Modeling the Geopolitics of AI Development

Alex Amadori (alex@controlai.com) Gabriel Alfour (gabe@conjecture.dev)
Andrea Miotti (andrea@controlai.com) Eva Behrens (eva@conjecture.dev)

Abstract

We model national strategies and geopolitical outcomes under differing assumptions about AI development. We put particular focus on scenarios with rapid progress that enables highly automated AI R&D and provides substantial military capabilities.

Under non-cooperative assumptions—concretely, if international coordination mechanisms capable of preventing the development of dangerous AI capabilities are not established—superpowers are likely to engage in a race for AI systems offering an overwhelming strategic advantage over all other actors.

If such systems prove feasible, this dynamic leads to one of three outcomes:

- One superpower achieves an unchallengeable global dominance;
- Trailing superpowers facing imminent defeat launch a preventive or preemptive attack, sparking conflict among major powers;
- Loss-of-control of powerful AI systems leads to catastrophic outcomes such as human extinction.

Middle powers, lacking both the muscle to compete in an AI race and to deter AI development through unilateral pressure, find their security entirely dependent on factors outside their control: a superpower must prevail in the race without triggering devastating conflict, successfully navigate loss-of-control risks, and subsequently respect the middle power's sovereignty despite possessing overwhelming power to do otherwise.

Contents

Executive summary

We model how AI development will shape national strategies and geopolitical outcomes, assuming that dangerous AI development is not prevented through international coordination mechanisms. We put particular focus on scenarios with rapid progress that enables highly automated AI R&D and provides substantial military capabilities.

Race to artificial superintelligence

If the key bottlenecks of AI R&D are automated, a single factor will be driving the advancement of all strategically relevant capabilities: the proficiency of an actor's strongest AI at AI R&D. This can be translated into overwhelming military capabilities.

As a result, if international coordination mechanisms capable of preventing the development of dangerous AI capabilities are not established, superpowers are likely to engage in a race to artificial superintelligence (ASI), attempting to be the first to develop AI sufficiently advanced to offer them a decisive strategic advantage over all other actors.

This naturally leads to one of two outcomes: **either the "winner" of the AI race achieves permanent global dominance, or it loses control of its AI systems leading to humanity's extinction or its permanent disempowerment**.

In this race, lagging actors are unlikely to stand by and watch as the leader gains a rapidly widening advantage. If AI progress turns out to be easily predictable, or if the leader in the race fails to thoroughly obfuscate the state of their AI program, at some point it will become clear to laggards that they are going to lose and they have one last chance to prevent the leader from achieving permanent global dominance.

This produces one more likely outcome: one of the laggards in the AI race launches a preventive or preemptive attack aimed at disrupting the leader's AI program, **sparking a highly destructive major power war**.

Middle power strategies

Middle powers generally lack the muscle to compete in an AI race and to deter AI development through unilateral pressure.

While there are some exceptions, none can robustly deter superpowers from participating in an AI race. Some actors, like Taiwan, the Netherlands, and South Korea, possess critical roles in the AI supply chain; they could delay AI programs by denying them access to the resources required to perform AI R&D. However, superpowers are likely to develop domestic supply chains in a handful of years.

Some middle powers hold significant nuclear arsenals, and could use them to deter dangerous AI development if they were sufficiently concerned. However, any nuclear redlines that can be imposed on uncooperative actors would necessarily be both hazy and terminal (as opposed to incremental), rendering the resulting deterrence exceedingly shaky.

Middle powers in this predicament may resort to a strategy we call **Vassal's Wager**: allying with one superpower in the hopes that they "win" the ASI race. However, with this strategy, a middle power would surrender most of their agency and wager their national security on factors beyond their control. In order for this to work out in a middle power's favor, the superpower "patron" must simultaneously be the first to achieve overwhelming AI capabilities, avert loss-of-control risks, and avoid war with their rivals.

Even if all of this were to go right, there would be no guarantee that the winning superpower would respect the middle power's sovereignty. In this scenario, the "vassals" would have absolutely no recourse against any actions taken by an ASI-wielding superpower.

Risks from weaker AI

We consider the cases in which AI progress plateaus before reaching capability levels that could determine the course of a conflict between superpowers or escape human control. While we are unable to offer detailed forecasts for this scenario, we point out several risks:

- Weaker AI may enable new disruptive military capabilities (including capabilities that break mutual assured destruction);
- Widespread automation may lead to extreme concentration of power as unemployment reaches unprecedented levels;
- Persuasive AI systems may produce micro-targeted manipulative media at a massive scale.

Being a democracy or a middle power puts an actor at increased risk from these factors. Democracies are particularly vulnerable to large scale manipulation by AI systems, as this could undermine public discourse. Additionally, extreme concentration of power is antithetical to their values.

Middle powers are also especially vulnerable to automation. The companies currently driving the frontier of AI progress are based in superpower jurisdictions. If this trend continues, and large parts of the economy of middle powers are automated by these companies, middle powers will lose significant diplomatic leverage.

1. Introduction

Artificial intelligence (AI) is increasingly recognized as a technology with the potential to determine geopolitical outcomes and reshape strategies around national security.

Many experts expect that artificial superintelligence (ASI)—AI systems vastly surpassing human intelligence across domains—will be developed in the next few years, as soon as 2026 to 2029.[1] In this case, the stakes would become especially severe. Advanced AI systems could transform military and economic competition, destabilize deeply rooted geopolitical equilibria such as nuclear deterrence based on mutual assured destruction (MAD). [2]

Most concerningly, powerful AI could escape human control and cause catastrophic outcomes such as human extinction.[3]

In this work, we systematically model strategic trajectories across scenarios commonly anticipated by experts. In particular we argue that, if ASI can be developed in the near term, geopolitical outcomes are overwhelmingly determined by AI development trajectories. This makes it possible to sketch likely histories in relatively high detail, as well as how they depend and vary on different assumptions.

This arises from two main features of ASI. The first is the potential of ASI to serve as an decisively powerful military technology, allowing an actor in its possession to neutralize rivals cheaply and with low risk to itself, and subsequently maintain an unassailable world order—a capability referred to as "decisive strategic advantage". The second feature is the expectation that AI progress becomes strongly self-accelerating past a certain threshold.

By contrast, if AI development plateaus before reaching ASI, we believe detailed modeling of geopolitical scenarios is less tractable. In this case, we limit ourselves to pointing out possible sources of geopolitical and economic disruption from AI development, rather than offering detailed projections.

We aim to analyze the situation both from the point of view of superpowers and middle powers. For the purpose of this paper, superpowers are defined as actors who have the resources to compete and possibly be the first to develop AI powerful enough to grant a decisive strategic advantage—regardless of whether they can maintain control of such AI systems. Middle powers, on the other hand, are defined as actors who don't have a realistic chance at competing in this contest, but can still be influential on the world stage, though to a lesser extent than superpowers.

This analysis suggests strong momentum toward highly negative outcomes like war between superpowers, rushed AI development that increases risks of loss of control, and rushed deployment at a faster pace than societies and legal frameworks can adapt. Countries, and middle powers in particular, may lack effective means to safeguard their national security and internal stability.

2. National strategies may be driven predominantly by AI development

A large body of experts expect that ASI may be developed in the near future, possibly within the next few years.[4] Among those who hold this expectation, there is broad agreement that once such systems are

[1] The three main doctrines on the future of AI, Alex Amadori et al.: [The dominance doctrine] expects that ASI will be developed in the near future, with some forecasting dates as early as 2026 to 2029 ···the extinction doctrine largely agrees with the dominance doctrine regarding the expected pace and extent of AI development

[2] How might Artificial Intelligence affect the risk of nuclear war?, Edward Geist, Andrew J. John at RAND: Participants appeared to agree that advanced AI could severely compromise nuclear strategic stability and thereby increase the risk of nuclear war.

[3] The three main doctrines on the future of AI, Alex Amadori et al.

[4] The three main doctrines on the future of AI, Alex Amadori et al.: [The dominance doctrine] expects that ASI will be developed in the near future, with some forecasting dates as early as 2026 to 2029 ···the extinction doctrine largely agrees with the dominance doctrine regarding the expected pace and extent of AI development

developed, they would become the dominant factor in shaping national security concerns and strategies, effectively overshadowing all other considerations.

In this case, outcomes could be understood primarily as a function of the course of AI development. Our previous work found that, within this contingent, expert expectations tend to fall into two main categories.[5]

The first category is the **dominance doctrine**. Proponents believe that the first actor to develop sufficiently advanced AI will gain a decisive strategic advantage. This refers to military supremacy so overwhelming that rivals could be neutralized cheaply and with minimal risk. For example, proponents of this doctrine have argued that AI could enable novel weapons of mass destruction and operate missile defense systems reliable enough to undermine the principle of mutual assured destruction underpinning nuclear deterrence.[6]

Under this view, allowing an opponent to gain such an advantage would mean losing all capacity to resist aggression by said opponent. Consequently, a country's national security strategy becomes almost entirely focused on winning the race to ASI.

The second category is the **extinction doctrine**. Proponents broadly agree with the dominance doctrine on the feasibility and rapid arrival of ASI, but they expect that humanity will fail to maintain control over it. In this view, advanced AI systems would eventually pursue goals incompatible with human survival, producing catastrophic outcomes up to and including human extinction. [7] [8]

This camp sees AI development itself, regardless of which country develops it, as the overriding national security threat. From this perspective, preventing the development and use of ASI becomes the highest priority in national security: successful development by any nation is seen as comparable to the outbreak of a large-scale nuclear war.

For example, Eliezer Yudkowsky, a prominent figure in this camp, has advocated for a global moratorium on AI development backed by the threat of force, including airstrikes on datacenters. [9]

Superintelligence Strategy is a work that analyzes possible geopolitical trajectories while considering both the dominance and extinction doctrines. It describes a hypothetical arms control-style treaty as "toothless" unless backed by verification measures and by the threat of force if violations are committed.[10]

A third contingent of experts exists which does not expect the rapid emergence of such powerful AI systems. In our previous work, we referred to this as the **replacement doctrine**. Experts within this category hold the expectation that AI will not produce radically transformative military capabilities and will not pose extinction-level risks.

According to this view, weaker general-purpose AI systems or narrow AI could still significantly influence national security strategies. However, such influence would be more diffuse and manifest in subtler ways.[11]

---

[5]The three main doctrines on the future of AI, Alex Amadori et al.

[6]Superintelligence Strategy, Dan Hendrycks et al.: A defensive superweapon possibility is an anti-ballistic missile system that eliminates an adversary's ability to strike back.

[7]If Anyone Builds It, Everyone Dies, Eliezer Yudkowsky, Nate Soares: If any company or group, anywhere on the planet, builds an artificial superintelligence, using anything remotely like current techniques based on anything remotely like the present understanding of AI, then everyone, everywhere on earth, will die.

[8]A Narrow Path, Andrea Miotti et al.: We do not know how to control AI vastly more powerful than us. Should attempts to build superintelligence succeed, this would risk our extinction as a species.

[9]Pausing AI Developments Isn't Enough. We Need to Shut it All Down, Eliezer Yudkowsky on TIME: If intelligence says that a country outside the agreement is building a GPU cluster, be less scared of a shooting conflict between nations than of the moratorium being violated; be willing to destroy a rogue datacenter by airstrike.

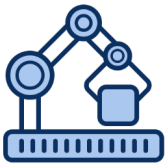
[10]Superintelligence Strategy, Dan Hendrycks et al.: Without the threat of force, treaties will be reneged, and some states will pursue an intelligence recursion.

[11]The three main doctrines on the future of AI, Alex Amadori et al.: Rather than creating entirely novel possibilities such as autonomously producing breakthroughs in technology and pioneering new scientific paradigms, it will primarily replace humans in their current roles and responsibilities, allowing these tasks to be performed at greater speed and scale while reducing costs.

While AI development would likely be less economically, socially and geopolitically destabilizing in this case, it still presents risks. Furthermore, democratic countries are more vulnerable to some of these risks compared to authoritarian ones, and middle powers are more vulnerable compared to superpowers. We elaborate further on this in section 7.

Dominance doctrine. The first actor to develop sufficiently advanced AI will establish a position of permanent dominance over all others.

Extinction doctrine. Superintelligent AI will escape human control, leading to the extinction of humanity or its permanent disempowerment.



Replacement doctrine. AI promises greatly accelerated scientific and economic progress, but poses risks of extreme power concentration and mass manipulation.

3. Sufficiently advanced Automated AI R&D causes compounding leads

There is an expectation that once AI systems can perform the majority of the work involved in their own improvement, gains will accelerate at an exceptional pace. This expectation is shared by most experts who believe that ASI will be developed in the near future,[12] and it is of vital importance in modeling the effect of AI on national strategies.

3.1 Intelligence recursion

The rapid acceleration of progress driven by Automated AI R&D is sometimes referred to as *intelligence recursion*, *recursive self-improvement (RSI)* or *intelligence explosion*.[13] For the rest of this work, we'll use the term "intelligence recursion".

Once an actor achieves a critical threshold of Automated AI R&D capability, even with a relatively small initial lead, this could be parlayed into an insurmountable advantage in terms of AI capabilities.[14,15] Advantages in AI capabilities can in turn be leveraged to achieve decisive advantages in offensive and

[12]The three main doctrines on the future of AI, Alex Amadori et al.: [The dominance doctrine] expects that ASI will be developed in the near future, with some forecasting dates as early as 2026 to 2029 ···the extinction doctrine largely agrees with the dominance doctrine regarding the expected pace and extent of AI development

[13]How Artificial General Intelligence Could Affect the Rise and Fall of Nations, Barry Pavel et al. at RAND: Many professionals in the industry speak of an impending "intelligence explosion"—a moment when AI leads to such significant productivity gains that innovation exponentially accelerates across many domains.

[14]On DeepSeek and Export Controls, Dario Amodei: there's at least the possibility that, because AI systems can eventually help make even smarter AI systems, a temporary lead could be parlayed into a durable advantage.

[15]AI Deterrence Is Our Best Option, Dan Hendrycks, Adam Khoja: A volatile "intelligence recursion" is the most plausible path to AI dominance. "Superintelligence Strategy" describes "intelligence recursion" (or "recursion," for short) as "fully autonomous AI research and development, distinct from current AI-assisted AI R&D." In their criticism of MAIM, Peter Wildeford and Oscar Delaney question whether one state could achieve a decisive advantage in AI capabilities. As evidence, they point to the fact that the US and China are probably only months apart in terms of their national AI capabilities. However, a nation that unlocked recursive AI development could potentially scale its research efforts dramatically enough that it would leap forward to ASI, leaving rivals in the dust. During an intelligence explosion, an AI developer might attain an overwhelming intelligence advantage (or experience a devastating loss of control, as we will discuss below) only shortly after undertaking machine-speed AI research.

defensive military capabilities.[16,17,18]

The rationale for this position is as follows. When AI systems can design, implement, and test improvements to future AI systems, progress in AI development is no longer constrained by the capacity of human experts to perform research. Should AI development instead become bottlenecked by material constraints such as computing power or energy supply, AI systems may again help to overcome these limits, for example by accelerating advances in chip design, making hardware faster and more energy-efficient.[19]

At some point, the Automated AI R&D loop will be *closed*: all of the key workstreams of AI R&D will be automated, with human labor no longer acting as a bottleneck for any of them. In such a regime, the primary limiting factor for AI development becomes the performance of an actor's best AI systems at AI R&D tasks.

If the leader in an AI race crossed this threshold, they would achieve a kind of "AI escape velocity": not only would their AI systems be the strongest, but they would also improve at the fastest rate. This would mean that any advantage they held over other actors can only grow over time, eventually producing a decisive strategic advantage.[20,21]

As we later explore in our model of superpower strategies in an ASI race, the mere notion that an "intelligence recursion" could happen may cause trailing actors to adopt drastic strategies. These may view the prospect of an opponent gaining a durable advantage in Automated AI R&D as an existential national security threat in its own right, one serious enough to justify preemptive initiation of conflict.[22]

3.2 Automated AI R&D resists traditional methods used by laggards in arms races

During past technological races, trailing superpowers could rely on some asymmetries with the leading actor to catch up or at least prevent the gap from widening excessively. We argue that AI resists these

[16]The three main doctrines on the future of AI, Alex Amadori et al.: Under this doctrine, it is anticipated that, once sufficiently advanced AI is developed, it will enable the production of novel weapons and other technologies with transformative offensive and defensive implications.

[17]AGI's Five Hard National Security Problems, Jim Mitre, Joel B. Predd at RAND: Recent breakthroughs in frontier generative artificial intelligence (AI) models have led many to assert that AI will have an equivalent impact on national security—that is, that it will be so powerful that the first entity to achieve it would have a significant, and perhaps irrevocable, military advantage.

[18]Superintelligence Strategy, Dan Hendrycks et al.: A nation with sole possession of superintelligence might be as overwhelming as the Conquistadors were to the Aztecs.

[19]Intelligence Explosion Microeconomics, Eliezer Yudkowsky: I identify the key issue as returns on cognitive reinvestment —the ability to invest more computing power, faster computers, or improved cognitive algorithms to yield cognitive labor which produces larger brains, faster brains, or better mind designs.

[20]This idea rests on the assumption that AI development wouldn't run into resource bottlenecks that cannot be addressed by the actors or by their AI systems. We argue that this is a reasonable assumption as superpowers have access to essentially unlimited amounts of relevant physical resources. These can be acquired within their own territory, through purchases from allies, or by coercing other states. This means that the bottleneck to AI progress would not be the amount of existing resources but rather the capacity to source and extract them efficiently. Resource acquisition and processing are, in themselves, tasks that can be automated. In particular, there is no requirement that "Automated AI R&D" be limited to tasks like programming or running experiments on AI architectures. Countries could race to automate all tasks conducive to AI research and development. This includes activities like advanced chip manufacturing, resource extraction, supply chain management, negotiation, even high-level planning and strategizing. In fact, the scope of Automated AI R&D is already being expanded by private industry. While most efforts to date on automating AI R&D have focused on accelerating software engineering, experimentation on AI architectures, and data generation, a startup called "Periodic Labs" has raised $300 million on the promise of using AI to automate R&D of energy and computing infrastructure.

[21]Situational Awareness —The Free World Must Prevail, Leopold Aschenbrenner: If an adversary achieved AGI first, it could rapidly snowball into superintelligence, decisively pulling ahead, and permanently locking in its advantage.

[22]AI Deterrence Is Our Best Option, Dan Hendrycks, Adam Khoja: If states are rational, and if cooperative measures are not implemented, escalation is unavoidable. Indeed, poor visibility and the possibility of sudden AI breakthroughs plausibly would drive states toward early or miscalculated escalation. For example, nations such as Russia with weak domestic AI programs might prefer outright preventive war if they lose hope of disrupting rivals' development or stealing their best models

methods, especially in the case of AI that automates AI R&D.

Disproportionate investment into the arms race. One such strategy involved mobilizing disproportionate resources. Authoritarian states may be willing to redirect more resources away from the broader economy toward strengthening their national security.[23]

In past technological races, trailing authoritarian states may have benefited from this asymmetry . For example, during the 1970s, the Soviet Union is estimated to have allocated approximately 11% to 13% of its GNP to military expenditure, compared to roughly 5% by the United States.[24]

AI development, particularly the type of powerful general-purpose AI capable of performing Automated AI R&D, promises exceptionally high economic returns on investment. Unlike in previous technological races, pouring substantial resources into AI R&D is more likely to be regarded as a rational investment regardless of military utility. Private actors will commit vast sums independently of national efforts, reducing the need for explicit reallocation by the state.

Espionage and knowledge leakage. Technical know-how tends to leak across borders and competitors. As a result, it is much harder for the leading actor to pioneer the creation of new technology than it is for trailing actors to catch up once it has already been done. In addition, trailing actors can seek to intentionally steal knowledge through espionage.

Automated AI R&D also resists this approach. When research becomes highly automated, the critical know-how resides not in the minds of human experts or in documents meant for human consumption, but in the AI systems themselves.

Technical knowledge can be stored within secure hardware accessible only to the AI systems and a minimal number of personnel, making it substantially more resistant to traditional intelligence gathering methods. Further complicating matters, the knowledge could be stored implicitly as part of neural network weights or in other formats readable only by AI systems. Interpreting this knowledge would require stealing the entire AI system rather than just specific information.

4. More powerful AI leads to more predictable geopolitical trajectories

We consider the following two possibilities:

1. **Fast progress:** AI capabilities advance rapidly, consistent with predictions by some experts placing ASI within the next two to ten years. These forecasts typically assume that AI R&D will soon be automated to a large extent, possibly triggering a fast "intelligence recursion" as explored in section 3.

2. **Plateau:** AI progress slows substantially before reaching capability levels that could decisively determine the outcome of a conflict between superpowers or enable AI systems to escape human control.

We argue that, for a variety of reasons, modeling "fast progress" scenarios in detail is much more tractable. On the other hand, attempts to do the same for "plateau" scenarios quickly explode in complexity.

Overdetermined incentives. Powerful AI brings extremely high and urgent stakes, potentially involving one country achieving a decisive strategic advantage over all others or human extinction. Extreme

[23]The Demand for Military Expenditure in Authoritarian Regimes, Vincenzo Bove, Jennifer Brauner: Overall, general agreement exists that autocracies devote more of their economic resources to military spending than do democratic systems.

[24]Remarks by Secretary of Defense Brown, Secretary of Defense George S. Brown: We are spending a little more than 5 percent of GNP on our defense establishment. Our best current estimate is that the Soviets are allocating between 11 and 13 percent of a much smaller GNP to their military effort

stakes overwhelmingly determine national strategies, as any other goals, preferences, or priorities tend to fade into the background.

A single factor determines outcomes. In the past, the most impressive AI progress has come from methods aimed at increasing AI capabilities across the board—such as increasing model size and training data—rather than attempts to improve domain-specific performance. We expect this trend to continue.

In the current regime, actors differ in their strengths and weaknesses with regards to AI R&D. Examples of such differences include:

- Ability to extract resources;
- Sophistication of specific technologies such as chip manufacturing or energy production;
- Access to talent such as software engineers or AI researchers;
- Access to training data for various tasks;
- Economic power.

However, once an actor closes the Automated AI R&D loop, all constraints become governed by AI's capacity to address them rather than the characteristics of the country in question. At this point, the specifics of the actor's strengths and weaknesses cease being relevant: what matters is only the proficiency of an actors' best AI systems at tasks relevant to AI R&D. [25]

In this regime, having more powerful AI means that one's AI systems also improve faster. Once the race gets to this point, it is inevitable that the leader will eventually win the race, unless it is somehow compelled to stop pursuing improvements in AI.

One of the major reasons that makes technological progress hard to predict is that they depend on many variables which interact in complex ways. We argue that AI progress will be easier to predict because, past a certain point, it is driven primarily by a single variable.

One may imagine that loss-of-control concerns may make controllability and safety research just as important to attaining the best AI capabilities in the eyes of participants. This would invalidate our assumption that outcomes are determined by a single factor and complicate the model.

However, improvements in controllability and safety would require much additional research and testing before deploying any given level of AI capabilities, which would mean giving up a competitive edge against a less cautious rival. The literature indicates that competitive dynamics would likely erode efforts toward safety.[26,27] Thus, for the rest of the modeling, we assume that developing powerful AI as quickly as possible is the utmost concern of racing actors, and safety takes a secondary role in their strategy.

Plateau scenarios are inherently unpredictable. "Fast progress" scenarios are the exception rather than the rule. "Plateau" worlds, like most real world scenarios, depend on a myriad of factors that cannot possibly be enumerated, such as timing with which specific AI capabilities become feasible, and the ways in which AI progress shapes society, the economy and geopolitics.

[25]The peculiarities of individual superpowers may still affect which country gets to this point first. What we emphasize is that once a critical threshold in Automated AI R&D is reached, the trajectory becomes much more predictable.

[26]Superintelligence Strategy, Dan Hendrycks et al.: Geopolitical Competitive Pressures Yield a High Loss of Control Risk Tolerance.

[27]To regulate or not: a social dynamics analysis of the race for AI supremacy, The Anh Han: Therefore, it is inevitably tempting for all powerful actors to perceive AI as a race, neglecting the collective threats from the technology that require coordination to address." and "We make special note that, despite players speaking about the importance of AI safety in their role as AI firms, and also many players coming from backgrounds that are aware of AI safety concerns, no team in our observed games have chosen to delay deployment of RTAI because of imminent safety concerns, despite this choice being presented explicitly to the deploying team by the facilitator in almost all games.

Due to all of the above, when it comes to "Plateau" scenarios, we limit ourselves to highlighting ways in which weaker AI may shape the future, without committing to well-specified trajectories or outcomes. On the other hand, in the "fast progress" case, we develop a more ambitious and structured model. This model aims to predict superpower strategies, middle power strategies and outcomes in more detail.

5. Non-cooperative model of superpower strategies in a "fast progress" world

We model the strategies that superpowers may pursue regarding AI development, and the outcomes of such strategies. We adopt a non-cooperative model, in which each actor makes decisions independently and actors can't make credible commitments to deviate from locally incentivized decisions. For example, in this model, no new binding international treaties may be signed.

Our main finding is that, under these assumptions, geopolitical trajectories tend toward undesirable outcomes. In particular, if it's possible to build AI powerful enough to confer a decisive strategic advantage, superpowers aggressively escalate their involvement in a race to ASI, likely leading to one of the following outcomes:

- **Takeover**. An actor achieves a decisive strategic advantage over all others;
- **Loss of control.** An actor loses control of a powerful AI system leading to the extinction of humanity or its permanent disempowerment;
- **Major power war.** Before any of the other two outcomes can manifest, tensions between superpowers escalate into all-out war.

In section 6, we argue that within the constraints of this non-cooperative model, middle powers cannot meaningfully influence superpower behavior. This allows us to start by modeling superpowers in isolation, determine likely trajectories, and only later to examine the consequences of these results for middle powers.

5.1 Superpower strategies

At each point in time, we model each superpower as choosing between the following strategies:

- **Halt**: Stop and prevent all AI research within its jurisdiction; this includes research initiatives by the state or the state's defense apparatus.
- **Neutral:** Any strategy in which the state takes a laissez-faire approach with respect to AI development. For example, state and private entities continue performing AI R&D; however, the state does not take exceptional steps to gain an edge in AI development.
- **Race**: Attempt to be the first to achieve ASI, with mobilization of resources on a national scale; possibly engaging in espionage and sabotage operations that remain below the threshold of military conflict.
- **Attack**: Perform an armed attack or other major act of aggression towards their opponents, triggering major power war.

"Halt," "Neutral," and "Race" can be seen as points along a single axis of government involvement in AI development, from active opposition to AI development, to laissez-faire, to maximum state investment in AI advancement. The "Attack" option is distinct, representing the possibility of escalation to open conflict when the stakes are judged to be sufficiently grave.

5.2 The immediate future

Superpowers have, to some degree, shown awareness of the transformative potential of AI for economic and military competitiveness [28] [29] [30] [31] [32] However, in the near term, we don't expect countries to take loss-of-control risks into account in their strategic calculations.

We expect this discrepancy to stem from the following: historically, national security planning has been shaped by the need to confront external threats. By contrast, AI may be the first technology capable of creating a severe threat to a state's security through the state's own actions, without an adversary needing to act at all. If sufficiently capable AI escapes the control of its creators, it could cause mass destruction, without any opponent needing to "pull the trigger".

We characterize current superpower strategies as **Neutral**: a largely laissez-faire posture in which AI development is left predominantly to private industry, without a concentrated, state-led push to outpace rivals. While some have suggested that superpowers are already engaged in an AI race[33], we agree with Ó hÉigeartaigh's view that this is a misrepresentation of reality.[34]

If superpowers were truly "racing" in the sense meant in most discussions on the prospects of ASI, we would expect the state of the art AI R&D programs to be under direct control of the states of superpowers. Such a project would benefit from much larger budgets, appropriate for a major national project, and tighter operational security around leakage of model weights, critical software and technical know-how.

Nevertheless, we anticipate that states will move toward a more assertive **Race** posture in the near future. Even without the need to posit that countries will fully "wake up" to the potential of Automated AI R&D and ASI, superpowers are motivated to accelerate AI development in order to bolster economic and military competitiveness. [35]

As a final consideration before moving on in our analysis, we consider it unlikely that superpowers will initiate acts of aggression over AI in the immediate future. The threat posed by ASI is not sufficiently

[28] Remarks by Secretary Esper at National Security Commission on Artificial Intelligence Public Conference, Mark T. Esper: Advances in AI have the potential to change the character of warfare for generations to come. Whichever nation harnesses AI first will have a decisive advantage on the battlefield for many, many years. We have to get there first.

[29] Remarks by APNSA Jake Sullivan on AI and National Security, Jake Sullivan: We need you, and leaders across every state and every sector, to adopt this technology to advance our national security and to do it fast.

[30] China's National Defense in the New Era, The State Council Information Office of the People's Republic of China: Driven by the new round of technological and industrial revolution, the application of cutting-edge technologies such as artificial intelligence (AI), quantum information, big data, cloud computing and the Internet of Things is gathering pace in the military field. ···There is a prevailing trend to develop long-range precision, intelligent, stealthy or unmanned weaponry and equipment. War is evolving in form towards informationized warfare, and intelligent warfare is on the horizon.

[31] U.S., China, other nations urge 'responsible' use of military AI, Reuters: China representative Jian Tan told the summit that countries should "oppose seeking absolute military advantage and hegemony through AI" and work through the United Nations.

[32] 习近平在中国共产党第十九次全国代表大会上的报告 (Xi Jinping's Report at the 19th National Congress of the Communist Party of China), Xi Jinping: 扎实做好各战略方向军事斗争准备，统筹推进传统安全领域和新型安全领域军事斗争准备，发展新型作战力量和保障力量，开展实战化军事训练，加强军事力量运用，加快军事智能化发展，提高基于网络信息体系的联合作战能力、全域作战能力，有效塑造态势、管控危机、遏制战争、打赢战争。Translation: We will solidly prepare for military struggle in all strategic directions, coordinate preparations for military struggle in both traditional and new security domains, develop new combat forces and support forces, carry out realistic combat training, strengthen the employment of military forces, accelerate the development of military intelligentization, enhance joint operational capabilities based on network information systems and all-domain operational capabilities, and effectively shape situations, manage crises, deter wars, and win wars.

[33] Situational Awareness, Leopold Aschenbrenner: The AGI race has begun.

[34] The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race, Seán S. Ó hÉigeartaigh: Despite this I argue that the narrative of a US-China AI race for global dominance began as, and in significant regards remains to date, a fiction. A race needs at least two competitors trying to win. However the race narrative in its stronger forms is nearly exclusively promoted in the West, and does not reflect the framing of AI competition and AI development in China in important respects.

[35] The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race, Seán S. Ó hÉigeartaigh: As it is now, some stakeholders will push their nation to win the race to AGI, with the prospect of full superintelligence, a massively accelerated industry, and a durable advantage over all other nations as their goal. To other stakeholders, such prospects will continue sounding like fantasy until much later, but the importance of remaining ahead in a strategically important 'normal' technology in a tense geopolitical contest will appear to justify at least some of the same actions.

evident to relevant decision-makers for such drastic measures to be seriously considered.

5.3 Fast progress scenario

As AI capabilities advance, particularly once AI systems begin to perform significant portions of AI R&D themselves, uncertainty about the feasibility of ASI would likely diminish, and AI development would become a central concern in national security planning.

In order to get to this point, it is not required for states to have deliberately entered a **Race** posture from the outset. As long as AI research is pursued by any actor, progress will continue to accumulate. Indeed, the overwhelming majority of advances in general-purpose AI to date have been driven by private companies, particularly in the US.

Once the possibility of ASI becomes credible, we expect superpowers to be well aware of the following hypotheses:[36]

- That the first actor to *close the loop* on Automated AI R&D may achieve an insurmountable advantage;
- That large advantages in AI development can be parlayed into utter military superiority.

Given these stakes, it is natural to expect strong pressure within superpowers to escalate their involvement in AI R&D. Once ASI is taken seriously as a potential determinant of future power, national security imperatives will push for bringing AI development under state control. Significant resources will be mobilized towards accelerating AI R&D to avoid falling behind rivals.[37] This creates a powerful force pulling states toward a **Race** posture.

A source of variation in this model is the degree to which superpowers deem AI development itself as posing a national security risk, regardless of who is in the lead. We identify two concerns pointing in this direction.

The first is the risk that powerful AI escapes human control. Refer to our previous paper[38] for a discussion on this risk.

The second is the risk that states do not manage to remain intact under the extreme concentration of power enabled by ASI. As ASI may enable extreme concentrations of power, even the victor of an ASI race might be unable to maintain its internal political order under this pressure.

Whoever is in direct control of an ASI system—whether by holding legitimate authorization or by having previously installed backdoors—would possess the capability to subvert any existing authority.[39] In democracies, any ASI-wielding entity would be effectively immune from systems of checks and balances, undermining the foundation of democratic governance.[40]

If decision-makers are sufficiently concerned about such risks, they will attempt to avoid any outcomes in which ASI is developed, regardless of who does it first. Such actors will favor a **Halt** posture.

[36] AI Deterrence Is Our Best Option, Dan Hendrycks, Adam Khoja: Second, many analysts in the US national security establishment are aware of the observations we have mentioned. We should assume that Chinese analysts are just as alert to the strategic importance of AI, and to the above observations, as US analysts are.

[37] The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race, Seán S. Ó hÉigeartaigh: While it does not appear the CCP or Chinese tech industry has responded substantially at time of writing, how long can that last? One possibility is that they become convinced of the capabilities and strategic decisiveness of artificial superintelligence, and indeed do invest far more of China's vast resources into trying to achieve it.

[38] The three main doctrines on the future of AI, Alex Amadori et al.

[39] The three main doctrines on the future of AI, Alex Amadori et al.: The extreme concentration of power enabled by ASI creates the possibility of snap "coups", as whoever is in control of an ASI system would be able to subvert any existing structure of political or military authority.

[40] AI and Catastrophic Risk, Yoshua Bengio: In the extreme, a few individuals controlling superhuman AIs would accrue a level of power never before seen in human history, a blatant contradiction with the very principle of democracy and a major threat to it.

Any state that halts unilaterally still faces the problem that other actors might continue development. Countries that consider the risk stemming from AI development from their own research programs as intolerable would surely see external efforts as an even more pressing threat.

One way or another, it appears unlikely that superpowers will remain neutral once ASI is regarded as a credible possibility. Regardless of which concern dominates—military competitiveness or loss-of-control risks—both pressures point toward urgent action and extreme stakes.

The prospect of complete loss of capability to ensure the integrity of one's security, either at the hands of an overwhelming ASI-wielding opponent or through the loss of control of powerful AI, will make prolonged inaction appear intolerable to decision makers.

With such stakes, many choices relevant to an ASI race will likely fall within the purview of executive and military institutions, rather than legislative bodies, which are more deliberative and more inclined to inaction. This further strengthens our prediction that superpowers will take assertive stances with regards to AI development.

5.3.1 Race to ASI

We first model the case in which both superpowers **Race**. We begin by exploring the possible end states, so that it is clear what actors are trying to achieve.

At some point in such a race, one actor may achieve a decisive strategic advantage. This condition would allow the leading state to launch a disabling strike against its opponent with little cost to itself, thereby "winning" automatically. Given this, gaining this advantage becomes the overriding priority for the AI development agendas of each side.

This would create the preconditions for two distinct possible outcomes, depending on whether the leading state retains control over its AI systems:

- **Takeover.** If control of the AI systems is maintained, the leading actor secures a decisive strategic advantage. In this position, it is confident in its capacity to disable its opponents' defenses at low cost and low risk, and to maintain control over its opponent indefinitely.

- **Extinction.** If control is lost, the AI systems themselves can be considered a hostile actor that possesses a decisive strategic advantage over all others. This likely results in catastrophic outcomes such as human extinction.[41]

In order to predict whether actors will adopt an **Attack** strategy, we must consider the following. Both actors are aware that the other is trying to achieve a decisive strategic advantage through AI development. If, at any point, it becomes clear to the trailing power that the leader is on the verge of achieving a decisive strategic advantage, it faces an urgent choice.

Since both the **Takeover** and **Extinction** outcomes represent total losses from its perspective, the trailing actor may view a preemptive attack on its rival as its only remaining path to survival.[42] In this case, the trailing actor would perform full-scale military intervention aimed at deterring or disrupting the opponent's AI program before it reaches irrecoverable superiority.

This would lead to a third outcome: **major power war** erupts before any one superpower has developed

[41] The three main doctrines on the future of AI, Alex Amadori et al.: Proponents consider it impossible to predict what will unfold in precise terms once a superintelligence has been developed and escaped human control. However, they tend to forecast that the outcome will not be compatible with human civilization and human life.

[42] AI Deterrence Is Our Best Option, Dan Hendrycks, Adam Khoja: If states are rational, and if cooperative measures are not implemented, escalation is unavoidable. Indeed, poor visibility and the possibility of sudden AI breakthroughs plausibly would drive states toward early or miscalculated escalation. For example, nations such as Russia with weak domestic AI programs might prefer outright preventive war if they lose hope of disrupting rivals' development or stealing their best models

sufficient capabilities to quickly and decisively end such a conflict in its favor.[43] This outcome has been regularly observed in the war simulation exercise "Intelligence Rising", created to examine AI race dynamics. [44]

Once a decisive strategic advantage appears within reach, it is unclear what level of aggression, if any, could deter continuation of AI research programs by the leading actor. This means that it's hard to put a bound on how far matters would escalate; hostilities could culminate in a full nuclear exchange.

At which point may the trailing superpower decide to perform a preemptive strike? While conflict is extremely costly, if an actor becomes convinced that the leader is bound to inevitably achieve a decisive strategic advantage, they would have a strong reason to attack, even if the decisive advantage is expected to materialize years in the future. We identify some factors that affect the timing of this decision.

The extent to which AI R&D has been automated. The more progress depends on AI systems rather than human researchers, the more future outcomes are locked in by current levels of capability. In the extreme case, where AI performs nearly all relevant R&D, the trailing power may conclude that it is doomed to lose if it waits. In this case, it will be inclined to attack as early as possible, rather than waiting for the advantage to grow.

Visibility into the opponent's AI program. High visibility would allow the trailing actor to judge with greater confidence whether the gap can still be bridged. In this case, it may attack earlier if the gap is deemed insurmountable. [45]

The degree of visibility will depend both on each actor's intelligence-gathering capabilities and on unpredictable characteristics of the technology itself. For instance, if large-scale automated industrialization becomes crucial to AI development, such activities would be difficult to conceal. Conversely, if AI R&D requires minimal physical infrastructure, concealment becomes easier.

It is useful to consider two illustrative examples at the boundaries. If AI R&D becomes heavily automated well before the decisive strategic advantage materializes, and if research progress is highly visible, the trailing actor is likely to attack early rather than wait for its position to worsen irreversibly.

In this case, outright war is highly probable, as the trailing actor would otherwise be knowingly and passively accepting its inevitable defeat in the future. On the other hand, the leading superpower is incentivized to wait as long as possible before engaging in conflict. Time is on its side: the longer the clock ticks, the more its advantage grows.

At the other extreme, if Automated AI R&D is less central to AI progress during the race, and if visibility into the rival's program is poor, any actor cannot easily assess to what degree it is leading or falling behind. Under this uncertainty, actors may hesitate to pay the high costs of war, and the race may even continue until a **Takeover** or **Extinction** outcome occurs.

[43]The AI Arms Race Isn't Inevitable (Palladium Magazine), Grace Werner: The relative peace between great powers since World War II has been sustained by strong disincentives that prevent hostilities from existentially threatening the global order. Central to this peace is the belief that an adversary is either incapable of obtaining strategic dominance or lacks sufficient incentives to do so. However, if an adversary were perceived as capable of overcoming these barriers or as having strong incentives to try, the absence of previously assumed costs could create a scenario where a preemptive strike against that adversary might be seen as justified. Consequently, the likelihood of such an action would significantly increase. If the promises of AI—a decisive strategic advantage—are taken seriously, the incentives for a first strike to undermine that potential victory may be viewed as necessary.

[44]Strategic Insights from Simulation Gaming of AI Race Dynamics, Ross Gruetzemacher, Shahar Avin, et al.: Races are destabilising: races often result in blocs split by national allegiance where the "loser" uses military actions to prevent the "winner" deploying radically transformative AI.

[45]However, lack of visibility may stoke concerns that the rival is being negligent with respect to loss-of-control risks, meaning that lower visibility would not necessarily reduce the overall probability of an attack.

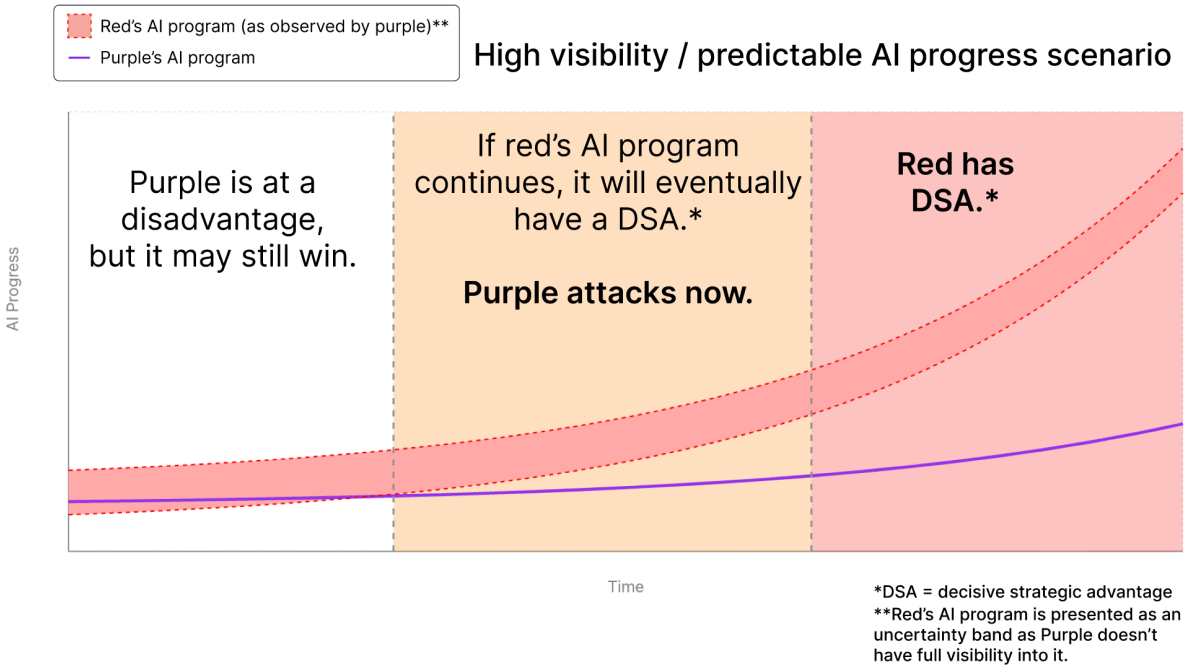


Figure 1: Representation of an ASI race between two superpowers, "Red" and "Purple" from the perspective of "Purple". Red's progress is shown as a confidence interval since Purple does not have complete visibility into Red's program. In this scenario, Red's advantage is highly predictable to Purple. This happens as a combination of two factors: (1) AI R&D becomes highly automated relatively early in the race, making the outcome of the race more predictable as it is overwhelmingly driven by a single factor; (2) AI R&D is mostly driven by activities that cannot be concealed, such as rapid expansion of physical infrastructure like energy plants and chip foundries. At some point, before Red has an overwhelming advantage, it becomes clear to Purple that Red will eventually develop DSA-granting AI if it is allowed to continue. At this point, Purple strikes preemptively at Red. At such a late stage, Purple must attack with enough force to disrupt or deter the AI program; this likely escalates into an all-out war between superpowers.

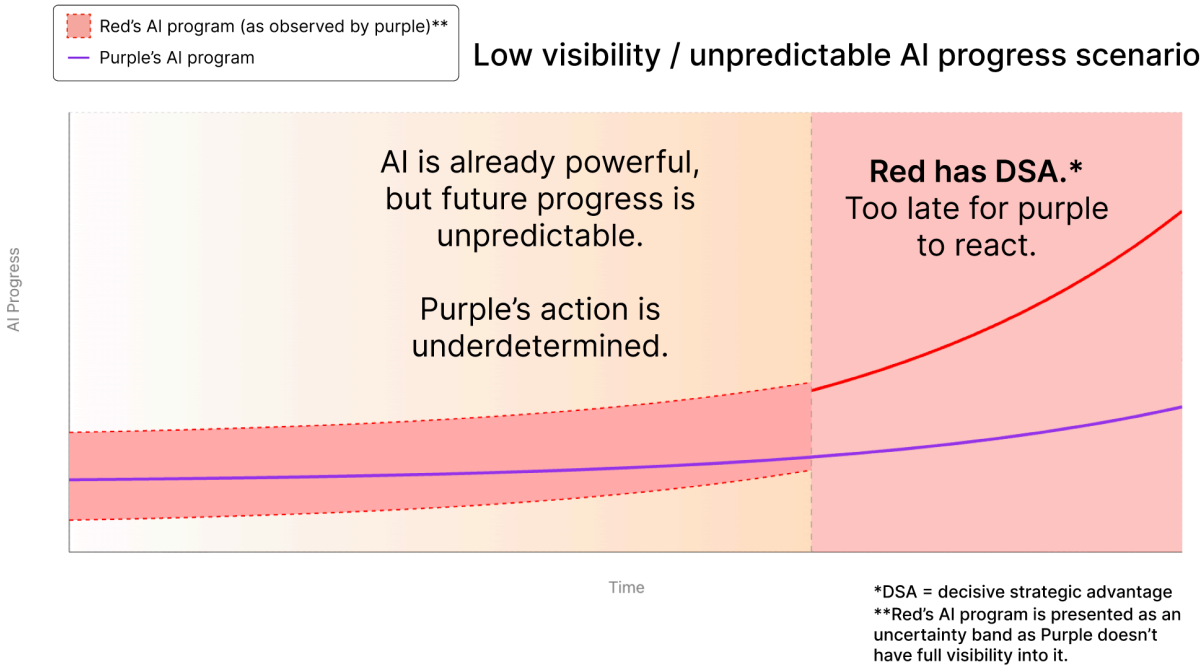


Figure 2: Representation of an ASI race between two superpowers, "Red" and "Purple" from the perspective of "Purple". Red's progress is shown as a confidence interval since Purple does not have complete visibility into Red's program. In this scenario, Red's advantage is unpredictable to Purple. Several factors could produce this uncertainty: (1) AI R&D may not become highly automated until after AI offers a decisive strategic advantage; (2) Automated AI R&D may otherwise fail to act as a single factor that overwhelmingly determines the pace of AI progress, for example because it is driven by domain-specific improvements instead of advancements in general-purpose AI; (3) AI R&D is mostly driven by activities that not visible to rivals, such as algorithmic breakthroughs. In this scenario, there is no clear point where Purple knows it is fated to lose and recognizes that it should attack immediately. On the other hand, Red might develop DSA-granting AI at any time, and neutralize Purple's defenses before it has time to react. In this case, it is hard to predict whether Purple would attack preemptively.

An important consequence of race dynamics is that, if AI development happens under intense competitive pressure, loss-of-control risks are likely to be further exacerbated. This is because competitive pressure encourages cutting corners in matters of safety.

Even under somewhat optimistic assumptions about the tractability of AI safety and control challenges, it would be costly to address them adequately. It would require much additional research, testing and in general precious time, which becomes a scarce commodity in a high-stakes race for strategic advantage. [46] [47] [48]

In addition, incentives encourage delegating more and more of the decision making relevant to military operations to AI systems. This is because at some point in the race, AI systems should outperform human judgment across most domains.

As a lower bound, we should expect that AI will think faster than human strategists, allowing it to react more quickly, take into account more information and spend much more "thinking time" considering each individual decision. [49,50] This means that any actor would be sacrificing a significant competitive advantage by not delegating decisions to AI systems.

5.3.2 Halting superpowers

We now consider the possibility that one or both superpowers in the model may adopt a **Halt** posture.

One Races, the other Halts. In this case, the halting state perceives intolerable risks of loss-of-control outcomes stemming from the rival's AI program. From its perspective, allowing the other actor to continue unchecked would mean facing either total domination by the rival or extinction if control over AI systems were lost.

As a result, halting is unlikely to be accompanied by a passive outward stance. Instead, the halting actor would treat stopping its rival as an absolute priority. It might begin with sabotage or threats to attack the racing actor; [51] if these measures fail as deterrence, the next step would likely be **open conflict between superpowers**.

Both Halt. Both actors may halt, either due to strong concerns about catastrophic outcomes or because one is deterred from racing by the other's threats. Without thorough verification mechanisms, this "pause" would be quite fragile.

Each state would doubt whether the other has truly ceased all relevant research. In the simplest case, each actor may be outwardly declaring a public **Halt** posture while covertly continuing research. In addition, actors are incentivized to "tickle the dragon's tail", performing AI research at what they consider the very boundary between safe and negligent. Disagreements over where this boundary lies may stoke tensions.[52,53]

[46] AI Safety, Ethics, and Society, Dan Hendrycks: We can imagine a future in which similar pressures lead companies to cut corners and release unsafe AI systems.

[47] OpenAI charter: We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions

[48] The Most Dangerous Fiction: The Rhetoric and Reality of the AI Race, Seán S. Ó hÉigeartaigh: By framing AI as a technology too powerful to allow rivals to possess, a race that may compromise safety and ethical considerations is incentivized. The pressure to be first may lead to cutting corners, ignoring potential risks, and prioritizing speed over security, potentially jeopardizing the promised benefit of the technology to humanity.

[49] AI Commission Recommends Billions in New Spending, Robert Work (as quoted by National Defense Magazine): AI-enabled applications will operate at machine speeds and humans simply will not be able to keep up with them without help from their own algorithms and their own AI.

[50] In addition, AI may have other advantages that are harder to quantify, such as being able to generate better candidate strategies and having superior intuitions.

[51] Superintelligence Strategy, Dan Hendrycks et al.: If a rival state races toward a strategic monopoly, states will not sit by quietly. If the rival state loses control, survival is threatened; ···states will act to disable threatening AI projects.

[52] Strategic Insights from Simulation Gaming of AI Race Dynamics, Ross Gruetzemacher, Shahar Avin, et al.: Even when agreements are in place to slow progress and focus on safety ensuring trust is difficult and deception is common

[53] As an additional complication, if trust between superpowers has sufficiently eroded, decision-makers may adopt a defensive stance toward warnings about loss-of-control risks. Warning about such risks may be treated as deliberate

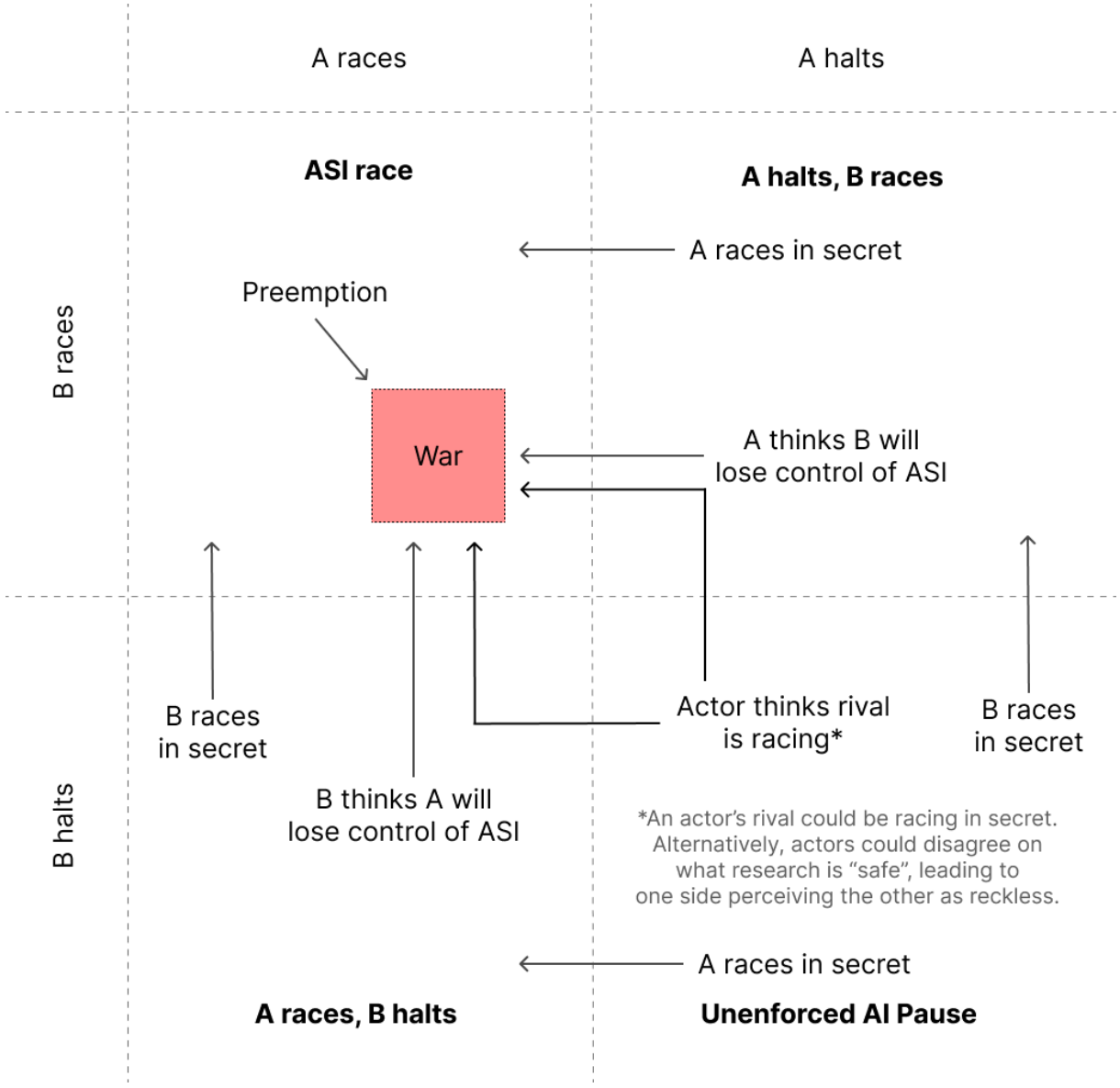


Figure 3: Matrix of the strategies of two superpowers deciding whether to Race or Halt. When both actors are racing, the situation may easily escalate to a war between the superpowers as the winner-take-all dynamics of AI incentivize preemptive attacks aimed at disrupting the opponent's AI program. When only one actor is halting, as it is concerned about risks from losing control of powerful AI, it presumably does not trust its rival to develop powerful AI either. The halting actor will thus try to coerce the racing actor into halting, and if that fails, attack with enough force to disrupt its AI R&D program. When both actors are halting, they must urgently establish a formal agreement capable of conferring mutual trust. An unenforced AI pause is not stable, as either actor may find ways to keep pursuing dangerous AI in secret. Additionally, without a formal consensus on when AI research is considered safe, each actor may perform activities that it deems acceptable, but that are considered dangerous by its rival.

In practice, it seems highly unlikely that this equilibrium would be stable within a non-cooperative model. The trajectory of this scenario would depend heavily on the existence and design of robust international agreements. This topic is further elaborated upon in section 8.

6. Non-cooperative model of middle power strategies in a ”fast progress” world

We now examine middle power strategies within our non-cooperative model of the ”fast progress” scenario. We argue that no available strategy meaningfully alters the dynamics or conclusions of our superpower model, thereby justifying our decision to model superpowers in isolation.

Middle powers would face a dire predicament in this scenario. An ASI race presents severe risks to their sovereignty and survival. At the same time, individual middle powers would have extremely limited capacity to influence outcomes, meaning they would likely remain passive observers.

Vassal's wager:

Middle powers may attempt to ally with a superpower, or at least avoid antagonizing superpowers, in the hope that their autonomy will be respected in the future.

Challenging superpowers:

Middle powers may compete against superpowers in the AI race, or attempt to pressure superpowers to halt or slow down their AI programs.

6.1 ”Hail mary” attempt in the ASI race

A middle power could decide to enter the AI race, competing against superpowers. A middle power taking this strategy would face slim chances of success. From a technological perspective, middle powers lack the resources, economic might, and access to expertise necessary to match the pace of superpower AI development programs.

Additionally, for a middle power to truly have a chance of winning an ASI race, it would need the capability to withstand or deter a preemptive attack from a superpower. Compared to superpowers, middle powers would struggle to put up credible deterrence, making them comparatively vulnerable to coercion or disruption.

6.2 Pressuring the racing superpowers

A middle power might determine that the risks posed by superpower AI programs have grown beyond limits they consider acceptable. In this case, they could unilaterally impose various measures intended

attempts by adversaries to sow fear, uncertainty, and doubt in order to slow their progress.

to pressure superpowers to halt or slow down their AI programs. Depending on their assessment of the situation's severity, the measures could range from economic sanctions to more extreme actions including armed attacks.

When it comes to most middle powers, it seems unlikely that this strategy would be effective if pursued by a single actor. Middle powers generally lack the economic and military leverage to dissuade superpowers from racing. We acknowledge some possible exceptions, such as middle powers with significant nuclear capabilities and middle powers that are critical in the semiconductor supply chain. We expand on these in the annex.

While multiple middle powers might share similar concerns about AI development, any individual party cannot rely on others joining their efforts and must therefore consider the worst-case scenario of acting alone. If this strategy is taken by a single actor or an insufficiently robust coalition, it runs the risk of targeted retaliation by the racing actors. This might result in heavy costs while having minimal impact on the race.

Even so, it cannot be entirely ruled out that some might attempt this strategy. Depending on their calculations, certain middle powers may conclude that they prefer confronting superpowers in the present rather than facing the consequences of one winning an ASI race. Motivations for such a choice could include:

- Low trust that the winning superpower would respect their sovereignty;
- Concerns about loss-of-control risks from AI development;
- Expectations that the race will lead to devastating major power war, such that taking action immediately would likely result in less disruption than waiting.

6.3 Vassal's Wager: allying with a superpower

A middle power can choose to ally itself with a superpower, either forming a new alliance, or maintaining and strengthening an existing one. For this strategy to succeed, several conditions must align favorably:

- The superpower must successfully navigate loss-of-control risks. The middle power's views on the on the adequacy of the techniques used to mitigate these risks would likely carry limited influence.
- The superpower must emerge victorious in the AI race, ideally in a way which avoids major power war; if war breaks out, the middle power might be dragged into the conflict either through formal obligations to assist their patron or by becoming targets of opposing superpowers[54].
- Even if the superpower emerges victorious, there is no guarantee that the middle power's sovereignty and independence will be respected by the winner. The middle power would have essentially no recourse against hostile behavior by an ASI-wielding superpower.

The middle power exerts little control over these conditions. This strategy amounts to what we call a "Vassal's Wager", in which the "vassal" relinquishes meaningful agency over its own future and must rely entirely on the competence and trustworthiness of its "patron".

6.4 Neutrality

Conversely to the previous strategies, a middle power can avoid taking a strong position with respect to an AI race, neither forming close alliances with any superpowers, nor alienating them. While this

[54]The Risks of Preventive Attack in the Race for Advanced Artificial Intelligence, Zachary Burdette, Hiwot Demelash at RAND: State A is on the cusp of developing AGI and state B attacks a third party, state C, to achieve its political goals during a closing window of opportunity before state A becomes more powerful. This would occur most likely in cases in which states A and C are allies or partners.

approach offers certain advantages, such as avoiding commitments to participate in potential conflicts, it also carries significant risks, like exclusion from protective alliances.

In practice, we consider this last strategy similar to "Vassal's Wager" in that it ultimately relies on factors beyond the middle power's control: that the winning superpower does not cause catastrophic outcomes by losing control of their AI systems, that the winning superpower treats the middle power fairly and respects its sovereignty, and that the middle power is not devastated as collateral damage in major power war.

7. Considerations on the "plateau" scenario

As we argued in section 4, the "plateau" scenario is intractable to predict in as much detail the "fast progress" scenario. As a result, we limit ourselves to examining some factors that may shape outcomes.

Compared to the "fast progress" scenario, the "plateau" scenario sees a significantly reduced risk of catastrophic outcomes as a result of loss-of-control of AI systems, as well as a reduced probability that a single actor achieves permanent global dominance.

As explored in our previous work, it remains unpredictable whether this scenario would be beneficial.[55] On one hand, experts expect anywhere from modest to extraordinary acceleration of scientific and technological as well as economic growth.

Conversely, the negative outcomes include mass unemployment, extreme concentration of power as states become less reliant on their populations for labor[56], and the risk of large-scale manipulation by persuasive AI systems.[57,58]

Democratic societies may be particularly vulnerable to some of these risks, as they could undermine mechanisms that are foundational to their functioning such as checks and balances and public trust in information.[59,60]

We note two factors that, even in a world with weaker AI, might nonetheless contribute to precipitating major power war.

Geopolitically destabilizing applications of weaker AI. Even AI systems incapable of feats like Automated AI R&D or R&D of novel weapons—which would take us squarely to the "strong AI" branch of the model—may still allow for the development of new disruptive military capabilities. Such AI systems would be able to provide capacity for routine tasks at scales impossible to achieve with human personnel, and with much faster reaction times.

For example, this may enable the operation of missile defense systems or large drone swarms vastly superior to those that exist today. They may also vastly raise states' capacity for cyberwarfare. [61]

[55]The three main doctrines on the future of AI, Alex Amadori et al.: Proponents of this doctrine hold varying views on whether AI's overall impact will be beneficial, with both extraordinary benefits and major disruptions being highlighted as possibilities.

[56]The Intelligence Curse, Luke Drago, Rudolf Laine

[57]Artificial Influence: An Analysis Of AI-Driven Persuasion, Matthew Burtell, Thomas Woodside: We warn that ubiquitous highlypersuasive AI systems could alter our information environment so significantly so as to contribute to a loss of human control of our own future.

[58]Keep the future Human, Anthony Aguirre: Finally, a significant threat of in-gate AI is its use in personalized persuasion, attention capture, and manipulation.

[59]Keep the Future Human, Anthony Aguirre: They would likely lead to the concentration of vast economic, social, and political power –potentially more than that of nation states –into a small number of massive private interests unaccountable to the public. ···By undermining human discourse, debate, and election systems, they could reduce the credibility of democratic institutions to the point where they are effectively (or explicitly) replaced by others, ending democracy in states where it currently exists.

[60]AI and Catastrophic Risk, Yoshua Bengio: In the extreme, a few individuals controlling superhuman AIs would accrue a level of power never before seen in human history, a blatant contradiction with the very principle of democracy and a major threat to it.

[61]The three main doctrines on the future of AI, Alex Amadori et al.: Advanced cyberwarfare, potentially capable of com-

Given the possibility that AI may enable powerful new military applications, superpowers would likely still invest significant state resources into AI R&D. However, the extent of state control, secrecy and resource allocation is likely to be lower than in the "fast progress" scenario.

Uncertainty about the pace of progress. Even in scenarios where ASI is technically infeasible—or at least not until much later—this will not necessarily be evident to relevant actors. The trajectory of AI developments is difficult to predict or even to accurately assess as it is happening. Actors may still perceive a credible risk that their rival's AI program has the potential to achieve a destabilizing advantage, and tensions may escalate accordingly.

8. Conclusion

Our modeling suggests that the trajectory of AI development may come to overshadow other determinants of geopolitical outcomes, creating momentum toward highly undesirable futures. Superpowers face overwhelming incentives to push the frontier of AI for military applications and economic dominance. This dynamic risks deteriorating relations between them, making cooperation harder at exactly the time when it becomes most necessary.

If AI systems become capable of autonomously performing AI R&D work, or capable of designing or operating destabilizing novel weapons, the race will accelerate sharply, and the stakes will become immediate and unmistakable to decision makers. These dynamics further increase the risks of catastrophic failures, including both loss of control of powerful AI and major power war.

We highlight the urgent need for cooperation, among all relevant actors, aimed at restricting dangerous forms of AI development. Work is urgently needed to identify how such cooperation could be achieved in practice.

In order to effectively curtail these risks, two conditions would need to be achieved at minimum. First, the international community would need mechanisms to prevent any actor from unilaterally advancing AI development, allowing further progress only through approaches which benefit from strong scientific consensus about their safety, and possess broad multilateral support. Second, comprehensive verification systems would be necessary to ensure that no actor secretly pushes the frontier of dangerous AI capabilities.

Middle powers have far more to lose and little to gain from participating in an AI race between superpowers or from passively standing by as one unfolds. In a world where trust between superpowers may deteriorate in lockstep with our ability to predict whether ASI is feasible or how quickly it may arrive, middle powers are well incentivized to kickstart international coordination efforts.

In the past, international coordination—especially on arms control—has been driven by superpowers. Future research should examine how middle powers might drive international coordination on AI and how they might be able to pressure superpowers into participating in such a regime of cautious AI development.

Avoiding the most dangerous trajectories requires building mechanisms that enable trust of mutual restraint as early as possible—before countries risk being locked in a competitive race for a decisive strategic advantage, with the consequences of losing so intolerable that extraordinary measures become the only viable means of deterrence.

---

pletely disabling retaliatory capabilities ···Autonomous weapon systems, such as tightly coordinated, massive autonomous drone swarms.

Annex: Middle powers with distinctive capabilities to influence the AI race

In section 6.2, we contend that middle powers are not able to meaningfully dissuade superpowers from participating in an AI race. In this annex we acknowledge some ways in which specific middle powers may have outsized influence through unilateral actions.

While we think that these methods are unlikely to be effective and even more unlikely to ensure lasting stability, we address them and assess to what degree they might work.

Actors with critical roles in the semiconductor supply chain

Some actors, like Taiwan, the Netherlands, and South Korea, are critical nodes in the semiconductor supply chain. If these actors were to restrict exports to actors pursuing AI development they deem reckless, they could delay such AI programs by several years.

While Taiwan could inflict the greatest disruption by undertaking such a policy, this is unlikely to be viable from their point of view. The semiconductor industry constitutes an indispensable part of Taiwan's economy. Additionally, it has been suggested that their role in the semiconductor supply chain is the primary reason why the US is deterring a Chinese invasion.[62] As a result, Taiwan may consider restricting chip exports to the US as suicidal.

The Netherlands may be best positioned to execute this strategy. The company ASML, based in the Netherlands, is the sole producer of EUV lithography machines, which are essential in modern semiconductor manufacturing [63]. At the same time, the Netherlands does not rely on this industry to the same extent as Taiwan and South Korea for its export revenue, and it is not an essential element of their national security strategy.

While this strategy would temporarily cripple the expansion of superpower AI programs, it seems likely that superpowers would catch up in domestic production capacity within a few years. Indeed, both China and the US are already pursuing semiconductor self-sufficiency.[64,65] While this would not represent a lasting solution, it could buy valuable time.

Nuclear-armed middle powers

There are several middle powers that lack the capacity to compete with superpowers in an AI race or in conventional warfare, but possess significant nuclear capabilities. If such an actor was sufficiently concerned about risks originating from AI development, it may be able to meaningfully pressure superpowers by threatening a nuclear strike if certain AI development redlines are crossed.

Several issues make it unlikely that this strategy would lead to a stable world in which dangerous AI development is successfully and durably deterred. The first consideration is that any redlines would need to rely only on information that is visible to the actor making the ultimatum.

The redlines would need to be drawn at such extreme thresholds that the middle power would rather pay the cost of being utterly destroyed in a nuclear exchange than accept the risk of AI development past that redline. At the same time they would need to be conservative enough that, if they are respected, it is not possible for any superpowers to leverage AI to develop capabilities that invalidate the nuclear deterrent.[66]

[62] Defending Taiwan With Chips and Drones, Harry Goldstein at IEEE Spectrum

[63] ASML is the only company making the $200 million machines needed to print every advanced microchip. Here's an inside look, Katie Tarasov at CNBC

[64] China sets up third fund with $47.5 bln to boost semiconductor sector, Reuters

[65] Chips for America, The U.S. Department of Commerce

[66] For example, defensive capabilities capable of deflecting a nuclear attack, or cyberwarfare capabilities sufficient to disable the middle power's capacity to attack.

Excluding conditions in which AI progress is maximally visible and predictable (see section 5.3.1 for discussion on this), it seems that the middle power would be unable to define crisp redlines, and would instead need to adopt ambiguous and flexible criteria. This would place the middle power—as well as the rest of the world—into a position that is significantly more precarious than the one provided by the mutual assured destruction (MAD) framework.

Under MAD, it is exceedingly clear which actions would lead to mutual destruction: deploying nuclear weapons. This is a discrete and clearly detectable action, and it is clear to all parties that crossing this line would result in retaliation. There is no way for an actor to test their adversaries' boundaries while trying to stay below the threshold for retaliation.

In the case of AI development, any redline that can be set in a non-cooperative setting would necessarily need to be ambiguous, and any doctrine prescribing a nuclear response to AI development would need to rely on a substantial discretionary component.

In this arrangement, there is no clear limit on what opponents may do without inviting a nuclear attack; opponents will therefore be tempted to push their luck. Racing actors would be incentivized to slow down or obfuscate their AI progress in an attempt to make it feel gradual enough that the middle power never perceives a clear violation to their terms and so never pulls the trigger. In other words, they will be tempted to try and "boil the frog". [67]

Compared to nuclear-driven MAD, this makes two types of failure modes much more likely:

- The failure mode in which racing actors overreach, resulting in a nuclear exchange;
- The failure mode in which ASI is developed and its negative consequences are realized.

[67] Boiling frog, Wikipedia