Seeking Stability in the Competition for AI Advantage

Commentary on *Superintelligence Strategy* by Dan Hendrycks, Eric Schmidt, and Alexandr Wang

Mar 13, 2025

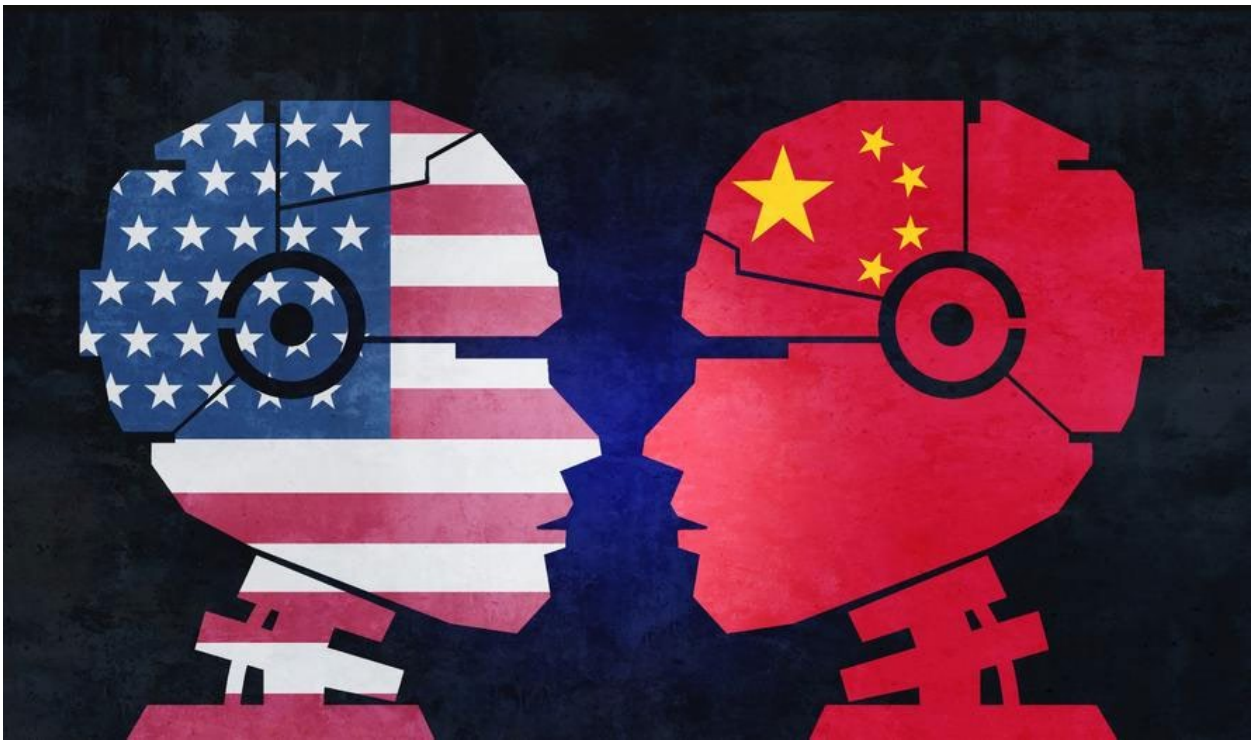

Photo by wildpixel/Getty Images

By Iskander Rehman, Karl P. Mueller, Michael J. Mazarr

In an important new report, *Superintelligence Strategy*, Dan Hendrycks, Eric Schmidt, and Alexandr Wang offer a bold vision for how the United States and China could compete securely and safely as they develop increasingly capable artificial intelligence (AI). The authors aim to synthesize national security imperatives, economic competitiveness, and AI governance into a coherent framework for urgent government action. Few experts have offered comprehensive strategies for managing the accelerating development of AI, so this essay makes a critical contribution to advancing the AI policy debate.

Many of the paper's recommendations are both sound and timely. Its most provocative proposal is a new concept to avert competition leading to instability among leading AI states that the authors dub “mutually assured AI malfunction,” or MAIM, analogizing it to nuclear mutual assured destruction (MAD). Under MAIM, they argue, “any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by its rivals.” Although it is valuable to compare the nuclear and AI revolutions in search of instructive parallels and insights, the differences between the technologies and their respective ecosystems have deep strategic implications. Taking these into account, we have concerns regarding both the practical viability of the MAIM concept as an

approach to overcoming instability risks in the AI race and the potential escalatory dangers that could follow from its core prescriptions.

Although it is valuable to compare the nuclear and AI revolutions in search of instructive parallels and insights, the differences between the technologies and their respective ecosystems have deep strategic implications.

Critical Elements of an AI Strategy

Superintelligence Strategy considers the increasingly likely scenario in which the United States and China race aggressively toward critical thresholds in the development of superintelligent AI—AI that is more capable than humans at just about every cognitive task and, when paired, or "embodied," with advanced robotics, at many other tasks as well. (This level of AI is often labeled "Artificial General Intelligence" or AGI, but that term is abstract and misleads as much as it informs.) Many AI enthusiasts argue (PDF)[1], and these authors acknowledge the possibility, that the first state to develop this technology could enjoy explosive economic growth and unprecedented military power as the AI recursively reprograms and duplicates itself at machine speed to be more and more capable. Other states could be left far behind. That scenario contains real risks: If both sides see superintelligent AI as the key to economic and technological dominance, and a potentially existential threat if it is in the hands of an adversary (whether through active hostile intent or if a rival loses control of their AI), the dangers of mutual hostility and aggression can be expected to rise the closer either side gets to achieving it.

Hendrycks, Schmidt, and Wang offer a three-part strategy to deal with this scenario. Two of the pillars are very convincing. They propose a policy of *AI nonproliferation*—an effort to limit the various risks of proliferating AI through policies and hardware-enabled mechanisms that restrict access to leading-edge AI chips and frontier model weights by nonstate actors. Given the damage that could be done by small groups with access to super-powerful AI, such an agenda—though incredibly challenging in practice (especially with the rise of open-weight AI models available to any user)—is an important priority.

The second component of their strategy is *managed competition*, which advocates and seeks to ensure that the United States continues to lead the world in AI development and diffusion. We agree and would also emphasize that competing in the AI Era involves more than building the technology stack—it's also about fashioning a society that can implement superintelligent AI effectively while managing its disruptive implications.

Many of the paper's recommendations tied to these two strategic lines of effort—compute security, export controls, information security, AI safeguards, investing in the economic foundations of AI advantage—ought to be part of any forward-looking U.S. AI strategy. The third leg of the proposed strategy—how to deter states from recklessly pursuing an advanced AI monopoly—raises serious questions, however.

Hendrycks, Schmidt, and Wang describe an international order in which threats to cripple the AI development of a rival if it sprints toward an AGI monopoly can prevent anyone from seeking such technological dominance:

---

[1] Key excerpts published here and here, with commentary.

"A state could try to disrupt such an AI project with interventions ranging from covert operations that degrade training runs to physical damage that disables AI infrastructure. Thus, we are already approaching a dynamic similar to nuclear Mutual Assured Destruction (MAD), in which no power dares attempt an outright grab for strategic monopoly, as any such effort would invite a debilitating response. This strategic condition, which we refer to as Mutual Assured AI Malfunction (MAIM), represents a potentially stable deterrence regime, but maintaining it could require care."

As they put it later in the report, "If a rival state races toward a strategic monopoly, states will not sit by quietly....Rather than wait for a rival to weaponize a superintelligence against them, states will act to disable threatening AI projects."

This is a thought-provoking idea. But we think it presents serious problems in two areas. First, it does not appear to be a feasible deterrent in practice. Second, a capability to disable an adversary's pursuit of advanced AI would exacerbate rather than dampen the instability of an AI race by creating potent first-strike incentives.

Targeting an Opaque, Distributed, General-Purpose Technology

The MAIM concept depends on several optimistic assumptions that underpin the posited "relative ease of sabotaging a rival's AI program." Many of the proposed "maiming" techniques would be extremely difficult to implement effectively, for a variety of reasons. Here we highlight just three.

First, the report refers loosely to an array of actions that states might take to cripple a rival's architecture for developing advanced AI. These range from cyberattacks and other, less clearly defined forms of sabotage to kinetic attacks on data centers and other AI-supporting facilities using hypersonic missiles. This assumes that adversary AI programs will have specific facilities that can be readily located and disrupted. However, distributed cloud computing, decentralized training, and algorithmic development increasingly may not require centralized physical locations, making AI systems more resilient to limited attacks, in addition to making adversary AI development more difficult to monitor. Moreover, as leading AI powers get closer to superintelligence, they will presumably work to harden their AI labs, data centers, and power generating infrastructures against cyber disruption and physical sabotage, and increasingly build redundancy into their systems. And if either side had to escalate from virtual disruption to outright attack, as MAIM assumes they might, a comprehensive attack to disable data centers, power generation, and AI labs in the other side's homeland is simply not a realistic option for the U.S. or Chinese militaries today short of the use of nuclear weapons.

A second practical challenge resides in the expectation that each party can accurately assess secretive AI progress by others and gauge when preventive action would be necessary. Contrary to what is averred in the report, it is unlikely that states will have a clear sense of when the moment has arrived to MAIM their opponent. The concept refers to a trigger point of "any state's aggressive bid for unilateral AI dominance," but the terms *aggressive* and *dominance* are frustratingly vague, and it's not clear that either could be defined precisely enough to appear in a planning document recommending large-scale preventive attacks on another country's homeland. Both recent developments and the history of great power competition suggest that it can be exceedingly difficult to know the exact state of an adversary's technological development, even with respect to technologies such as nuclear weapons development that involve distinctive infrastructure, well-understood science, and relatively clear developmental thresholds.

Much of the difficulty lies in a basic difference between cold war and contemporary nuclear deterrence and the MAIM concept. MAD did not seek to deter the development of weapons but instead their use, which made the threshold for response vastly simpler (though it could still be problematic in cases such as false or ambiguous warnings of attacks). No similarly clear and obvious brink exists to justify a MAIM strike as a defensive measure. From a Chinese perspective, for example, one could make a compelling case—considering public U.S. AI strategies, statements of senior officials, massive private-sector funding of AI Labs, the CHIPS Act, export controls, and the new, government-supported "Stargate" project—that the United States is *already* racing as fast as it can toward a monopoly on superintelligence. Yet we rarely hear even the most optimistic observers of U.S. AI progress predicting a Chinese first strike or even large-scale cyberattacks against the labs. Nor do we hear those ideas from the U.S. side despite growing fears of accelerating Chinese progress in AI.

Third and finally, 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. Arms prohibition regimes that have been reasonably effective in the past (such as the Non-Proliferation Treaty and the Biological Weapons Convention) draw their strength from the explicit or implicit threat of suffering broader economic, political, or military punishment for violations, not just destruction of the proscribed technological investments.

A state getting close to superintelligence would have many options to foil or avert a MAIM threat from being carried out apart from giving in. It could move to hide its research, threaten proportional responses to any attacks, recruit global support, and much else. It could also decide to accelerate rather than pause its AI work on the theory that, given the clearly hostile and suppressive intent of its rival, the only route to true security is to get there first.

Superintelligent AI would be the most powerful general-purpose technology in human history. It promises to transform societies, economies, and militaries. It seems hard to believe that a great power driven by either ambition for global dominance or fear of being the victim of someone else's decisive technological advantage would allow its AI development to be hamstrung by such a limited threat.

Superintelligent AI would be the most powerful general-purpose technology in human history. It promises to transform societies, economies, and militaries.

MAIM's Problematic Lack of MADness

In addition to these concerns over the MAIM concept's feasibility, an even more serious problem lies in its strategic logic. For all its good intentions, MAIM stands to exacerbate some of the very stability problems it aims to solve.

Although the authors often portray MAIM as an emergent reality—akin to Robert Jervis's aphorism that "MAD is a fact, not a policy"—they seem to be recommending that the United States embrace it as a path to secure AI development. But actively endorsing MAIM would declare U.S. willingness to go to war, if necessary, to prevent Chinese acquisition of a general-purpose technology with profound social and human benefits. In making the threat, the United States would encourage China to push its most advanced AI development underground (both literally and metaphorically), thus reducing potential U.S. intelligence insight into it. Other countries might not take kindly to

such a nakedly unilateral and coercive approach—especially if China offered the world access to open-source versions of its AI.

More worryingly, once a state of mutual AI vulnerability was in place, the risk of crisis and war could easily rise rather than fall. Because, as argued above, it will be so difficult to know when the critical moment of danger has arrived in a rival's development of advanced AI, a MAIM-like balance could be highly unstable, with both sides constantly misreading signals, adjusting red lines for responses, and struggling to "guesstimate" the right thresholds. Each side would be scrutinizing the others' AI development with an intensity verging on paranoia, and the potential for misperception could become very great. Highly escalatory steps could hang on bitterly contested technical—and algorithmically opaque—interpretations of AI advances. The result could be a hair-trigger balance of AI terror.

Such instability would be magnified by the fact that a state approaching the recursive AI threshold might assume it could nullify the other side's attacks, removing the mutuality from MAIM. The result would be two (or more) sides leaning forward, terrified of waiting too long to act. If tasked to avoid a superintelligence leap on the part of a rival, risk-averse national security establishments will be wary of relying on perfect intelligence in the critical moment. They might well press for dangerously preemptive action.

A selective strike on advanced AI infrastructure might also be interpreted as the opening wave of a broader attack on the targeted state's national security. AI infrastructure is likely to be deeply intertwined with economic and military power, so a strike on AI data centers (let alone electrical power grids) could be perceived as a clear act of war, triggering rapid escalation. Differentiating an incoming missile (or other) attack on its data centers from a more general attack would be difficult if not impossible.

Part of the problem with the strategic logic of MAIM is that it does not reflect how nuclear deterrence functioned in the Cold War. The paper presents MAIM as analogous to MAD, in which "any state's aggressive bid for unilateral AI dominance is met with preventive sabotage by rivals." But that is not how MAD worked. MAD was (and is) something close to the opposite: *Neither* side was able to execute effective and comprehensive preventive sabotage or strike. That was the whole point; no one could hit first and avoid a devastating response—a state of affairs that decades of arms control efforts eventually sought to reinforce.

Nor was MAD a settled, stable relationship once it emerged. The United States and the USSR embraced assured destruction, but not the mutual part—they often tried to escape the constraints of MAD with first strike-capable nuclear systems and doctrines of nuclear warfighting and "damage limitation," and consistently feared that the other side might manage to do so. The nature of the technology and its enormous destructive power prevented them from overcoming MAD, but it wasn't for lack of trying. There is every reason to think that an "AI balance of terror" would quickly generate a whole series of steps to recapture advantage.

It is also worth noting that the MAIM world described by Hendrycks, Schmidt, and Wang is entirely inconsistent with the current reality of private sector–driven AI development. In a world of existential threats to homelands from AI progress, states would surely find it untenable to allow private sector actors to take steps which could provoke war. Government control of AI research and development would seem the only conceivable answer at that point.

In a world of existential threats to homelands from AI progress, states would surely find it untenable to allow private sector actors to take steps which could provoke war.

Conclusion

Superintelligence Strategy offers an important reminder of the need to establish more coherent and whole-of-government strategy toward AI. Moreover, it serves a useful catalytic function by opening up a number of interesting questions for inquiry.

There is more to understand, and a full agenda for promoting stability on the road to superintelligence needs much more study if, to quote Isaac Asimov, we wish to ensure that science does not gather knowledge faster than we collectively gain wisdom. Even now, however, we believe it is possible to identify several important steps short of deterrent threats to attack AI infrastructure that could help.

For example, the United States could propose a multilateral dialogue on the issue of geopolitical stability on the road to AI, with the goal of identifying clear risk points and generating shared commitments of restraint. It could suggest a military-to-military dialogue on destabilizing military technologies arising from superintelligence and measures to reduce their effects. In line with the recommendations of Hendrycks, Schmidt, and Wang, the United States could propose collaboration among AI leaders (which would necessarily include private industry) on controlling proliferation in the common interest. Finally, and most importantly, the United States could announce its intention to reject some specific applications of increasingly capable AI—such as interfering in others' nuclear command and control—to offer some reassurances about the potential impact of superintelligence, although it would be naïve to expect China to be broadly reassured about U.S. intentions in the near term.

One thing that *Superintelligence Strategy* makes abundantly clear is the need for a strategic concept for managing stability among great powers on the road to transformationally powerful AI. The threat of states attempting preventive actions against their rivals' AI development merits serious concern. However, there are real questions about whether a deterrent doctrine centered on preemption would heighten rather than reduce instability, whether the concept is practically feasible, and about the weakness of analogies to Cold War–era concepts like MAD. MAIM may thus not be the right answer—but the United States and China will need more such willingness to explore bold and new ideas, and soon, to manage this perilous transition.