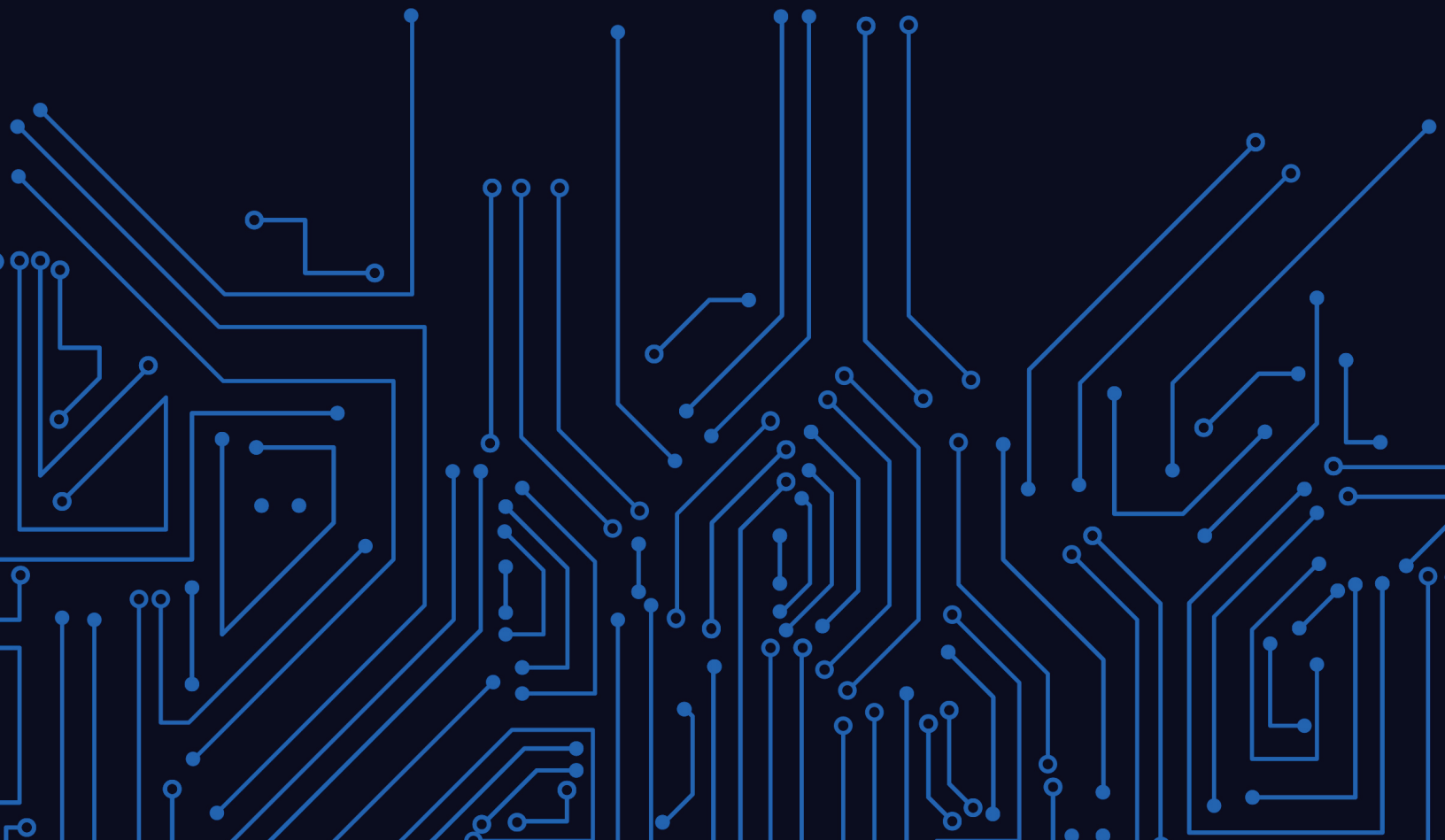
SUPERINTELLIGENCE STRATEGY

DAN HENDRYCKS ERIC SCHMIDT ALEXANDR WANG



Superintelligence Strategy

Dan Hendrycks Eric Schmidt Alexandr Wang

Abstract

Rapid advances in AI are beginning to reshape national security. Destabilizing AI developments could rupture the balance of power and raise the odds of great-power conflict, while widespread proliferation of capable AI hackers and virologists would lower barriers for rogue actors to cause catastrophe. Superintelligence—AI vastly better than humans at nearly all cognitive tasks—is now anticipated by AI researchers. Just as nations once developed nuclear strategies to secure their survival, we now need a coherent superintelligence strategy to navigate a new period of transformative change. We introduce the concept of Mutual Assured AI Malfunction (MAIM): a *deterrence* regime resembling nuclear mutual assured destruction (MAD) where any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals. Given the relative ease of sabotaging a destabilizing AI project—through interventions ranging from covert degradation of training runs to potential kinetic strikes on datacenters—MAIM already describes the strategic picture AI superpowers find themselves in. Alongside this, states can engage in *nonproliferation* to rogue actors to keep weaponizable AI capabilities out of their hands, and they can increase their *competitiveness* by bolstering their economies and militaries through AI. Taken together, the three-part framework of deterrence, nonproliferation, and competitiveness outlines a robust strategy to superintelligence in the years ahead.

1 Introduction

From geopolitical conflict to catastrophic misuse, the challenges AI poses are far too broad, and far too serious, for piecemeal measures. What is needed is a comprehensive approach, one that does not shy from the unsettling implications of advanced AI. As with Herman Kahn's famous analysis of nuclear strategy [1], superintelligence strategy requires "thinking about the unthinkable." An effective strategy should draw from a long history of national security policy because superintelligence is inescapably a matter of national security.

AI lowers barriers for acts of mass destruction once the exclusive domain of major powers. Individuals armed with an expert-level AI virologist could create novel pathogens, while advanced hacking AIs might target energy grids at a national scale. Defense often lags behind offense in both biology and critical infrastructure, leaving large swaths of civilization vulnerable. The relative security we enjoyed when only nation-states were capable of sophisticated attacks will no longer hold if highly capable AIs can guide extremist cells from plan to execution.

Militaries see in AI the key to a decisive edge, igniting a race to develop capabilities that could overturn existing balances of power. Beyond powering next-generation drones, AI raises the specter of strategic breakthroughs that might upend nuclear deterrence. A superintelligence—if its creators can control it—could conceivably deliver a "superweapon" and hand its wielders a strategic monopoly on power [2]. Superintelligence is not merely a new weapon, but a way to fast-track all future military innovation. A nation with sole possession of superintelligence might be as overwhelming as the Conquistadors were to the Aztecs. In a winner-take-all race, any hint of a looming superweapon fuels tensions, pushing rivals toward actions that

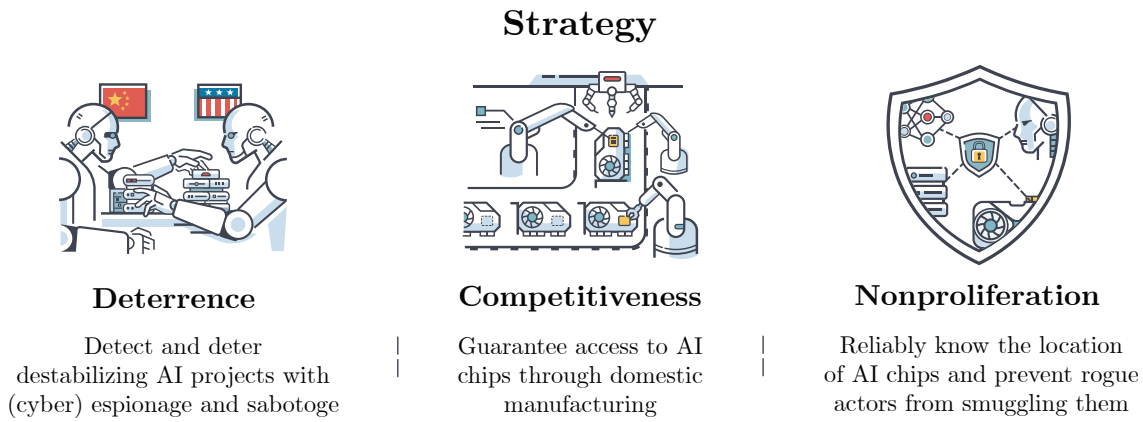


Figure 1: Effective strategies for managing advanced AI can draw from national security precedents in handling previous potentially catastrophic dual-use technology.

could quickly escalate into open conflict. AI development is fast becoming a matter of survival rather than merely technological ambition.

Another hazard emerges when AI systems can autonomously develop the next generation of AIs. Systems that can fully automate the AI research process could telescope a decade of progress into a year. A fast feedback loop could create an "intelligence explosion," resulting in AIs as uncontainable to us as an adult would be to a group of three-year-olds. Racing countries might forgo human oversight of automated research, since it would slow research from machine speed to human speed. If AI systems outpace the safeguards designed for them, we may accidentally unleash an AI which does not follow a commander's intent.

In this paper, we propose a superintelligence strategy emphasizing three key pillars of deterrence, competitiveness, and nonproliferation. We introduce Mutual Assured AI Malfunction (MAIM), arguing that when states possess common knowledge of the national security implications of AI, they will each act to sabotage rival AI projects that threaten their security. MAIM might be used to create a stable deterrence regime, preventing mutual assured AI destruction from escalating into mutual assured human destruction. Meanwhile, governments which rely exclusively on Taiwan for crucial chips leave themselves vulnerable to crippling disruption if tensions escalate. In order to retain economic and military competitiveness in a shifting landscape, nations must secure domestic supply chains for AI chips and drones. Finally, all nations have a shared interest in nonproliferation efforts to limit the AI capabilities accessible to rogue actors. Through these pillars, states can safeguard their security while opening the door to unprecedented prosperity.

Though this abridged paper describes the core of our strategy, we recommend the comprehensive version of this paper—"Superintelligence Strategy: Expert Version"—for those interested in a more detailed strategic analysis covering a broader range of challenges.

2 Existing Strategies

States grappling with terrorist threats, destabilizing weaponization capabilities, and the specter of losing control to AI face difficult choices on how to preserve themselves in a shifting landscape. Against this backdrop, three proposals have gained prominence: the first lifts all restraints on development and dissemination, treating AI like just another computer application; the second envisions a voluntary halt when programs cross a danger threshold, hoping that every great power will collectively stand down; and the third advocates concentrating development in a single, government-led project that seeks a strategic monopoly over the globe. Each path carries its own perils, inviting either malicious use risks, toothless treaties, or a destabilizing bid for dominance. Here we briefly examine these three strategies and highlight their flaws.

1. **Hands-off ("Move Fast and Break Things", or "YOLO") Strategy.** This strategy advocates for no restrictions on AI developers, AI chips, and AI models. Proponents of this strategy insist that the U.S. government impose no requirements—including testing for weaponization capabilities—on AI companies, lest it curtail innovation and allow China to win. They likewise oppose export controls on AI chips, claiming such measures would concentrate power and enable a one-world government; in their view, these chips should be sold to whoever can pay, including adversaries. Finally, they encourage that advanced U.S. model weights continue to be released openly, arguing that even if China or rogue actors use these AIs, no real security threat arises because, they maintain, AI's capabilities are defense-dominant. From a national security perspective, this is neither a credible nor a coherent strategy.

2. **Moratorium Strategy.** The voluntary moratorium strategy proposes halting AI development—either immediately or once certain hazardous capabilities, such as hacking or autonomous operation, are detected. Proponents assume that if an AI model test crosses a hazard threshold, major powers will pause their programs. Yet militaries desire precisely these hazardous capabilities, making reciprocal restraint implausible. Even with a treaty, the absence of verification mechanisms [3] means the treaty would be toothless; each side, fearing the other's secret work, would simply continue. Without the threat of force, treaties will be reneged, and some states will pursue an intelligence recursion. This dynamic, reminiscent of prior arms-control dilemmas, renders the voluntary moratorium more an aspiration than a viable plan.

3. **Monopoly Strategy.** The Monopoly strategy envisions one project securing a monopoly over advanced AI. A less-cited variant—a CERN for AI reminiscent of the Baruch Plan from the atomic era—suggests an international consortium to lead AI development, but this has gained less policymaker interest. By contrast, the U.S.-China Economic and Security Review Commission [4] has suggested a more offensive path: a Manhattan Project to build superintelligence. Such a project would invoke the Defense Production Act to channel AI chips into a U.S. desert compound staffed by top researchers, a large fraction of whom are necessarily Chinese nationals, with the stated goal of developing superintelligence to gain a strategic monopoly. Yet this facility, easily observed by satellite and vulnerable to preemptive attack, would inevitably raise alarm. China would not sit idle waiting to accept the US's dictates once they achieve superintelligence or wait as they risk a loss of control. The Manhattan Project assumes that rivals will acquiesce to an enduring imbalance or omnicide rather than move to prevent it. What begins as a push for a superweapon and global control risks prompting hostile countermeasures and escalating tensions, thereby undermining the very stability the strategy purports to secure.

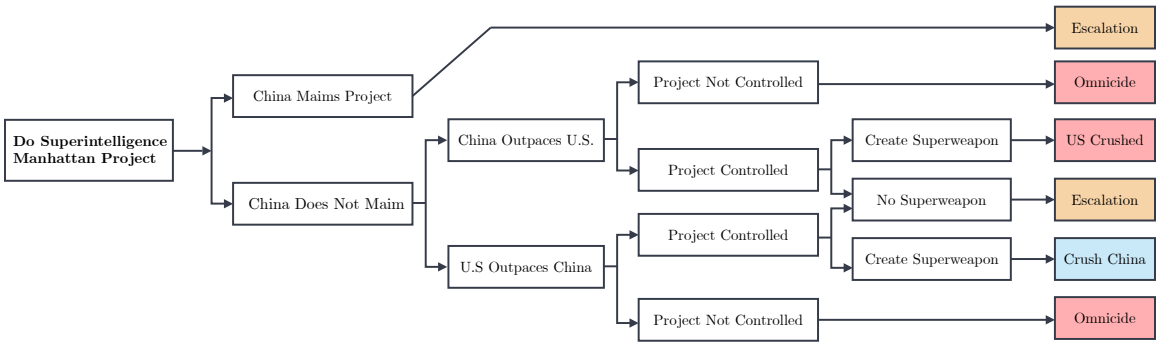


Figure 2: Possible outcomes of a U.S. Superintelligence Manhattan Project. An example pathway to Escalation: the U.S. Project outpaces China without being maimed, and maintains control of a destabilizing AI project but doesn't achieve superintelligence or a superweapon. Though global power shifts little, Beijing condemns Washington's bid for strategic monopoly as a severe escalation. The typical outcome of a Superintelligence Manhattan Project is extreme escalation, and omnicide is the worst foreseeable outcome.

Rival states, rogue actors, and the risk of losing control call for more than a single remedy. We propose three interconnected lines of effort. First, *deterrence*: a standoff akin to the nuclear stalemate of MAD, in which no power can gamble human security on an unbridled grab for dominance without expecting disabling sabotage. Next, *nonproliferation*: just as fissile materials, chemical precursors, and biological agents have long been denied to terrorists by great powers, AI chips and weaponizable AI systems can similarly be kept from rogue actors. Finally, *competitiveness*: states can protect their economic and military power through a variety of measures including legal guardrails for AI agents and domestic AI chip and drone manufacturing. Our superintelligence strategy, the **Multipolar Strategy**, echoes the Cold War framework of deterrence, nonproliferation, and containment, adapted to AI's unique challenges.

3 Deterrence with Mutual Assured AI Malfunction (MAIM)

In the nuclear age, an initial pursuit of monopoly—one nation seeking unchallenged command of nuclear weapons—eventually gave way to the standoff of deterrence known as mutual assured destruction (MAD). As nuclear arsenals matured and the capability for mutual destruction became undeniable, nations eventually accepted that any bold attempt to dominate all opposition risked drawing a preemptive strike. A similar state of mutual strategic vulnerability looms in AI. If a rival state races toward a strategic monopoly, states will not sit by quietly. If the rival state loses control, survival is threatened; alternatively, if the rival state retains control and the AI is powerful, survival is threatened. A rival with a vastly more powerful AI would amount to a severe national security emergency, so superpowers will not accept a large disadvantage in AI capabilities. Rather than wait for a rival to weaponize a superintelligence against them, states will act to disable threatening AI projects, producing a deterrence dynamic that might be called Mutual Assured AI Malfunction, or MAIM.

3.1 MAIM Is the Default Regime

Paths to Disabling a Rival's AI Project. States intent on blocking an AI-enabled strategic monopoly can employ an array of tactics, beginning with *espionage*, in which intelligence agencies quietly obtain details about a rival's AI projects. Knowing what to target, they may undertake *covert sabotage*: well-placed or blackmailed insiders can tamper with model weights or training data or AI chip fabrication facilities, while hackers quietly degrade the training process so that an AI's performance when it completes training is lackluster. This is akin to Stuxnet which aimed to covertly sabotage Iran's nuclear enrichment program. When subtlety proves too constraining, competitors may escalate to *overt cyberattacks*, targeting datacenter chip-cooling systems or nearby power plants in a way that directly—if visibly—disrupts development. Should these measures falter, some leaders may contemplate *kinetic attacks* on datacenters, arguing that allowing one actor to risk dominating or destroying the world are graver dangers, though kinetic attacks are likely unnecessary. Finally, under dire circumstances, states may resort to *broader hostilities* by climbing up existing escalation ladders or threatening non-AI assets. We refer to attacks against rival AI projects as "maiming attacks."

Infeasibility of Preventing Maiming. Since above-ground datacenters cannot currently be defended from hypersonic missiles, a state seeking to protect its AI-enabled strategic monopoly project might attempt to bury datacenters deep underground to shield them. In practice, the costs and timelines are daunting, and vulnerabilities remain. Construction timelines can stretch to three to five times longer than standard datacenter builds, amounting to several additional years. Costs balloon as well, diverting funds away from the project's AI chips and pushing total expenditures into the several hundreds of billions. Cooling the world's largest supercomputer underground introduces complex engineering challenges that go well beyond what is required for smaller underground setups. Should the supercomputer require an order-of-magnitude AI chip expansion, retrofitting the facility would become prohibitively difficult. Even those with the wealth and foresight to pursue this route would still face the potent risks of insider threats and hacking. In addition, the entire project could be sabotaged during the lengthy construction phase. Last, states could threaten non-AI assets to deter the project long before it goes online.

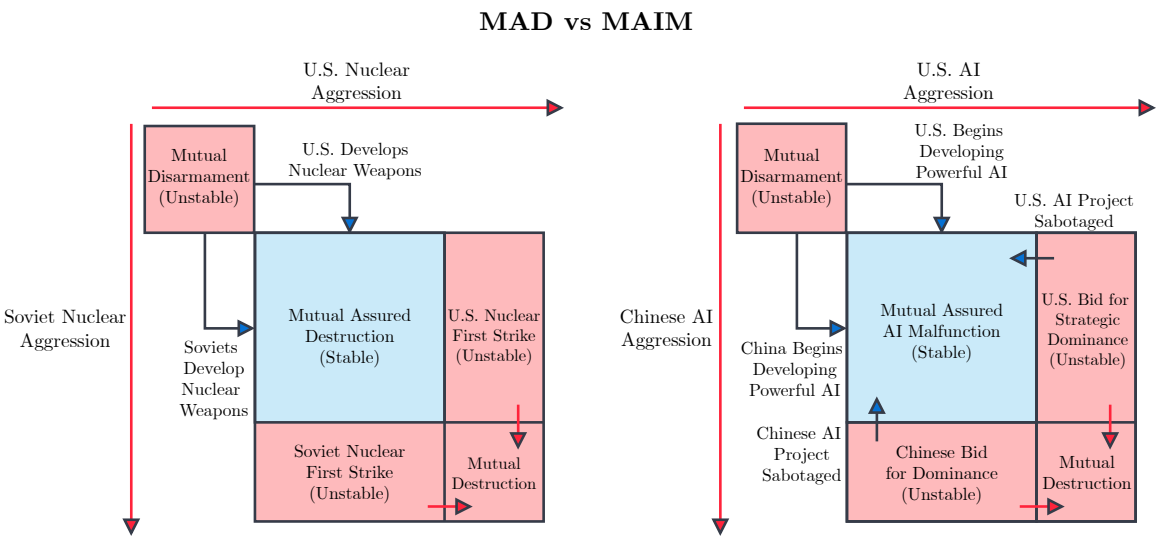


Figure 3: The strategic stability of MAIM can be paralleled with Mutual Assured Destruction (MAD). Note MAIM does not displace MAD but characterizes an additional shared vulnerability. Once MAIM is common knowledge, MAD and MAIM can both describe the current strategic situation between superpowers.

MAIM Is the Default. The relative ease of (cyber) espionage and sabotage of a rival's destabilizing AI project yields a form of deterrence. Much like nuclear rivals concluded that attacking first could trigger their own destruction, states seeking an AI monopoly while risking a loss of control must assume competitors will maim their project before it nears completion. A state can expect its AI project to be disabled if *any* rival believes it poses an unacceptable risk. This dynamic stabilizes the strategic landscape without lengthy treaty negotiations—all that is necessary is that states collectively recognize their strategic situation. The net effect may be a stalemate that postpones the emergence of superintelligence, curtails many loss of control scenarios, and undercuts efforts to secure a strategic monopoly, much as mutual assured destruction once restrained the nuclear arms race.

3.2 How to Maintain a MAIM Regime

States eventually came to accept that mutual deterrence, while seemingly a natural byproduct of nuclear stockpiling, demanded deliberate maintenance. Each superpower recognized that advanced defensive measures—particularly anti-ballistic missile (ABM) systems—could unravel the fragile balance that restrained either side from a catastrophic first strike. They responded by safeguarding mutual vulnerabilities, culminating in the 1972 ABM Treaty. By analogy, we should not leave to chance today's default condition of MAIM: where would-be monopolists, gambling not to cause omnicide, can expect their projects to be disabled. Even if attempting to harden massive datacenters is extraordinarily prohibitive and unwise, rumors alone can spark fears that a rival is going to risk national security and human security. Formal understandings not to pursue such fortifications help keep the standoff steady. We now discuss additional measures that curb unintended escalation and limit collateral damage, so that MAIM does not unravel into broader conflict.

Preserve Rational Decision-Making. Just as nuclear rivals once mapped each rung on the path to a launch to limit misunderstandings, AI powers must *clarify the escalation ladder* of espionage, covert sabotage, overt cyberattacks, possible kinetic strikes, and so on. For deterrence to hold, each side's readiness to maim must be common knowledge, ensuring that any maiming act—such as a cyberattack—cannot be misread and cause needless escalation. However, clarity about escalation holds little deterrence value if rogue regimes or extremist factions acquire large troves of AI chips. Measures to *prevent smuggling* of AI chips keep decisions in the

hands of more responsible states rather than rogue actors, which helps preserve MAIM's deterrent value. Like MAD, MAIM requires that destabilizing AI capabilities be restricted to rational actors.

Expand the Arsenal of AI Project Cyberattacks. To avoid resorting to kinetic attacks, states could improve their ability to maim destabilizing AI projects with cyberattacks. They could identify AI developers' projects or collect information on the professional activities of AI developers' scientists. To spy on AI projects at most companies, all that is necessary is a Slack or iPhone zero-day software exploit. States could also poison data, corrupt model weights and gradients, disrupt software that handles faulty GPUs, or undermine cooling or power systems. Training runs are non-deterministic and their outcomes are difficult to predict even without bugs, providing cover to many cyberattacks. Unlike kinetic attacks, some of these attacks leave few overt signs of intrusion, yet they can severely disrupt destabilizing AI projects with minimal diplomatic fallout.

Build Datacenters in Remote Locations. During the nuclear era, superpowers intentionally placed missile silos and command facilities far from major population centers. This principle of *city avoidance* would, by analogy, advise placing large AI datacenters in remote areas. If an aggressive maiming action ever occurs, that action would not put cities into the crossfire.

Distinguish Between Destabilizing AI Projects and Acceptable Use. The 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. Coordinating can help states reduce the risk of maiming datacenters that merely run consumer-facing AI services. The approach of mutual observation echoes the spirit of the Open Skies Treaty, which employed unarmed overflights to demonstrate that neither side was hiding missile deployments. In a similar spirit, increased transparency spares the broader ecosystem of everyday AI services and lowers the risk of blanket sabotage.

Large scale attack on many datacenters; threatening non-AI-related assets
Escalation to Broader Hostilities
Kinetic attacks on datacenters or corresponding power plants
Kinetic Threshold
Cyberattacks on datacenters or corresponding power plants; deleting code
Overt Sabotage Threshold
Model weights stolen; covert attacks to degrade training runs of destabilizing AI projects; cyberattacks causing GPUs to fail more often
Covert Sabotage Threshold
Espionage of AI developer workspace communications, personnel devices, and facilities

Figure 4: An example MAIM escalation ladder with maiming actions.

AI-Assisted Inspections. Speculative but increasingly plausible, confidentiality-preserving AI verifiers offer a path to confirming that AI projects abide by declared constraints without revealing proprietary code or classified material. By analyzing code and commands on-site, AIs could issue a confidentiality-preserving report or simple compliance verdict, potentially revealing nothing beyond whether the facility is creating new destabilizing models. Humans cannot perform the same role as easily, given the danger of inadvertently gleaning or leaking information, so AIs could reshape the classic tension between security and transparency [5]. Information from these AI inspections could help keep any prospective conflict confined to the disabling of AI development programs rather than escalating to the annihilation of populations. Such a mechanism can help in the far future when AI development requires less centralization or requires fewer computational resources.

MAIM can be made more stable with unilateral information acquisition (espionage), multilateral information acquisition (verification), unilateral maiming (sabotage), and multilateral maiming (joint off-switch). Mutual assured AI malfunction, under these conditions, need not devolve into mutual assured human destruction.

A standoff of destabilizing AI projects may arise by default, but it is not meant to persist for decades or serve as an indefinite stalemate. During the standoff, states seeking the benefits from creating a more capable

AI have an incentive to improve transparency and adopt verification measures [6], thereby reducing the risk of sabotage or preemptive attacks. In the Conclusion, we illustrate how this standoff might end, allowing AI's benefits to grow without global destabilization.

4 Competitiveness

Although deterrence may prevent destabilizing AI projects, competitiveness remains a decisive factor, and nowhere is this more evident than in the looming risk of a Chinese invasion of Taiwan. Multiple assessments now place the probability of a Chinese invasion or blockade in the double digits within this decade. Such a conflict would carry enormous stakes for the West: Taiwan's semiconductor fabrication plants are the source of the advanced AI chips outside China. A blockade or invasion would instantly deprive the West of AI chips, eliminating its primary edge in AI. China, having poured resources into domestic chip manufacturing for years [7], would be better insulated from this shock, while many Western nations have hesitated to build or subsidize new foundries at home. Worse, on the battlefield, China's lead in drone technologies would pose a second challenge, introducing a cheap and agile weapon system against which Western forces are not fully prepared. These events could potentially vault China ahead economically and militarily and become a unipolar force.

These vulnerabilities must be addressed before they become crises. Economic strength depends on securing access to AI chips and ending sole-source dependence on one of the world's most volatile regions. Military strength requires that states rapidly scale up drone manufacturing. By expanding and securing supply chains for AI chips and drones, governments can ensure that competitiveness is not lost the instant a single pillar—Taiwan—topples.

4.1 Economic Strength

AI Chips as the Currency of Economic Power. As AI becomes more deeply integrated in the economy, the possession of advanced AI chips may define national power. Historically, wealth and population size underpinned a state's influence; however, the automation of tasks through AI alters this dynamic. A collection of highly capable AI agents, operating tirelessly and efficiently, rivals a skilled workforce, effectively turning capital into labor. In this new paradigm, power will depend both on the capability of AI systems and the number of AI chips on which they operate. Nations with greater access to AI chips could outcompete others economically. This shift raises the stakes for any disruption to AI chip supplies. An invasion of Taiwan, always a concern, could become more tempting for China if its leaders conclude that shutting down Taiwan's semiconductor production would cripple Western AI capabilities.

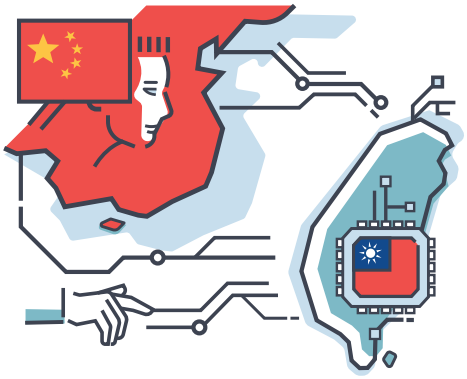


Figure 5: A Chinese invasion of Taiwan would remove the West's access to new AI chips.

Create a Secure AI Chip Supply Chain. To guard against this vulnerability, nations could invest in building advanced AI chip fabrication facilities within their own borders and reinforce a more resilient supply chain. Constructing such facilities domestically entails higher costs—over 30% more than in Taiwan—but government subsidies can bridge this gap. This effort is urgent because fabrication facilities take years to become operational.

This strategic move mirrors historical efforts to control critical technologies. During the Manhattan Project, significant investment was made not only in the development of nuclear weapons at Los Alamos but also in uranium enrichment at Oak Ridge. Similarly, ensuring access to AI chips requires substantial investment in both innovation and manufacturing infrastructure. By making the AI chip supply chain more robust, states can prevent a foreseeable devastating defeat.

4.2 Military Strength

Even if a state pioneers a breakthrough, it risks falling behind if it fails to integrate that capability into real-world operations [8]. Britain introduced the first tanks during World War I but was soon eclipsed by Germany's systematic adoption of tanks in the second World War. We turn next to three short-term imperatives for AI diffusion in the military: securing a reliable drone supply chain, carefully weaving AI into command and control, and integrating AI into cyber offense.

Secure Drone Supply Chains. Although general-purpose AI can pose larger-scale dangers, drones occupy a more conventional yet increasingly pivotal role on modern battlefields [9, 10]. Drones are cheap, agile, lethal, and decentralized, attributes that make them indispensable for states determined to keep pace with military trends. Yet many states remain heavily reliant on Chinese manufacturers for key drone and robotics components, leaving them vulnerable if those parts are withheld or disrupted at a decisive juncture.

Diffuse AI Into Command and Control and Cyber Offense. Modern battlefields demand rapid decisions drawn from torrents of data across land, sea, air, and cyber domains. AI systems can sift through these streams faster than human officers, enticing commanders to rely on automated judgments. Similarly, AI hacking systems which outpace humans in speed and cost could greatly expand a military's capacity to perform cyberattacks.

Incorporating AI into command and control and cyberoffense would significantly enhance military capabilities, yet this dynamic risks reducing "human in the loop" to a reflexive click of "accept, accept, accept," with meaningful oversight overshadowed by the speed of events. Demanding human approval of all individual lower-level engagements may be less important than ensuring explicit human approval for more severe or escalatory attacks. However, human oversight of key military decisions is nonetheless crucial. A human backstop can reduce the risk of a "flash war" [11], akin to the 2010 flash crash [12], where a minor AI mistake might spiral into destructive reprisals before any human can intervene.

By securing and expanding AI chip and drone supplies, states can safeguard their economic and military competitiveness. AI and an invasion of Taiwan are poised to reshape global power dynamics. Those who fail to act decisively will find themselves at the mercy of those who do.

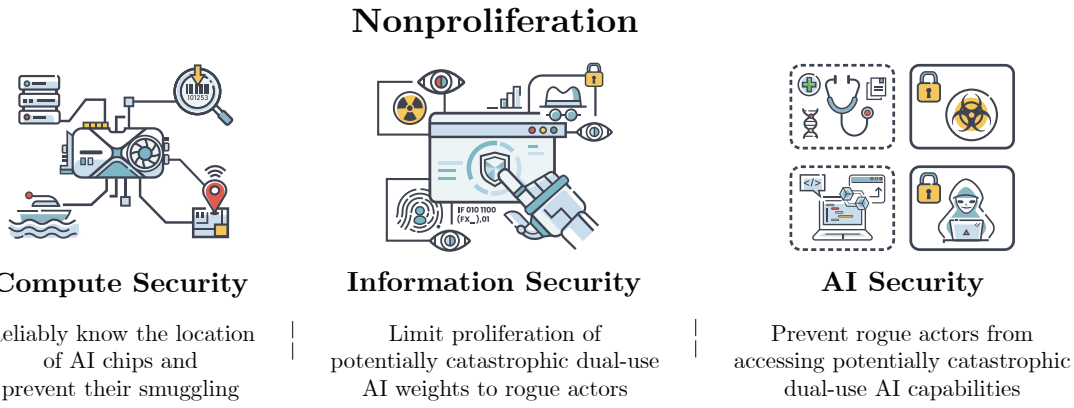


Figure 6: Due to the unprecedented scale of harm that terrorists armed with AI could cause, several lines of defense are necessary to realistically prevent proliferation.

5 Nonproliferation

For decades, states that sparred on nearly every front nonetheless found common cause in denying catastrophic dual-use technologies to rogue actors. The Non-Proliferation Treaty, the Biological Weapons Convention, and the Chemical Weapons Convention drew the U.S. the Soviet Union, and China into an unlikely partnership, driven not by altruism but by self-preservation. None could confidently manage every threat alone, and all understood that accidental proliferation would imperil them equally. Today, the same logic applies to advanced AI: if rogue actors gain the ability to launch a cyberattack on critical infrastructure or unleash engineered pandemics, no major state remains secure. Self-preservation may again impel rivals to cooperate, as each side faces grave risks if AI technology slips beyond their control.

In the context of AI, nonproliferation proceeds along three lines of defense. *Compute security* confines advanced AI chips to authorized users, mirroring the way nations restrict fissile material. *Information security* safeguards the model weights—a digital file describing the individual connections between an AI's neurons—paralleling the measures taken to secure sensitive information in the context of WMDs. Finally, *AI security* imposes guardrails on the AIs themselves, akin to safety protocols for nuclear or chemical plants, so a single actor cannot turn a civilian system into a weapon. By securing the core elements of AI in this way, states can preserve their own power and avert a world in which one rogue actor can threaten entire populations.

5.1 Compute Security

Securing high-end AI chips ("compute") is a direct outgrowth of nonproliferation principles. Just as states restrict fissile materials, chemical weapons, and biological agents, they can also restrict the specialized hardware that powers advanced AI (Figure 7). By treating AI chips as they would fissile material—cataloging each unit, supervising its destination, and guarding against unauthorized diversion—states can deny terrorists or other rogue actors the raw power to train their own weaponizable AI. This pillar of nonproliferation, which we call "compute security," has the two goals of knowing where high-end AI chips are located, and preventing them from falling into the hands of rogue actors.

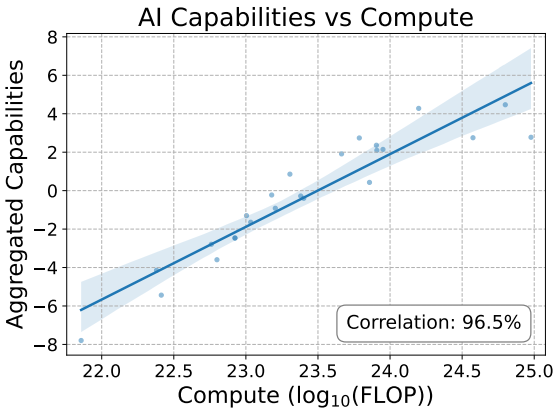


Figure 7: Compute is the most robust tracker of AI capabilities [13].

Export Controls. Export controls offer a practical way to fulfill that purpose. By drawing on existing frameworks and agencies such as the Bureau of Industry and Security, sellers of high-end AI chips would apply for a license that identifies the chips, their recipient, and any intended transfers. Entities with a strong record of compliance might earn exemptions on the condition that they notify authorities of every resale or relocation. Because a licensing regime relies on familiar infrastructure, it can be introduced swiftly, enabling officials to track chips without stalling legitimate commerce.

Enforcement. Export controls can be made more robust through stronger enforcement. A facility in Singapore, for example, might initially acquire AI chips under a valid license, only to reroute them illegally to China. More enforcement officers, assigned to in-person compliance visits and end-use checks, would detect any such deviation since the actual location of the chips would no longer match declared inventories. Any chip discovered in unauthorized hands would trigger penalties such as fines, criminal charges, or a ban on future shipments to the offending party. In addition, any AI chip declared inoperable or obsolete would undergo verified decommissioning, much like the disposal of chemical or nuclear materials, ensuring these supposedly

defunct AI chips do not get quietly resold. By tightening inspections and imposing meaningful consequences for violations, states raise the cost of covert transfers and limit the spread of advanced compute to groups that could threaten security.

These steps need not be the United States' burden alone. If the U.S. begins to treat advanced AI chips like fissile material, it may likewise encourage China to do the same. The rationale is akin to why the U.S. wants Russia to track its fissile material: no one gains from letting these capabilities slip into uncertain hands. After the Soviet Union's collapse, unsecured enriched uranium and chemical weapons in Russia posed a global threat until the U.S. initiated the Nunn–Lugar Cooperative Threat Reduction program which helped contain them. Similarly, urging China to safeguard its AI chips would acknowledge the shared imperative of avoiding catastrophes that serve no one's ambitions.

5.2 Information Security

Information security forms a second line of defense, ensuring weaponizable model weights stay out of rogue hands. If the weights of potentially catastrophic dual-use AIs are publicly leaked, they are *irreversibly proliferated*, giving rogue actors permanent catastrophic power. Hackers from terrorist cells, or even insiders hoping to "liberate" AI, could achieve this with a single successful exfiltration. Denying these groups access is feasible [14], yet an all-out push to prevent espionage by other major states may prove self-defeating. A large share of the leading AI researchers in the U.S. for instance, are Chinese nationals; dismissing them en masse might only drive talent abroad, damage competitiveness, and destabilize deterrence. Efforts to safeguard weights must therefore focus on blocking terrorists and ideologues, rather than attempting to seal every channel against a rival superpower's intelligence apparatus.

5.3 AI Security

In dual-use domains like chemistry, biology, and nuclear technology, safeguards prevent malicious use without stifling legitimate work. DNA synthesis services, for instance, screen their clients to avert the production of lethal pathogens. Analogously, AI systems can be trained to refuse destructive requests and fitted with filters that incercept "jailbreak" attempts aimed at exploiting them for advanced virology or cyberattacks. These safeguards need not impede legitimate work: authorized companies and trusted researchers can be granted broader access with fewer restrictions. Ultimately, such safeguards deny rogue actors the power to inflict mass harm without stifling innovation for lawful users.

By fortifying compute, information, and AI security, states can adapt the familiar logic of nonproliferation to the AI era, keeping AI systems from becoming instruments of terror while preserving their constructive potential.

6 Conclusion

Some observers have adopted a *doomer* outlook, convinced that calamity from AI is a foregone conclusion. Others have defaulted to an *ostrich* stance, sidestepping hard questions and hoping events will sort themselves out. In the nuclear age, neither fatalism nor denial offered a sound way forward. AI demands sober attention and a *risk-conscious* approach: outcomes, favorable or disastrous, hinge on what we do next.

A risk-conscious strategy is one that tackles the wicked problems of deterrence, nonproliferation, and strategic competition. Deterrence in AI takes the form of Mutual Assured AI Malfunction (MAIM)—today's counterpart to MAD—in which any state that pursues a strategic monopoly on power can expect a retaliatory response from rivals. To preserve this deterrent and constrain intent, states can expand their arsenal of cyberattacks to disable threatening AI projects. This shifts the focus from "winning the race to superintelligence" to deterrence. Next, nonproliferation, reminiscent of curbing access to fissile materials, aims to constrain the

capabilities of rogue actors by restricting AI chips and open-weight models if they have advanced virology or cyberattack capabilities. Strategic competition, echoing the Cold War's containment strategy, channels great-power rivalry into increasing power and resilience, including through domestic AI chip manufacturing. These measures do not halt but stabilize progress.

States that act with pragmatism instead of fatalism or denial may find themselves beneficiaries of a great surge in wealth. As AI diffuses across countless sectors, societies can raise living standards and individuals can improve their wellbeing however they see fit. Meanwhile leaders, enriched by AI's economic dividends, see even more to gain from economic interdependence and a spirit of détente could take root. During a period of economic growth and détente, a slow, multilaterally supervised project to build superintelligence—marked by a low risk tolerance and negotiated benefit-sharing—could proceed and further increase human wellbeing. By methodically constraining the most destabilizing moves, states can guide AI toward unprecedented benefits rather than risk it becoming a catalyst of ruin.

Acknowledgements

We would like to specially thank Adam Khoja for his close involvement in the creation of this paper. We would also like to thank Suryansh Mehta for his contributions to the analysis and drafting process. We would like to thank Corin Katzke, Daniel King, and Laura Hiscott for contributing to the draft. We would also like to thank Iskander Rehman, Jim Shinn, Max Tegmark, Andrew Critch, Jaan Tallinn, Nathan Labenz, Aidan O'Gara, Nathaniel Li, Richard Ren, Will Hodgkins, Daniel King, Avital Morris, Joshua Clymer, Long Phan, and Thanin Dunyaperadit.

Expert Version of *Superintelligence Strategy*

This covers a fraction of the content in "Superintelligence Strategy: Expert Version" which discusses more topics including gradual disempowerment from automation, open-weight AIs, wicked problems, firmware-level geolocation, how to better control a fleet of autonomous AI researchers, international red-lines, when AIs should refuse to help, how to align collectives of AI agents, stabilizing mass automation, and more.

References

[1] Herman Kahn. *On Thermonuclear War*. Princeton University Press, Princeton, NJ, 1960. ISBN 978-0-691-02343-1.

[2] Jim Mitre and Joel B. Predd. *Artificial General Intelligence's Five Hard National Security Problems*. RAND Corporation, Santa Monica, CA, 2025. doi: 10.7249/PEA3691-4.

[3] Akash R. Wasil, Tom Reed, Jack William Miller, and Peter Barnett. Verification methods for international ai agreements, 2024.

[4] 2024 report to congress: Executive summary and recommendations, November 2024. URL https://www.uscc.gov/sites/default/files/2024-11/2024_Executive_Summary.pdf.

[5] Andrew J. Coe and Jane Vaynman. Why arms control is so rare. *American Political Science Review*, 114(2):342–355, 2020. doi: 10.1017/S000305541900073X.

[6] Aaron Scher and Lisa Thiergart. Mechanisms to verify international agreements about ai development. Technical report, Machine Intelligence Research Institute, 2024.

[7] Reuters. China sets up third fund with $47.5 bln to boost semiconductor sector. *Reuters*, 2024.

[8] Michael C Horowitz. *The Diffusion of Military Power: Causes and Consequences for International Politics*. Princeton University Press, 2010.

[9] Stacie Pettyjohn et al. Swarms over the strait: Drone warfare in a future fight to defend taiwan, June 2024.

[10] Daniel M. Gerstein and Erin N. Leidy. *Emerging Technology and Risk Analysis: Unmanned Aerial Systems Intelligent Swarm Technology*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/RRA2380-1.

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks, 2023.

[12] Liam Vaughan. *Flash Crash: A Trading Savant, a Global Manhunt, and the Most Mysterious Market Crash in History*. Doubleday, New York, 2020. ISBN 978-0385543651.

[13] Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do ai safety benchmarks actually measure safety progress?, 2024. URL https://arxiv.org/abs/2407.21792.

[14] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RAND Corporation, Santa Monica, CA, 2024. doi: 10.7249/RRA2849-1.