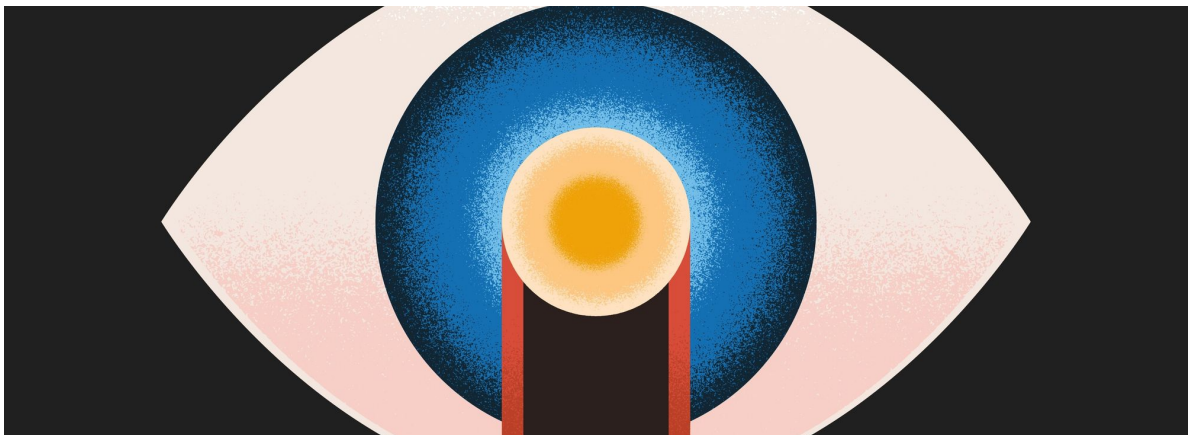
Superintelligence Deterrence Has an Observability Problem

Mutual Assured AI Malfunction (MAIM) hinges on nations observing one another's progress toward superintelligence — but reliable observation is harder than MAIM's authors acknowledge.

Aug 14, 2025

Jason Ross Arnold



In an age of heightened political division, countering China's efforts to dominate AI has emerged as a rare point of alignment between US Democratic and Republican policymakers. While the two parties have approached the issue in different ways, they generally agree that the AI "arms race" is comparable to the US-Soviet strategic competition during the Cold War, which encompassed not just nuclear weapons but also global security, geopolitical influence, and ideological supremacy. There is, however, no modern deterrence mechanism comparable to the doctrine of Mutually Assured Destruction (MAD), which prevented nuclear war between the US and Soviet Union for four decades — and which is arguably the reason no other nation has used nuclear weapons since.

The bridge from MAD to MAIM. Earlier this year, co-authors Dan Hendrycks, Eric Schmidt, and Alexandr Wang proposed a framework called Mutual Assured AI Malfunction (MAIM), hoping to fill that dangerous strategic vacuum. The authors present MAIM as a modern analogue to MAD: a grim equilibrium sustained by credible threats of mutual sabotage and the risk of uncontrollable escalation. However, unlike MAD, MAIM does not have the threat of retaliatory strikes as its core deterrence mechanism. Instead, under MAIM, "any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals."

MAIM is a compelling framework for preventing a runaway AI arms race — a race that could result in irreversible strategic dominance by either the US or China, potentially leading to an "unshakable totalitarian regime" and the subordination of democratic institutions, while also posing risks of losing control of superintelligent systems or their proliferation to dangerous actors.

effectively observe one another's frontier AI development. Because of fundamental monitoring limitations, both countries are prone to fail in their observation efforts. This creates two dangerous possibilities: (1) missing important signs of advancement that allow one country to achieve dominance, or (2) misinterpreting normal activity as a threat, triggering unnecessary sabotage. Without additional work, the concept of mutual observation as laid out in MAIM threatens to undermine the framework's strategic logic and stability.

The Logic of MAIM

Hendrycks, Schmidt, and Wang lay out the following logical progression in the emergence of MAIM (note that the language is mine):

1. Monopoly control of superintelligence is coming, and will likely lead to geopolitical dominance.
2. Major powers understand this, and are closely tracking their rivals' AI development because it's an existential national security threat
3. States which detect that their rivals are making significant progress towards superintelligence will act to sabotage them, forcing both sides to navigate a perilous but structured escalation ladder

Monopoly control of superintelligence is coming, and will likely lead to geopolitical dominance. The authors define superintelligence as "AI surpassing humans in nearly every domain." Rapid developments in machine learning during the past decade have made superintelligence much more likely. Credible forecasts predict it will arrive as early as 2027. While all the major U.S. AI labs have stated that AGI or superintelligence is their goal, China's top AI labs couch their ambitions in the language of national strategy of achieving global AI leadership by 2030. Public policies in both the US and China have tended toward acceleration, with limited constraints on frontier model

intelligence advantages, including superweapons such as AI-enabled cyber attacks, electromagnetic pulses (EMPs), next-generation drones, sophisticated antiballistic missile systems, and advanced nuclear weapons systems. These would give the superintelligence-powered country what Hendrycks et al. call "a strategic monopoly" on power.

At the same time, that country would likely become an economic behemoth, through ingenious innovations in medicine, neuroscience, finance, energy, materials science, and so many other fields — as well as a "collection of highly capable AI agents, operating tirelessly and efficiently, rival[ing] a skilled workforce, effectively turning capital into labor." The state's economic strength and unprecedented military dominance would probably make it difficult for other countries to resist its global power, enabling it to expand and consolidate its gains, including by eliminating or limiting a rival's AI program. If China reaches this dominant position first, then the prospect of "an unshakable totalitarian regime" supported by "an AI-driven surveillance apparatus" would not seem far-fetched.

Major powers understand this, and are closely tracking their rivals' AI development because it's an existential national security threat. In the national security context, the term "existential threat" does not necessarily involve the destruction of humanity. Instead, it can encompass the loss of a country's powers, and all of the attendant disadvantages. Obviously, this scenario is bad for that country and its people, but it's not quite as bad as species death.

States therefore have a strong incentive to understand (through monitoring) how far along their rivals are in developing superintelligence. The US has closely monitored China's AI progress for several years, probably going back to 2015, when Chinese Communist Party leadership issued the Made in China 2025 plan, followed by 2017's New Generation Artificial Intelligence

research facilities and US-aligned corporations like OpenAI, Google, and Anthropic.

States which detect that their rivals are making significant progress towards superintelligence will act to sabotage them, forcing both sides to navigate a perilous but structured escalation ladder. No rational state would allow its rivals to develop a technology that allowed them to usurp global dominance and pose an existential national security threat. "Rather than wait for a rival to weaponize a superintelligence against them, states will act to disable threatening AI projects," Hendrycks et al. write, with cyber attacks, missile strikes, or other military actions. Such attacks would initiate a perilous, yet theoretically manageable, escalatory sequence. Mutual awareness of this dynamic deters aggressive bids for unilateral AI dominance.

The Observability Problem Is Bigger Than MAIM's Authors Acknowledge

MAIM emerges naturally, but requires maintenance. Hendrycks and his co-authors posit that strategic deterrence, *a la* MAIM, is a natural outcome of AI race dynamics. However, in order to maximize the effectiveness of this deterrence regime — and to prevent accidental escalation — the authors argue that the US and China would need to work both independently and collectively to maintain it.

The MAIM strategic deterrence framework focuses on four main areas: communication and verification between the rivals, developing effective sabotage techniques, building data centers in rural areas where they could be safely targeted without civilian casualties, and — crucially — mutual observation.

The four problems with observability in MAIM. MAIM's observability

forecast overall progress toward superintelligence. Second, such progress will probably happen in rapid spurts; MAIM's proposed mechanisms are too slow. Third, superintelligence development will be decentralized, making it difficult to track progress linearly. Fourth, observability implies some degree of espionage, which each rival could use to fast-track its own superintelligence progress — potentially accelerating the escalation between the two rivals.

We'll explore each of these problems below; but first let's explore the dangers of having poor observability.

False positives, false negatives. The national security risks from under- or overestimating a rival could themselves be existential.

A false positive would lead a state to conclude something is present when it isn't. The US, for example, might observe that China is stockpiling high-end chips and erroneously conclude that the Chinese are surging ahead, when in fact they were preparing to roll out a nationwide cloud computing initiative. Based on this incomplete data, alarm bells would go off in Washington, where officials would activate sabotage plans (unless they're already automatically activated), possibly leading to retaliation and escalation.

A false negative error fails to detect something that is present. In this scenario, the US might miss important developments in China (e.g., US spy agencies might fail to detect the implementation of a powerful new model architecture), leading to a catastrophic failure to respond. China would have superintelligence first, and it would likely become an unmatched global power very quickly.

"distinguish between destabilizing AI projects and acceptable use," which would "reduce the risk of maiming datacenters that merely run consumer-facing AI services." But I'm not optimistic that the US or China would make this distinction both accurately and in a timely manner. So, while the authors were clearly aware of the limitations of observability, it's not clear that policymakers will appreciate their nuance.

Challenge One: Using Appropriate Proxies for AI Progress

MAIM's overreliance on the big three observables. The MAIM report doesn't come with a list of key observables, but it underlines the so-called big three: *compute, chips, and data centers*. It also notes others, such as "the professional activities of AI developers' scientists." There is no question that states' monitoring efforts should closely track rivals' compute, chip, and data center developments. But an overly limited focus on those key variables could be grievously insufficient for the task of accurately observing the rival's superintelligence progress. Such misleading or inconclusive data could prompt unnecessary preemptive actions based on misleading data, which would likely start a dangerous run up the escalation ladder. Or one country could miss critical advances leading the other to superintelligence. Game over.

AI progress could be determined by a second big three. Developmental milestones in the big three (compute, chips, and data centers) are critically important but potentially unreliable proxies for AI progress. A second "big three" — novel algorithms, architectures, or efficient data learning innovations — could provide equally important metrics.

The Chinese company DeepSeek's unexpected January 2025 breakthrough, DeepSeek-R1, is a good example of this point. DeepSeek developed R1 for

algorithmic design, among other reasons, for their success. Its emergence shocked financial markets, the tech industry, and possibly even the US intelligence community.

As stated in *The Cipher Brief*, "senior US policymakers from both the Biden and Trump administrations professed to have been surprised, as were many of the country's leading AI developers, suggesting that the [intelligence community] had not gone beyond initial research to provide a warning of what DeepSeek had accomplished."

Observability should track many proxies for superintelligence progress. This isn't to say that US intelligence should ignore compute, chips, and data centers. However, superintelligence could emerge as an unpredictable black swan, or, more likely, a semipredictable grey swan via algorithm, architecture, or data innovations. The point is, predicting an intelligence explosion will require tracking a wide range of variables: the big three, the second big three, plus others such as talent and energy (production, consumption, and innovation).

However, even if the U.S. and China had a comprehensive list of observables, and the resources to track them, their monitoring capabilities might still be insufficient. This would destabilize the MAIM deterrent system. The other three challenges to observability all follow from the core challenge of determining which proxies for AI progress are most important.

Challenge Two: Observation Must Keep Up With Rapid Progress

AI progress could outpace observability. AI progress can be as rapid as it is unexpected (e.g., DeepSeek, or a black swan scenario). A lab might achieve a breakthrough and deploy it (or lose control) before rivals can react or even gain

relies on a detectable window of progress, but rapid development threatens to shrink that window to a point where a breakthrough could occur before it can be reliably identified and acted upon.

The possibility of fast takeoff creates instability. A rapid breakthrough undermines the stability of MAIM, by increasing the risk of strategic surprise. Both the US and China will understand this nonzero probability of their rival's fast takeoff, and the difficulty of detecting it, as well as the steps leading to it.

If an uncertain but fearful superpower collects information that suggests its rival is on the cusp of a breakthrough, it will have two options. First, it can choose to wait, in the hopes of obtaining additional, clarifying information. Or it can attack, understanding that the costs of retaliation and escalation might be lower than those of relentless global dominance by the rival. The cost of a false negative (failing to detect or respond to a true threat) becomes existential. Table 1 outlines the dilemma a state in this situation would face.

	Outcome	
State A's action	State B is *not* on the cusp	State B is on the cusp
Wait	Good (No superintelligence, retaliation, or escalation from false alarm)	Catastrophic (False negative; B achieves superintelligence, dominance)
Attack	Bad (False positive; no superintelligence, but retaliation and possible escalation)	Good (A stops B's superintelligence, introducing MAIM stalemate)

Table 1: The consequences of uncertainty in the race toward superintelligence

How uncertainties could drive escalation. Worse, there might even be a kind of race to the bottom, a preemptive spiral where the uncertainties drive each

lose. Therefore, we must attack them first, especially if we suspect they're close." Each side will understand this and know that its rival understands it, creating additional pressure to preempt. This scenario makes for a deeply unstable system.

Hendrycks et al. clearly recognize rapid AI progress as a problem. Their concept of "intelligence recursion" from superintelligence speaks directly to this. Yet I question whether the proposed MAIM maintenance mechanisms (espionage, future AI-assisted verification, and response capability and readiness) can operate effectively within the tight timelines that a rapid breakthrough might present. For example, in a "fast takeoff" scenario (which, as philosopher Nick Bostrom describes in his 2014 book, "Superintelligence," would occur in "minutes, hours, or days"), the stakes are too high, and the proposed mechanisms too fragile, to have confidence in the "relative ease of (cyber) espionage and sabotage of a rival's destabilizing AI project," especially given the timeliness of effective detection and response.

Challenge Three: Superintelligence Development Will Likely Be Widely

Distributed R&D will stretch intelligence resources. The difficulty of comprehensively monitoring AI progress across many qualitative and quantitative measures makes the central MAIM task of observation difficult enough. Doing so across multiple labs, some of them with numerous worksites and decentralized development methods, multiplies the challenge. This fragmented landscape creates an incredibly large surface area to monitor. It's not clear whether US or Chinese intelligence services can succeed.

Western R&D is currently less secure. At the moment, the Chinese might have the easier task. To be sure, at least a dozen labs in the US are developing frontier AI models, each with some degree of decentralization. Adding to these are several more labs in closely allied European countries (e.g., Mistral, in France). But security remains shockingly lax, including in the biggest, most established labs. Some signs indicate that the tide is turning, but the Chinese have so far benefitted from the private sector's lack of attention to securing some of the most important US secrets.

That is not to say that US intelligence doesn't have some relative advantages. Leading Chinese firms are increasingly offering open source models, which of course provide transparency, even if the trend poses other national security risks, including proliferation and its downstream cyber and bioweapon risks. Infiltrating Chinese AI labs, however, is probably much more difficult, compared with US and allied Western labs. Beyond posing the common problem of multiple labs (some with multiple sites, decentralized methods, etc.), leading Chinese firms have close ties to the government, which they are required to assist at any time.

Open societies are more vulnerable to espionage. China's closed, authoritarian system, in which there are far fewer restrictions on domestic counterintelligence activities, provides numerous additional opportunities for the state to surveil individuals and organizations. All of this places the US and

al. briefly touch upon the decentralization issue, but they suggest it won't be a problem because AI-assisted verification "can help in the far future when AI development requires less centralization or requires fewer computational resources." However, the fact that these tools don't yet exist, and might not exist until progress towards superintelligence has already passed critical thresholds, leaves an all-important hole in the MAIM observability framework.

Challenge Four: Intelligence Activities Themselves Could Lead to Escalation

Espionage is inherently destabilizing. Espionage is one of the central mechanisms underpinning mutual observation in the MAIM deterrence model. But in a world where a state's acquisition of superintelligence leads to global domination and the subordination of its rival, each side has strong incentives to employ extreme secrecy and intensive counterespionage measures, just as the US and the USSR did during the Cold War.

One state might catch glimpses of what the other is doing, but, with all of the secrecy and counterespionage, neither will be able to observe the other with perfect transparency. Plus, it's an open question whether either or both sides would accept the other's espionage as the price for MAIM stability. The US and the USSR gave each other some leeway, in the form of a tacit understanding, to underpin MAD. But, even during the Cold War, there were instances where botched spy missions almost led to dangerous escalations.

With advanced observation comes another challenge: the type of observability required for MAIM might also enable industrial espionage between the rivals, thus escalating tensions. After all, there's a fine line between spying to monitor

Propaganda blurring progress could trigger unwarranted reactions. One or both states might also wage disinformation campaigns to complement secrecy and counterespionage efforts. The purpose would be to understate progress. China, for example, might release false information to mislead the US into believing that it is less advanced than it really is. If co-designed and co-orchestrated by advanced AI, this propaganda might be especially effective, particularly when paired with successful secrecy and counterespionage programs — all of which would help to keep the truth at bay. Although it would destabilize the MAIM regime (especially if the US detected the disinformation as such), China's effort would be rational. It would want to do everything it could to prevent sabotage. And, of course, China wants to win the race, just as the US does.

MAIM Started the Conversation on Superintelligence Deterrence, but More Dialogue Is Needed

The MAIM-based deterrence system proposed in "Superintelligence Strategy" provides a vital, contemporary analogue to MAD: a grim equilibrium sustained by credible threats. We owe Hendrycks, Schmidt, and Wang a debt of gratitude for advancing the conversation in this area. But their framework for maintaining MAIM has a critical vulnerability: multiple information problems that undermine its observability requirements and potentially exacerbate the dangerous instabilities in the race to superintelligence.

Cumulative errors undermine MAIM's efficacy. MAIM could work, but only after plugging the holes in its observability framework. The critical flaws in the MAIM framework add up, creating a narrow and unreliable model for mutual observation. Relying on such a model would lead to dangerous errors. False

— and, with it, global dominance.

No deterrence framework is perfect, but MAIM can be improved. Overall, MAIM (like MAD) can never be a perfect deterrence framework. Indeed, the observability challenges described above are the most substantial problem with MAIM; the excellent RAND analysis authored by Iskander Rehman, Karl P. Mueller, and Michael J. Mazarr outlines others. MAIM's stability as a deterrence framework depends on a set of assumptions that are, in the real world, dangerously fragile.

Despite its flaws, MAIM serves as a necessary starting point for a perilous new era of AI competition, much like early nuclear deterrence theory paved the way for MAD. Strategists and policymakers must work to make the grim equilibrium as stable as possible. Bolstering mutual observation with concrete additional metrics (like the second big three) and establishing clear thresholds for action are the necessary first steps. The global stability of the 21st century may depend on it.

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

WRITTEN BY

Jason Ross Arnold

Jason Ross Arnold is Professor and Chair of Political Science at Virginia Commonwealth University, with an affiliated faculty appointment in the Computer Science Department. He is the author of Secrecy in the Sunshine Era: The Promise and Failures of US Open Government Laws (2014), Whistleblowers, Leakers, and Their Networks, from Snowden to Samizdat (2019), and Uncertain Threats: The FBI, the New Left, and Cold War Intelligence (forthcoming in

Cover image: Ghariza Mahavira / 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

artificial intelligence.

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.