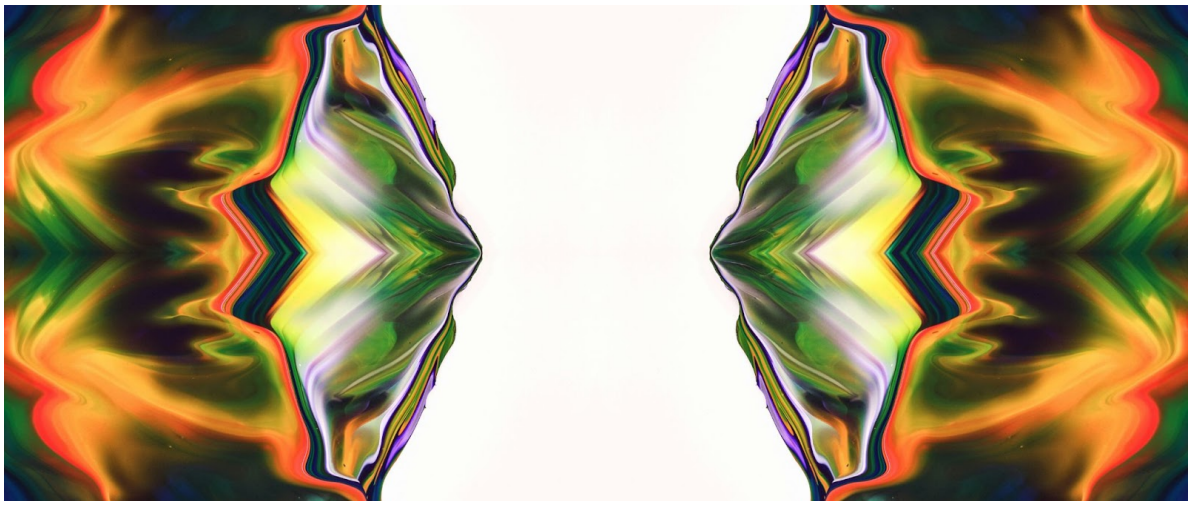
All › Peace & Security

AI Deterrence Is Our Best Option

A response to critiques of Mutually Assured AI Malfunction (MAIM).

Sep 22, 2025

Dan Hendrycks and **Adam Khoja**



“AI Deterrence Is Our Best Option” by Dan Hendrycks, Adam Khoja
Sep 21 • AI Frontiers

21:19

implications of states racing to develop artificial superintelligence (ASI) — AI systems that vastly exceed human capabilities across nearly all cognitive tasks. The paper argued that no superpower would remain passive while a rival transformed an AI lead into an insurmountable geopolitical advantage. Instead, capable nations would likely threaten to preemptively sabotage any AI projects they perceived as imminent threats to their survival. But with the right set of stabilizing measures, this impulse toward sabotage could be redirected into a deterrence framework called Mutual Assured AI Malfunction (MAIM).

Since its publication, "Superintelligence Strategy" has sparked extended debate. This essay will respond to several critiques of MAIM, while also providing context to readers who are new to the discussion. First, we'll argue that creating ASI incentivizes state conflict and the tremendous tensions that its development produces are not confined to MAIM. Second, we'll consider whether MAIM's proposals reduce instability. Third, we'll explore the issue of redlines, and determine whether MAIM can effectively shape states' perceptions of risk.

Building Superintelligence Amplifies Tensions, and Could Be Considered an Act of War

In this section, we'll walk through the arguments of "Superintelligence Strategy", discussing some of its core observations while responding to several objections. We'll begin with the main premise: states have a strong incentive to threaten sabotage against rival development of ASI.

A state with "superweapons" may become overwhelmingly powerful. If one state attains sole access to superhuman AI scientists and engineers, it may

undermine nuclear deterrence altogether, allowing it to obliterate its rivals and seize total geopolitical power. More important than any specific superweapon (such as advanced anti-ballistic missile technology, next-generation drones, and comprehensively destructive cyberweapons) is the advantage that a state with vastly superior AI capabilities would have in developing and fielding new breakthroughs more quickly than others. Unlike many traditional military capability asymmetries, which are costly but bearable, a state with sole access to ASI would likely have the power to completely dominate its rivals.

States cannot trust that rivals won't use ASI against them. Russia, for example, would not take solace in American assurances that it would use an ASI only for peaceful purposes. We should thus expect powerful nations to view a rival with unilateral access to ASI as an imminent threat to their continued existence.

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.

The majority of frontier AI companies are in fact planning for a recursion. Anthropic CEO Dario Amodei wrote: "[B]ecause AI systems can eventually help make even smarter AI systems, a temporary lead could be parlayed into a durable advantage." Or consider Sam Altman's discussion of AI-led research:

security on the hope that frontier AI companies are wrong and that fully automated AI research won't amount to much.

A recursion is also a plausible path to losing control of ASI. ASI would not be a mere tool, and it may not remain under the control of the actor who develops it. The machine learning pioneer and Nobel Prize–winner Geoffrey Hinton remarked that "there is not a good track record of less intelligent things controlling things of greater intelligence." During an intelligence recursion — when AI systems would be given broad license to design their successors and human oversight would be minimal — there is a heightened risk that no state will be able to control what results. If a state undertakes a recursion, even allies would justifiably be gravely concerned, because they would also be threatened by a loss of control.

National security establishments will be aware of the stakes. Wildeford and Delaney, alongside David Abecassis, also argue that powerful states might not have enough strategic foresight to consider aggressive measures in response to a racing rival.

We disagree. First, people in the intelligence community do not need much excuse to spy on rival AI development. Likewise, cyber command does not need much excuse to develop attacks against AI data centers. Performing espionage and preparing disruptions against their geopolitical rivals' tech stack is their mandate even in low-stakes situations. 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.

Pursuing ASI dominance can be considered an act of war. RAND senior political scientist Iskander Rehman and colleagues argue that threatening to sabotage ASI development is "unilateral and coercive." This mistakes self-defense for aggression. *Pursuing total dominance through ASI* is unilateral and

justifiable response when a rival is bidding for complete power, which could circumvent mutual assured destruction's (MAD) power to deter omnicide.

Norms will adapt to strategic realities. Critics correctly point out that building ASI is not yet widely viewed as a grave escalation. But geopolitical norms adapt to strategic realities; imagine how absurd MAD would have sounded to many from the year 1930. As national security experts grasp the alarming implications of ASI, norms should shift toward greater vigilance about ASI's creation, and more accepting of efforts to disrupt it in the name of self-preservation. At that point, it will not be difficult for states to diplomatically and domestically justify their efforts to disrupt another state's unilateral efforts to attain what Anthropic's Amodei describes as an exclusive "country of geniuses" — in this case, military geniuses — "in a datacenter."

Modern AI development is currently vulnerable to state sabotage. States' strong motivations to disrupt a rival building ASI can be actualized, because states are indeed capable of sabotaging frontier AI development. Frontier development requires billions of dollars' worth of hardware, in datacenters so large and heat-emitting that they cannot easily be placed underground to escape conventional kinetic attacks. Security at most labs is remarkably lax, leaving their information robustly accessible to well-resourced nations. The supply chains for the best AI hardware are so fragile that powerful states could block them for months or years if they desired. Most importantly, defending against all of these vulnerabilities in the near term would require significant time and effort, which would trade off heavily with competitiveness.

This concludes a discussion of the core observations that underlie MAIM. Some

paper's proposals. We'll take a moment to address these conflations.

Complaints about MAIM are often complaints about ASI races in general. Rehman et al. highlight the escalation risks that could stem from premature maiming attacks, based on poor intelligence of a rival's ASI progress. And Jason Ross Arnold points to the destabilizing implications of poor observability of rival AI development and the potential for sudden shifts in AI capability to catch states off guard. We agree that each of these issues raises the odds of conflict, but we disagree that they apply to MAIM uniquely rather than inherently to the development of advanced AI.

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. As we will discuss in the next section, MAIM proposes several mechanisms to reduce the chance of preventive war and other default destabilizing dynamics that ASI development creates.

MAIM's Proposals Increase Stability

The feasibility of AI disruption poses both risks and opportunities. We have already established that ASI projects would invite conflict, a risk that is not confined to MAIM. The opportunity is that states can channel these incentives for conflict towards a more stable deterrence dynamic, namely MAIM. We argue that MAIM actually reduces the volatility that ASI development introduces as well as the risk of unjustified escalation.

MAIM recommends escalation ladders, verification, and more. MAIM advocates for escalation ladders of sabotage; formalizing ties between AI

sabotage in ways that could be seen as a drastic escalation; and implementing multilateral transparency and verification measures. These efforts attempt to convert the default regime of unstructured preemption into a more stable deterrence dynamic. More minimal forms of MAIM do not require new treaties, but more stable forms include verification and enforcement.

Comparing nuclear and AI deterrence. Nuclear MAD is perhaps the most famous example of deterrence. "Superintelligence Strategy" draws a pedagogical parallel between MAIM and MAD, but the argument for MAIM is based on AI's unique features.

The structures of the two frameworks differ significantly. For example, MAD is built around nuclear *retaliation*, whereas MAIM represents AI *preemption*, analogous to the ongoing US preemption of Iran's nuclear-development efforts. MAD and MAIM also have different constraints and standards of success: MAD has been successful in helping to prevent nuclear war throughout this century and much of the last, while MAIM hopes to forestall destabilizing ASI development and the conflict it would engender for some years until states can transition to a more verifiably stable deterrence regime.

Some critics of "Superintelligence Strategy" argue against MAIM by claiming that it does not satisfy the same assumptions as MAD. We have generally found criticisms of MAIM based on comparison to MAD assumptions (rationality, secure second-strike, and overwhelming arsenals) to be unhelpful. MAD itself has succeeded despite having its assumptions strained in a number of ways. That said, Wildeford and Delaney argue that MAIM would fail to work as MAD has, in areas such as reliability of retaliation and attack attribution. Objections like these improperly judge MAIM by the standards of retaliatory deterrence. MAIM doesn't depend on retaliation. ASI projects are disrupted whenever a rival perceives them as imminent threats, *not* in response to an attack. MAIM also doesn't depend on attribution — the risk of ASI development being sabotaged can act as a deterrent regardless of whether the

we will discuss why MAIM increases stability, regardless of whether states act unilaterally or cooperatively.

To be considered useful, MAIM does not need to prevent ASI development. Rehman et al. write: "Even a credible MAIM threat might not deter a rival from pursuing superintelligent AI. Halting one's AI development would entail essentially the same costs as being the victim of a MAIM attack — loss of the program." And Abecassis similarly argues that a maiming attack would not impose sufficient costs to deter a dedicated rival from racing for ASI.

However, MAIM's proposals can still be useful even if a state chooses to race toward ASI in the face of strong deterrents. For the sake of argument, suppose that a rival is pursuing ASI development at all costs. MAIM suggests that other states apply escalating interventions to communicate rising discomfort with the rival's activity and demonstrate their willingness to escalate further. MAIM suggests tying threats of sabotage to especially risky development pathways, such as an intelligence recursion — this would shape the rival's risk assessment of those techniques and encourage them to pursue less risky methods.

MAIM suggests that states should signal the situations where they would threaten more severe sabotage. The aggressive rival would need to undertake

Finally, if a rival chooses to harden its development — shrugging off demands for transparency and negotiations, despite intermediate escalations — states will have ample confirmation of the rival's belligerent intentions and could then justifiably consider sober next steps.

In this way, MAIM's recommendation to clarify the escalation ladder reduces information problems, slows bids for dominance, and increases stability. In practice, we do not believe a state would pursue a bid for dominance through ASI at all costs if it anticipated the robust response we have described. That is to say, MAIM deters destabilizing ASI projects.

Deterrence is an organic pathway toward cooperation. States would likely prefer to have mutual visibility and leverage between their own and rivals' ASI projects over an all-out race that each side frames as a national security emergency, which would bring tremendous risks of escalation. MAIM's tools — escalation ladders, transparency, and verification — are the scaffolding for legitimately enforceable agreements. Wary rivals would be unlikely to accept proposals for voluntary halts on development, for example, without a robust mechanism to detect and deny violations of the agreement. And states that are reluctant to cooperate may still be pressured through deterrence into stabilizing win-win arrangements.

"Superintelligence Strategy" lists several verification measures that states could pursue, including location verification and AI-assisted datacenter inspections, and work since the publication of that piece has painted an early yet promising picture of how they could be implemented. These measures could eventually supplement unilateral espionage and sabotage techniques as the basis of robust deterrence. But states are not going to rush to implement verification mechanisms without a strong cause; upholding a deterrence regime is a realistic organic pathway to aligning state interests toward transparency.

enforceable international agreements.

MAIM Facilitates Redlines

Several critics of "Superintelligence Strategy" have argued that MAIM does not have *redlines* — in this case, clear thresholds for AI development that would provoke sabotage — and so it cannot function. We disagree on two fronts.

First: MAIM does specify useful redlines. "Superintelligence Strategy" presents an intelligence recursion as a specific activity worth deterring. If a huge fleet of AIs in a rival state are fully autonomously engineering the next generation of AI, disruption would clearly be warranted. Second, and more importantly: precise, prespecified redlines are not necessarily essential to strategic deterrence. Specifying precise redlines is one of many ways that actors can communicate to shape rivals' perceptions of risk. In this section, we discuss redlines and related objections.

Transparency and verification can aid increasingly precise deterrence. Abecassis argues that intelligence recursion could happen so fast (on a scale of days to weeks) that rivals wouldn't have time to respond; the redline would be irrelevant. Most projections, however, identify the time for a recursion to produce ASI as months to years. This would give rivals plenty of time to disrupt. Even if recursion proves too close to ASI to serve as a viable redline, the transparency infrastructure proposed by MAIM could still help states credibly commit to other, earlier redlines with wider safety margins. Transparency and verification — backed by the robust threat of disabling advances that violate agreements — are prerequisites for enforcing finer-grained restrictions. Through these mechanisms, MAIM's proposals help reinforce new redlines.

Unpacking the importance of redlines. Deterrence is the process of shaping others' perception of risk to influence their behavior. Prespecified, highly

Deterrents can be ambiguous, and ambiguity can be an asset. Generals in the US have indicated that they would respond with nuclear force to sufficiently destructive cyberattacks against its critical infrastructure, and it appears no rival has an interest in testing exactly where that line is marked. The US power grid has not been taken down by a cyberattack even though cyber deterrence does not have highly precise redlines. Meanwhile, in law, intentionally underspecified terms like "reasonable" is commonplace; law based on such language can still constrain behavior effectively. Tax law is one of the only parts of law that does not make use of such underspecified legal standards, and that area of law is particularly fragile.

This isn't to say that redlines can be arbitrarily vague. In the case of MAIM, we agree that a state threatening sabotage upon vacuous thresholds like "AGI" would not effectively shape rival expectations. But with strategically ambiguous redlines as a starting point, detailed ongoing dialogues between states can continually clarify how they would respond to specific developments. Deterrence does not need to be a guessing game. In the worst case, where all other communication has failed to align states' understandings of unacceptable behavior, early escalations themselves act as a message that a line has been crossed. In total, it would be categorically false to say that deterrence requires highly precise, prespecified redlines.

Our Best Option

Deterrence can be a source of stability. A MAIM framework keeps rival ASI projects visible and vulnerable to each other, weakening bids for dominance. Deterrence can begin with nations' unilateral capabilities and mature into a system of international verification. An opaque, heavily hardened race for superintelligence would instead invite panic and uncontrolled escalation. As we develop systems that could disrupt nuclear deterrence, MAIM is our best

See things differently? *AI Frontiers welcomes expert insights, thoughtful critiques, and fresh perspectives. Send us your pitch.*

WRITTEN BY

Dan Hendrycks

Dan Hendrycks is the Editor-in-Chief of AI Frontiers. He is also the founder and Director of the Center for AI Safety, which funds this publication.

Adam Khoja

Adam Khoja does technical and policy research at the Center for AI Safety. He studied math and computer science at UC Berkeley.

Image: Susan Wilkinson / Unsplash

CONTINUE READING

Exporting Advanced Chips Is Good for Nvidia, Not the US

The White House is betting that hardware sales will buy software loyalty — a strategy borrowed from 5G that misunderstands how AI actually works.

Laura Hiscott Dec 15, 2025

AI Could Undermine Emerging Economies

AI automation threatens to erode the "development ladder," a foundational economic pathway that has lifted hundreds of millions out of poverty.

Deric Cheng Dec 11, 2025

AI Frontiers is a platform for expert dialogue and debate on the impacts of artificial intelligence.

Donate

The views expressed in our articles reflect the perspectives of individual authors, not necessarily those of the editors or the publication as a whole. Our editorial team values intellectual variety and believes that AI is a complex topic demanding a range of viewpoints, carefully considered.