Refining MAIM: Identifying Changes Required to Meet Conditions for Deterrence

April 11, 2025(https://intelligence.org/2025/04/11/)
David Abecassis(https://intelligence.org/author/david/)

This is part of the MIRI Single Author Series. Pieces in this series represent the beliefs and opinions of their named authors, and do not claim to speak for all of MIRI.

In an op-ed (https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/) published in TIME Magazine in 2023, Eliezer Yudkowsky called for an international agreement to establish a moratorium on frontier AI development. If a country outside of the agreement was in violation, signatories should "be willing to destroy a rogue datacenter by airstrike."

The key assumptions are that there is a level of AI capabilities progress which can be **monitored** for and **interdicted**, such that successful interdiction forestalls the development of smarter-than-human systems.

In their recent work, *Superintelligence Strategy* (https://www.nationalsecurity.ai/), Dan Hendrycks, Eric Schmidt, and Alexandr Wang propose a deterrence regime in which the threat of **sabotage** is used to motivate countries (specifically, the U.S. and China) to forestall dangerous AI development. This regime is called Mutual Assured AI Malfunction (MAIM), and includes limiting AI development domestically and using the threat of sabotage to coerce rivals to do the same.

This essay argues that MAIM, as conceived in *Superintelligence Strategy*, is

unlikely to provide an effective or stable deterrence regime. While MAIM correctly frames the strategic dynamic, this analysis reveals significant flaws. Specifically, MAIM struggles with **unclear and unmonitorable red lines** for AI development, **questionable threat credibility** (particularly the capability of sabotage to deny progress), and a **volatile deterrence calculus** driven by immense stakes and uncertainty.

Still, *Superintelligence Strategy* deserves credit for advancing the key problem: states cannot be expected to sit idly by and accept the consequences of unrestricted AI development. **What characteristics should alternatives to MAIM possess?** As a precursor to answering this question, this analysis introduces a framework based on established conditions for successful deterrence, drawing insights from works such as Michael Mazarr's *Understanding Deterrence* (https://www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE295/RAND_PE295.pdf). The following sections will outline this framework, apply it systematically to the MAIM proposal, and reveal where MAIM, in its current form, fails to meet the necessary conditions for stable deterrence.

A Deterrence Framework

Simply put, *deterrence* is the imposition of costs on some action to make it less attractive.

Deterrence means dissuading one party, the **deterred**, from taking an **unwanted action**. A **red line** is clearly communicated, which acts as a trigger for a **response** by the **deterrer**. The deterrer makes a **threat**: a credible commitment to impose a **cost** via the response.

In some literature, the deterred party is called the aggressor, and the deterrer is called the defender, and these terms will be used interchangeably.

Effective deterrence requires a red line which can be **monitored**; the deterrer must be able to accurately detect infractions. The red line must also be **effective**: by constraining behavior away from the red line, the unwanted action is avoided.

An effective threat must be credible: this requires the deterred to believe in the deterrer's **capability** and **will**. Capability is the technical means to carry out the threat. Will is the resolve to do so in spite of costs the deterrer's response imposes on itself.

An effective threat imposes a cost which must have sufficient **potency**: the perceived magnitude of the cost on the deterred must be sufficient to offset gains from the unwanted action. This makes the desirability of the **gains** (from the unwanted action, to the deterred) a key consideration.

Lastly, an underlying assumption for deterrence is **rationality**: the deterred needs to be responsive to the costs and benefits of different actions. This is not a requirement for perfect rationality, merely that the deterred is considering these factors and their consequences in its decision-making at all. Indeed, the application of deterrence to this domain is owed a stronger incorporation of (Chinese) differences in perception, culture, values, and risk tolerance.

An example may help to bring this into focus: let's examine U.S. efforts to deter a potential Chinese invasion of Taiwan. In this scenario, China is the deterred party, and the United States (often alongside allies) acts as the deterrer. The unwanted action is a large-scale military attack on Taiwan. The U.S. maintains a policy of "strategic ambiguity," but the implicit red line is the initiation of such an attack. The threat includes potential U.S. military intervention and severe economic consequences, designed to impose costs that Beijing would deem unacceptable compared to the perceived gains of seizing Taiwan.

Conditions for Successful Deterrence

Using the framework, we can create a deterrence checklist:

1. The unwanted action cannot be taken without crossing the red line.
2. The red line can be monitored
3. The red line can be clearly communicated
4. The deterrer's capability is credible
5. The deterrer's will is credible
6. The threat's potency eclipses the potential net gains
7. The deterred party is sufficiently responsive to costs and benefits

We'll return to this checklist after we first describe MAIM, in more detail, using the framework. (/)

Placing MAIM into the Framework

In *Superintelligence Strategy's* vision of MAIM, we can consider the **deterrer** and the **deterred** to be the governments of the U.S. and China. In the main expectation, and as envisaged by *Superintelligence Strategy*, each fulfills both roles. There are some scenarios in which this is or begins without such symmetry, or a third party fulfills the role of deterrer, but these are outside the scope of this analysis.

The **unwanted action** is the progress of "destabilizing AI projects": those which could upset the balance of power through the creation of powerful new capabilities. This **red line** is not precisely defined within *Superintelligence Strategy* but in a recent post on X (https://x.com/DanHendrycks/status/1907495379685486812), author Dan Hendrycks gives the example of automated AI research. He says, "A fleet of thousands of AIs doing fully automated AI research (an "intelligence recursion") is an example red line and is the most credible path to a superintelligence. A recursion would likely take months and be discernible and disruptable through sabotage."

The appendix to *Superintelligence Strategy* also describes certain "critical capabilities" including "highly sophisticated cyberattack, expert-level virology, and fully autonomous AI R&D".

The **response** to tripping the red line is "maiming attacks" which can disrupt AI progress. There are many options referenced in *Superintelligence Strategy*. Specific responses mentioned include covert sabotage, overt cyberattacks, kinetic attacks, and broader hostilities which ultimately threaten non-AI assets. These are organized into an escalation ladder; a progression from smaller, covert, and focused attacks to larger, overt, and more broadly damaging attacks which can be useful because it gives the deterred a chance to back down. It can also be useful because the deterrer lacks confidence about the effectiveness of earlier approaches and so needs to try them to see if further measures are even necessary.

The **threat** is to use these maiming attacks in response to the detection of destabilizing AI projects. The **cost** ranges from the disruption of these projects to

broader disruption and attacks on non-AI assets. The inclusion of broader hostilities is not a red line because it crosses the boundary between *deterrence by denial* and *deterrence by punishment*, which we'll return to later.

The **gains** are notably amorphous in the realm of superintelligence. By *Superintelligence Strategy's* telling, the gains potentially cover the immense distance between permanently "dominating all opposition" and being so dominated or being destroyed outright by a reckless rival's project leading to loss of control. Evaluations of threat **potency** will have to contend with this uncertainty.

Some further details must be elaborated in order to apply the deterrence framework.

Red Lines

The first three conditions relate to red lines:

1. The unwanted action cannot be taken without crossing the red line.
2. The red line can be monitored
3. The red line can be clearly communicated

Does drawing a red line around fully automated AI research work? Even though Dr. Hendrycks claims this would take months, others find it plausible that an intelligence recursion could proceed too quickly for the recursion to be identified and responded to. Either way, reacting to the deployment of AI systems capable of intelligence recursion is as late in the game as one could possibly react, and leaves little margin for error. This extremely short "breakout" distance between acceptable use (e.g., autonomous software development) and the conversion of that acceptable use into a decisive strategic advantage (e.g., a cyberweapon which can disable a nuclear arsenal) is a weakness which undermines deterrence by demanding a level of monitoring fidelity and response speed that seems practically unattainable.

There is also the issue, familiar from contemporary deterrence scenarios in Ukraine and Taiwan, of "salami slicing" approaches in which an aggressor makes incremental progress toward the unwanted action in a way that makes it difficult to ever justify that the red line has been tripped. Incremental progress toward recursive self improvement is an ongoing process and the effective speedup which state-of-the-art models grant to AI researchers will increase smoothly over time.

Can the red line be drawn further back from the brink of deployed automated research, perhaps to the development of capabilities which would enable such a deployment? Even doing so still presents several problems.

Frontier AI capabilities advance in broad, general ways. For example, a new model's development does not have to specifically aim at autonomous R&D to advance the frontier of relevant capabilities. If development is proceeding on a model which is expected to be state-of-the-art at programming tasks, that likely also entails novel capabilities relevant to AI development. This suggests that the red line would need to be drawn in a similarly broad and general way, but *Superintelligence Strategy* seems to imply that we can gain the benefits of safe AI progress while avoiding escalation.

Frontier AI capabilities can be difficult to predict even within a narrow domain. Just how good will the next state-of-the-art model be at identifying software vulnerabilities? This is not something that AI developers have been able to predict.

Lastly, frontier AI capabilities can advance considerably post-training. Even if a model is exposed to competent evaluations at various stages of development, its capabilities can be extended through post-training techniques and agentic scaffolding, in ways that are not entirely foreseeable to the developers of those models early in the process. (https://techgov.intelligence.org/research/what-ai-evaluations-for-preventing-catastrophic-risks-can-and-cannot-do)

These present problems even to the developers themselves, but a foreign government faces the additional problem of monitoring these advances at a distance. The uncertainties introduced when trying to interpret the sentiments of foreign developers through espionage further complicate the difficulty of monitoring.

Superintelligence Strategy says that, "[t]he threat of a maiming attack gives states the leverage to demand transparency measures from rivals, such as inspection, so they need not rely on espionage alone to decide whether maiming is justified." But this is circular if we require threatening a maiming attack in response to some red line that we cannot monitor without the transparency measures we'd only gain through a successful threat. To resolve this circularity, states would need to make threats which target assets that are not clearly associated with dangerous AI development, which brings us to broader attacks.

Broader attacks, which hit non-destabilizing AI projects or even non-AI assets, rather than maiming attacks, could be used to gain leverage to demand transparency

measures. This seems unprecedented and wants further research. States have reasons to resist this sort of blackmail but would-be deterrers might have no better recourse short of war. Transparency measures gained through such coercion would also require the utmost scrutiny to be trusted.

This lack of effective, monitorable, and clear red lines is the major issue with MAIM as described, and would be the starting point for future efforts to further develop effective deterrence.

Credible Threats

The next two conditions relate to the credibility of threats:

1. The deterrer's capability is credible
2. The deterrer's will is credible

First, we return to the distinction of deterrence by denial and deterrence by punishment. This distinction is about the costs imposed by the threat.

In deterrence by denial, the cost is an increased expectation that the unwanted action will fail. A classic example is providing weapons to an ally to improve their ability to resist hostile military action.

In deterrence by punishment, the cost goes beyond the scope of the unwanted action, attempting to signal that said action, even if successful, would leave the deterred worse off. One example is the threat of economic sanctions in response to military aggression.

Superintelligence Strategy asserts that "states seeking an AI monopoly while risking a loss of control must assume competitors will maim their project before it nears completion." How effective can they be? This is deterrence by denial and so we can simplify our examination to look at methods which would interfere with the unwanted act directly.

Such attacks include **interfering with model development**. Model performance is tracked meticulously during pre-training so interference is plausibly noticeable long before training is "complete." Such an attack at best leads to a loss of progress, but developers can restart or revert to last known best prior to sabotage, even given that

they might not notice the damage immediately.

(/)

Kinetic-effect cybernetic weapons, in the style of the Stuxnet worm, are designed to destroy hardware directly. Yet, even Stuxnet likely only affected 11% of the centrifuges (https://www2.cs.arizona.edu/~collberg/Teaching/466-566/2012/Resources/presentations/topic9-final/report.pdf) it could have in its successful attack. For disrupting AI compute, there are substantial difficulties in addressing distributed hardware which could be of different physical makeup and operated by different organizations. This likely means that any attack could be identified and interdicted (even with simple methods like turning off the power) before it reaches anything like full coverage of the target hardware, meaning this is likely to provide only a slowdown.

Conventional weapons, including the use of hypersonics, drones, or space-based weapons face similar issues with getting sufficient coverage. As overt attacks, they also call into question the defender's will to use such attacks and risk substantial diplomatic and escalatory consequences. It is possible that the employment of an escalation ladder will help address the issue of will, as it lets the defender make incremental progress on demonstrating their determination (https://volty.substack.com/p/forecasting-chinas-response-to-the) and so reach the point where the aggressor will take them seriously.

In the tabletop exercises Superintelligence Rising (https://www.intelligencerising.org/) and AI 2027 (https://ai-2027.com/about?tab=tabletop-exercise#tab-box-tabletop-exercise), it has been a common occurrence for the U.S. and China to apply their (covert) methods of sabotage to each other's AI projects. This creates a *stable* dynamic where each side continues AI progress albeit with an ongoing impact to velocity as a result of sabotage efforts. It's obviously a limited sample, but I haven't yet encountered the scenario where unrestricted espionage, cybernetic warfare, or even kinetic strikes can actually *halt* AI development.

Understanding Deterrence suggests another important aspect of the denial/punishment distinction: denial strategies, when credible, are often considered more stable and less escalatory than punishment strategies. On the other hand, denial strategies may lack sufficient potency if they cannot all but guarantee the failure of the unwanted action. In military deterrence, denial can often be a credible capability for the defender to attain. In the context of MAIM, deterrers may not be able to credibly threaten a sufficient degree of denial.

The Calculus of Deterrence

The final two conditions relate to the deterred party's decision-making process:

1. The threat's potency eclipses the potential net gains
2. The deterred party is sufficiently responsive to costs and benefits

Evaluating these conditions reveals the immense difficulty inherent in the MAIM concept.

First, consider potency versus potential gains (condition 6). *Understanding Deterrence* establishes that a key factor is the aggressor's level of motivation, which is directly tied to the perceived gains of the unwanted action versus the perceived costs of the threat. *Superintelligence Strategy* portrays the stakes as total: the potential gain is decisive global dominance and permanent security, while the potential loss is subjugation, destruction by a rival, or omnicide through a careless project's loss of control. These perceived existential stakes create extraordinarily high motivation for states to pursue superintelligence if they believe a rival might get there first, or if they see it as the only path to long-term security.

Against these immense potential gains, the potency of the MAIM threat appears questionable. As analyzed under "Credible Threats," the denial component of MAIM (sabotaging the specific AI project) likely lacks the capability to guarantee failure, instead only imposing delays and costs. The potency of merely delaying a rival from achieving potential world dominance seems insufficient to deter a highly motivated state facing existential stakes. Therefore, for the MAIM threat to possess sufficient potency, it would likely need to rely on the threat of broader punishment—escalating to attacks on non-AI assets or even general hostilities. This shifts MAIM away from a relatively stable denial posture towards a far more dangerous punishment-based deterrence, with significantly higher risks of uncontrolled escalation.

Second, consider the deterred party's responsiveness (condition 7). Deterrence does not assume perfect rationality, but rather that the deterred party considers costs, benefits, and consequences in its decision-making. A key factor here is risk tolerance. States, like individuals, are often risk-averse. Even if the potential payoff of achieving superintelligence first is enormous, a state might be deterred from pursuing a destabilizing project if the perceived probability of failure (due to sabotage) is

sufficiently high, even if the *expected value* calculator technically favors taking the risk. The uncertainty of losing a massive investment, or sparking an escalation, could deter a risk-averse leadership. This factor potentially strengthens the case for MAIM's effectiveness.

However, several factors complicate this reliance on risk aversion. The unprecedented nature of superintelligence could induce leaders to accept risks they normally wouldn't, driven by fear or ambition. That AI development is driven by private firms may invert the government's perception of inaction; states may permit firms to forge ahead on superintelligence as a result of fear that any intervention risks upsetting what chance they have of winning. Domestic political pressures, specific leadership psychologies, or a perception of impending crisis (as Mazarr notes) could lead to decisions that appear "irrational" or highly risk-seeking from an external perspective. Furthermore, the profound uncertainties surrounding AI development speed, the level of capabilities required for dominance, the probability of loss of control, the effectiveness of sabotage, and the likelihood of escalation make a clear cost-benefit analysis extremely difficult.

This deep uncertainty cuts both ways. It might induce caution and favor deterrence, as states shy away from gambling with unknown risks. Conversely, it could fuel worst-case thinking, assumptions of hostile intent, and preemptive risk-taking to avoid perceived inevitable doom.

Finally, we can ask what each side's level of satisfaction with the status quo is. For China, there is a clear intent to create and use an advantage in science and technology to meet or exceed the U.S. in all areas of international achievement. At the same time, the U.S. also has reasons to be the aggressor, pursuing a vision of using AI in defense of Western values placed in jeopardy by a rising China. With neither power likely content to accept potential inferiority, this deterrence regime is highly unstable.

Conclusion

Applying deterrence theory reveals critical flaws in Mutual Assured AI Malfunction (MAIM) as a stable strategy. While *Superintelligence Strategy* valuably frames the problem, MAIM struggles with core requirements: unclear and unmonitorable red lines for AI development, questionable threat credibility (sabotage only likely delays, not denies), and a volatile deterrence calculus driven by immense stakes and uncertainty.

MAIM's reliance on denying rivals AI progress toward destabilizing capabilities seems insufficiently given the likely limits of sabotage capability. Achieving necessary potency would likely require dangerous escalation toward deterrence-by-punishment. The profound uncertainties surrounding AI and the high motivation of competing states further challenge the predictability needed for effective deterrence.

These weaknesses can be addressed by an alternative deterrence regime centered on earlier, more monitorable red lines. To be clear, this requires a shift from seeking deterrence *during* advanced AI development toward deterrence which *forestalls* advanced AI development altogether. For example, by operating at the level of advanced chip fabrication or concentration of compute (into datacenters) required for frontier AI development. Shifting to such a regime would enable practical monitoring and denial at the cost of delaying some other benefits of advanced AI.

Why don't we already observe threats or sabotage, beyond export controls, in service of stopping AI progress today? Another promising area for future analysis is identifying the circumstances, not yet in effect, under which we would predict the U.S. or China to instantiate deterrence.

We also see promise in integrating differing national perspectives, especially U.S.-China differences in strategic culture, risk tolerance, and technological priorities. Adjacent to this is research into deterrence regimes which integrate broader dissuasion strategies, involving inducements and credible assurances for benefits sharing alongside threats of cost imposition.

The stakes are too high for anything less than rigorous, ongoing analysis and careful statecraft tailored to AI's unique challenges.

(//)

Analysis (https://intelligence.org/category/analysis/)

Conversations (https://intelligence.org/category/conversations/)

Guest Posts (https://intelligence.org/category/guest-posts/)

MIRI Strategy (https://intelligence.org/category/miri/)

News (https://intelligence.org/category/news/)

Newsletters (https://intelligence.org/category/newsletters/)

Papers (https://intelligence.org/category/papers/)

Video (https://intelligence.org/category/video/)

Subscribe

Email

Follow us on

(https://www.facebook.com/MachineIntelligenceResearchInstitute)

(https://x.com/MIRIBerkeley)

(https://intelligence.org/feed/)

(/)

Contact

Donate

Careers

Team

Transparency

Privacy

Subscribe to our Newsletter

Email Address

Subscribe

(https://www.facebook.com/MachineIntelligenceResearchInstitute) (https://x.com/MIRIBerkeley)

Machine Intelligence Research Institute
Berkeley, California