How middle powers may prevent the development of artificial superintelligence

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

Executive summary

Artificial intelligence (AI) development, particularly the pursuit of artificial superintelligence (ASI), presents unprecedented challenges to international security. Expert analyses suggest that the first actor to develop ASI could achieve a decisive strategic advantage—the ability to neutralize all rivals' defenses at minimal cost and maintain indefinite control over them.

Alternatively, humanity may lose control of AI systems entirely, resulting in its extinction or permanent disempowerment. These outcomes could materialize through AI R&D activities alone, without any actor —whether a country or a renegade AI system—committing a conventional act of war until it has absolute confidence of success.

In previous work we investigated the strategies available to countries in the absence of international coordination focused on preventing dangerous AI development. We found that, in the context of a race to ASI, middle powers have little to gain and much to lose.

In this context, middle powers face catastrophic risks to their national security, with the worst case entailing the country's complete destruction. At the same time, they lack the means to unilaterally influence superpowers to halt their attempts to develop ASI.

Middle powers may ally with a superpower, a strategy we term ”Vassal's Wager”. Even provided that their patron ”wins” the race and averts catastrophic outcomes, this strategy offers no guarantee that the middle power's sovereignty will be respected once the superpower achieves an overwhelming strategic advantage over all the actors.

We argue that it is in the interest of most middle powers to collectively deter and prevent the development of ASI by any actor, including superpowers. In this work, we lay out the design of an international agreement which would enable middle powers to achieve this goal.

This agreement focuses on establishing a coalition of middle powers that is able to collectively pressure other actors to join a regime of restricted AI R&D backed by extensive verification measures. Eventually, this coalition should be able to deter even uncooperative superpowers from persisting in pursuing AI R&D outside of this verification framework.

Key Mechanisms

Trade restrictions. The agreement imposes comprehensive export controls on AI-relevant hardware and software, and import restrictions on AI services from non-members, with precedents ranging from the Chemical Weapons Convention and the Nuclear Non-Proliferation Treaty.

Reactive deterrence. Escalating penalties—from strengthened export controls to targeted sanctions, broad embargoes, and ultimately full economic isolation—are triggered as actors pursue more and more dangerous AI R&D outside of the verification framework.

Preemptive self-defense rights. The coalition recognizes that egregiously dangerous AI R&D constitutes an imminent threat tantamount to an armed attack, permitting members to claim self-defense rights in extreme cases.

Escalation in unison. The agreement would establish AI R&D redlines as well as countermeasures tied to each breach. These are meant to ensure that deterrence measures are triggered in a predictable manner, in unison by all participants of the agreement. This makes it clear to actors outside of the agreement which thresholds are not to be crossed, while ensuring that any retaliation by actors receiving penalties are distributed among all members of the coalition.

Though these measures represent significant departures from established customs, they are justified by AI's unique characteristics. Unlike nuclear weapons, which permit a stable equilibrium through mutually assured destruction (MAD), AI R&D may lead to winner-take-all outcomes. Any actor who automates all the key bottlenecks in Automated AI R&D secures an unassailable advantage in AI capabilities: its lead over other actors can only grow over time, eventually culminating in a decisive strategic advantage.

Path to Adoption

We recommend that the agreement activates once signatories represent at least 20% of the world's GDP and at least 20% of the world's population. This threshold is high enough to exert meaningful pressure on superpowers; at the same time, it is reachable without assuming that any superpower champions the initiative in its early stages.

This threshold enables middle powers to build common knowledge of their willingness to participate in the arrangement without immediately antagonizing actors in violation of the redlines, and without paying outsized costs at a stage when the coalition commands insufficient leverage.

As the coalition grows, network effects may accelerate adoption. Trade restrictions make membership increasingly attractive while non-membership becomes increasingly costly.

Eventually, the equilibrium between competing superpowers may flip from racing to cooperation: each superpower could severely undermine the others by joining the coalition, leaving the final holdouts facing utter economic and strategic isolation from the rest of the world. If this is achieved early enough, all other relevant actors are likely to follow suit and join the verification framework.

Urgency

The agreement's effectiveness depends critically on timing. Earlier adoption may be achieved through diplomatic and economic pressure alone. As AI R&D is automated, superpowers may grow confident they can achieve decisive strategic advantage through it. If so, more extreme measures will likely become necessary.

Once superpowers believe ASI is within reach and are willing to absorb staggering temporary costs in exchange for total victory, even comprehensive economic isolation may prove insufficient and more extreme measures may be necessary to dissuade them.

The stakes—encompassing potential human extinction, permanent global dominance by a single actor, or devastating major power war—justify treating this challenge with urgency historically reserved for nuclear proliferation. We must recognize that AI R&D may demand even more comprehensive international coordination than humanity has previously achieved.