human and the vast majority was written by a human.

References

[1] Onni Aarne, Tim Fist, and Caleb Withers. Secure, governable chips. *Center for a New American Security. https://www.cnas.org/publications/reports/secure-governable-chips*, 2024.

[2] Anthony Aguirre. Keep the Future Human: Why and How We Should Close the Gates to AGI and Superintelligence, and What We Should Build Instead. *arXiv preprint arXiv:2311.09452*, 2025. URL https://arxiv.org/abs/2311.09452v4.

[3] Sam Altman. Machine intelligence, part 1. Blog post, February 2015. URL https://blog.samaltman.com/machine-intelligence-part-1. Accessed: 2025-10-10.

[4] Dario Amodei. Machines of Loving Grace, October 2024. URL https://darioamodei.com/machines-of-loving-grace.

[5] Dario Amodei. Interview at Axios AI Summit. Axios AI Summit, September 2025. URL https://x.com/axios/status/1968387815726891268. Interviewed by Jim VandeHei.

[6] Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

[7] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

[8] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of LLM agents. *arXiv preprint arXiv:2410.09024*, 2024.

[9] Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J. Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J. Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot Jailbreaking. November 2024. URL https://openreview.net/forum?id=cw5mgd71jW.

[10] Anthropic. How difficult is AI alignment? | Anthropic Research Salon. YouTube, January 2025. URL https://youtu.be/IPmt8b-qLgk?t=772. Jan Leike speaking about the Alignment Science Team. Accessed: 2025-10-10.

[11] Anthropic. Claude 3.7 Sonnet System Card, February 2025. URL https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf. Accessed: 2025-10-10.

[12] Anthropic. Responsible scaling policy, version 2.2. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf.

[13] Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: a model of artificial intelligence development. *AI & society*, 31(2):201–206, 2016.

[14] Artificial Analysis. AI Trends: Countries. Artificial Analysis, 2025. URL https://artificialanalysis.ai/trends#countries. Accessed: 2025-10-10.

[15] Artificial Analysis. Comparison of models: Intelligence, performance & price analysis. https://artificialanalysis.ai/models/, 2026. Accessed: 2026-05-11.

[16] Leopold Aschenbrenner. Situational Awareness: The Decade Ahead, June 2024. URL https://situational-awareness.ai/.

[17] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[18] Mauricio Baker, Gabriel Kulp, Oliver Marks, Miles Brundage, and Lennart Heim. Verifying International Agreements on AI: Six Layers of Verification for Rules on Large-Scale AI Development and Deployment. *arXiv preprint arXiv:2507.15916*, 2025.

[19] Peter Barnett and Aaron Scher. AI Governance to Avoid Extinction: The Strategic Landscape and Actionable Research Questions. *arXiv preprint arXiv:2505.04592*, 2025.

[20] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

[21] Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, September 2024. Published: 09 Sept 2024.

[22] Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, and Markus Anderljung. Societal Adaptation to Advanced AI, January 2025. URL http://arxiv.org/abs/2405.10295. arXiv:2405.10295 [cs].

[23] Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. Demonstrating specification gaming in reasoning models. *arXiv preprint arXiv:2502.13295*, 2025.

[24] Miles Brundage. My recent lecture at Berkeley and a vision for a "CERN for AI", December 2024. URL https://milesbrundage.substack.com/p/my-recent-lecture-at-berkeley-and.

[25] Justin B Bullock, Samuel Hammond, and Seb Krier. AGI, Governments, and Free Societies. *arXiv preprint arXiv:2503.05710*, 2025.

[26] Vitalik Buterin. My techno-optimism, November 2023. URL https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html.

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

[28] Satish Chandra and Maxim Tabachnyk. AI in software engineering at Google: Progress and the path ahead. Google Research Blog, June 2024. URL https://research.google/blog/ai-in-software-engineering-at-google-progress-and-the-path-ahead/. Accessed: 2025-10-10.

[29] Neil Chowdhury, Daniel Johnson, Vincent Huang, Jacob Steinhardt, and Sarah Schwettmann. Investigating truthfulness in a pre-release o3 model. Transluce, April 2025. URL https://transluce.org/investigating-o3-truthfulness. Accessed: 2025-10-10.

[30] ControlAI. What leaders say about AI, 2025. URL https://controlai.com/quotes. Accessed: 2025-11-15.

[31] Forrest W. Crawford, Kyle Webster, Gerald L. Epstein, Derek Roberts, Joseph Fair, and Sella Nevo. *Securing Commercial Nucleic Acid Synthesis*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/RRA3329-1.

[32] Tom Davidson, Lukas Finnveden, and Rose Hadshar. AI-Enabled Coups: How a Small Group Could Use AI to Seize Power. 2025. URL https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power. Accessed: 2025-04-17.

[33] Tom Davidson, Rose Hadshar, and Will MacAskill. Three types of intelligence explosion. 2025. URL https://www.forethought.org/research/three-types-of-intelligence-explosion. Accessed: 2025-03-28.

[34] Luke Drago and Rudolf Laine. The intelligence curse. https://intelligence-curse.ai/, April 2025. Accessed: 2025-10-10.

[35] Brian David Ehrhart, Dusty Marie Brooks, Alice Baca Muna, and Chris Bensdotter LaFleur. Evaluation of risk acceptance criteria for transporting hazardous materials. 2020.

[36] Epoch AI. AI Benchmarking Hub, 11 2024. URL https://epoch.ai/benchmarks/. Accessed: 2025-10-10.

[37] Epoch AI. Data on AI Models, 07 2025. URL https://epoch.ai/data/ai-models. Accessed: 2025-10-10.

[38] A. Erben, M. Negele, L. Heim, J. Sevilla, D. Fernandez Llorca, and E. Gomez. Training Compute Thresholds - Key Considerations for the EU AI Act. Policy Development KJ-01-25-458-EN-N, Luxembourg (Luxembourg), 2025.

[39] Ege Erdil, Andrei Potlogea, Tamay Besiroglu, Edu Roldan, Anson Ho, Jaime Sevilla, Matthew Barnett, Matej Vrzla, and Robert Sandler. GATE: An Integrated Assessment Model for AI Automation. *arXiv preprint arXiv:2503.04941*, 2025.

[40] Kevin M Esvelt. Delay, detect, defend: preparing for a future in which thousands can release new pandemics. *GCSP Geneva Papers*, 2022.

[41] Daniel Eth and Tom Davidson. Will AI R&D Automation Cause a Software Intelligence Explosion? 2025. URL https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion. Accessed: 2025-04-11.

[42] Google DeepMind. Frontier safety framework, version 3.0. Technical report, Google DeepMind, September 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf.

[43] Katja Grace, Julia Fabienne Sandkühler, Harlan Stewart, Benjamin Weinstein-Raun, Stephen Thomas, Zach Stein-Perlman, John Salvatier, Jan Brauner, and Richard C Korzekwa. Thousands of AI authors on the future of AI. *Journal of Artificial Intelligence Research*, 84, 2025.

[44] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI Control: Improving Safety Despite Intentional Subversion, July 2024. URL http://arxiv.org/abs/2312.06942. arXiv:2312.06942 [cs].

[45] Ben Harack, Robert F. Trager, Anka Reuel, David Manheim, Miles Brundage, Onni Aarne, Aaron Scher, Yanliang Pan, Jenny Xiao, Kristy Loke, Sumaya Nur Adan, Guillem Bas, Nicholas A. Caputo, Julia C. Morse, Janvi Ahuja, Isabella Duan, Janet Egan, Ben Bucknall, Brianna Rosen, Renan Araujo, Vincent Boulanin, Ranjit Lall, Fazl Barez, Sanaa Alvira, Corin Katzke, Ahmad Atamli, and Amro Awad. Verification for international AI governance. Technical report, Oxford Martin School, AI Governance Initiative, July 2025. URL https://aigi.ox.ac.uk/publications/verification-for-international-ai-governance/. Accessed: 2025-10-10.

[46] Jason Hausenloy, Andrea Miotti, and Claire Dennis. Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI, October 2023. URL http://arxiv.org/abs/2310.09217. arXiv:2310.09217 [cs].

[47] Lennart Heim and Leonie Koessler. Training Compute Thresholds: Features and Functions in AI Regulation, August 2024. URL http://arxiv.org/abs/2405.10799. arXiv:2405.10799 [cs].

[48] Dan Hendrycks, Eric Schmidt, and Alexandr Wang. Superintelligence Strategy: Expert Version, March 2025. URL http://arxiv.org/abs/2503.05628. arXiv:2503.05628 [cs].

[49] Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models, March 2024. URL http://arxiv.org/abs/2403.05812. arXiv:2403.05812 [cs].

[50] Sara Hooker. On the limitations of compute thresholds as a governance strategy. *arXiv preprint arXiv:2407.05694*, 2024.

[51] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[52] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

[53] Leonie Koessler, Jonas Schuett, and Markus Anderljung. Risk thresholds for frontier AI. *arXiv preprint arXiv:2406.14713*, 2024.

[54] Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman. *Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/WRA3056-1.

[55] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, et al. Measuring AI Ability to Complete Long Tasks. *arXiv preprint arXiv:2503.14499*, 2025.

[56] Ange Lavoipierre. AI's dark in-joke. ABC News, July 2023. URL https://www.abc.net.au/news/2023-07-15/whats-your-pdoom-ai-researchers-worry-catastrophe/102591340. Accessed: 2025-10-10.

[57] Jan Leike and Ilya Sutskever. Introducing superalignment. OpenAI Blog, July 2023. URL https://openai.com/index/introducing-superalignment/. Accessed: 2025-10-10.

[58] Simon Lermen. Why I don't believe Superalignment will work. LessWrong, September 2025. URL https://www.lesswrong.com/posts/kyBGcHfzfZziHm5xL/why-i-don-t-believe-superalignment-will-work. Accessed: 2025-10-10.

[59] Eli Lifland, Nikola Jurkovic, and FutureSearch. Timelines forecast: Forecasting time to automated superhuman coders. AI 2027, April 2025. URL https://ai-2027.com/research/timelines-forecast. Accessed: 2025-10-10.

[60] Andrew London. Elon Musk warns 'AI is a fundamental risk to the existence of human civilization', July 2017. URL https://www.techradar.com/news/elon-musk-warns-ai-is-a-fundamental-risk-to-the-existence-of-human-civilization.

[61] Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J Ritchie, Soren Mindermann, Ethan Perez, Kevin K Troy, and Evan Hubinger. Agentic Misalignment: How LLMs Could Be Insider Threats. *arXiv preprint arXiv:2510.05179*, 2025.

[62] METR. Recent frontier models are reward hacking, 2025. URL https://metr.org/blog/2025-06-05-recent-reward-hacking/. Accessed 10 October 2025.

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

[64] Elon Musk. Interview with Garry Tan. X (Twitter), June 2025. URL https://x.com/SawyerMerritt/status/1935809018066608510. Accessed: 2025-10-10.

[65] Richard Ngo, Lawrence Chan, and Sören Mindermann. The Alignment Problem from a Deep Learning Perspective, March 2025. URL http://arxiv.org/abs/2209.00626. arXiv:2209.00626 [cs].

[66] Michael Nielsen. Notes on Differential Technological Development, January 2024. URL https://michaelnotebook.com/dtd/index.html.

[67] Jordan Novet and Jonathan Vanian. Satya Nadella says as much as 30% of Microsoft code is written by AI. CNBC, April 2025. URL https://www.cnbc.com/2025/04/29/satya-nadella-says-as-much-as-30percent-of-microsoft-code-is-written-by-ai.html. Accessed: 2025-10-10.

[68] OpenAI. Expanding on what we missed with sycophancy. https://openai.com/index/expanding-on-what-we-missed-with-sycophancy/, May 2025. Accessed: 2025-10-10.

[69] OpenAI. OpenAI Model Spec, February 2025. URL https://model-spec.openai.com/2025-02-12.html.

[70] Ananthi Al Ramiah, Raymond Koopmanschap, Josh Thorsteinson, Sadruddin Khan, Jim Zhou, Shafira Noh, Joep Meindertsma, and Farhan Shafiq. Toward a global regime for compute governance: Building the pause button. *arXiv preprint arXiv:2506.20530*, 2025.

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

[72] Jaime Sevilla and Edu Roldán. Training compute of frontier AI models grows by 4-5x per year, 2024. URL https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year. Accessed: 2025-10-10.

[73] G Smith. Licensing Frontier AI Development: Legal Considerations and Best Practices. *The Lawfare Institute website (2024). Publisher: The Lawfare Institute*, 2024.

[74] Superintelligence Statement. Statement on superintelligence, oct 2025. URL https://superintelligence-statement.org/.

[75] The Logan Bartlett Show. Anthropic CEO on Leaving OpenAI and Predictions for Future of AI. Podcast, October 2023. URL https://www.youtube.com/watch?v=gAaCqj6j5sQ. Accessed: 2025-10-10.

[76] U.S. Nuclear Regulatory Commission. An approach for using probabilistic risk assessment in risk-informed decisions on plant-specific changes to the licensing basis. Draft Regulatory Guide DG-1285, U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, July 2017. URL https://www.nrc.gov/docs/ML1721/ML17213A676.pdf. Accessed: 2025-10-10.

[77] Tony T. Wang, John Hughes, Henry Sleight, Rylan Schaeffer, Rajashree Agrawal, Fazl Barez, Mrinank Sharma, Jesse Mu, Nir Shavit, and Ethan Perez. Jailbreak Defense in a Narrow Domain: Limitations of Existing Methods and a New Transcript-Classifier Approach, December 2024. URL http://arxiv.org/abs/2412.02159. arXiv:2412.02159 [cs].

[78] John Wentworth. The Case Against AI Control Research. LessWrong, January 2025. URL https://www.lesswrong.com/posts/8wBN8cdNAv3c7vt6p/the-case-against-ai-control-research. Accessed: 2025-10-10.

[79] Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts, November 2024. URL http://arxiv.org/abs/2411.15114. arXiv:2411.15114 [cs].

[80] Wikipedia. P(doom) — Wikipedia, the free encyclopedia, 2025. URL https://en.wikipedia.org/w/index.php?title=P(doom)&oldid=1314921337. [Online; accessed 10-October-2025].

[81] Eliezer Yudkowsky. AGI Ruin: A List of Lethalities. June 2022. URL https://www.alignmentforum.org/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities.

[82] Eliezer Yudkowsky and Nate Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Hachette Book Group, September 2025. ISBN 9780316595643.

[83] Philip Zelikow, Mariano-Florentino Cuéllar, Eric Schmidt, and Jason Matheny. Defense against the AI dark arts: Threat assessment and coalition defense. Report, Hoover Institution, Stanford, CA, 12 2024. A Publication of the Hoover Institution.

[84] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.