A significant gap exists between the expectations and fears of public opinion, policymakers, and global defense communities about artificial intelligence (AI) and its actual military capabilities, particularly in the nuclear sphere. The misconceptions that exist today are largely caused by the hyperbolic depictions of AI in popular culture and science fiction, most prominently the Skynet system in *The Terminator.* Misrepresentations of the potential opportunities and risks of military AI can obscure constructive and crucial debate on these topics—specifically, the challenge of balancing the potential operational, tactical, and strategic benefits of leveraging AI while managing the risks posed to stability and nuclear security.

This article demystifies the hype surrounding AI in the context of nuclear weapons and, more broadly, future warfare. Specifically, it highlights the potential, multifaceted intersections of this disruptive technology with nuclear stability. The inherently destabilizing effects of military AI may exacerbate tension between nuclear-armed great powers, especially China and the United States, but not for the reasons you may think.

I will begin this assessment by first examining the destabilizing characteristics of AI, such as machine-speed, bias, vulnerability, and non-human decision-making. I will then describe how these characteristics can worsen some of the key conditions for nuclear instability (or strategic instability) in the age of AI, particularly because of competitive and contested nuclear multipolar world order and differing...
escalation risk tolerance between great military powers—especially China and the United States. I will then conclude by evaluating what can be done to limit these risk factors in a future that will undoubtedly become more AI-oriented.

**AI’s Destabilizing Characteristics**

AI does not exist in a vacuum. That is, in isolation, AI is unlikely to be a strategic game-changer. Instead, it will likely reinforce the destabilizing effects of existing advanced conventional—especially counterforce—capabilities, thereby increasing the speed of warfare and compressing the decision-making timeframe. Although AI-augmented command and control (C2) systems could mitigate many of the shortcomings inherent to human strategic decision-making such as the susceptibility to invest in sunk costs, skewed risk judgment, cognitive heuristics, and group-think, its effect at a strategic level remains uncertain.

**In isolation, AI is unlikely to be a strategic game-changer.**

**Machine-speed AI warfare**

AI introduces a unique means to operate and respond at machine speed in the use of military force. In military arenas where a premium on autonomy and speed exist (e.g., missile defense, autonomous weapon systems [AWS], and cyberspace), faster reaction times will likely have outsized strategic effects. While the current generation of automated missile defense systems (e.g., terminal high-altitude area defense [THAAD] and Patriot systems) are capable of tracking and prosecuting hostile targets without human intervention, they are unable to autonomously monitor and improve their performance, independent of human intervention. AI would give the next generation that ability.

AI technology fused with advanced weapons (potentially simultaneously across multiple combat zones) could allow these systems to react at machine speed and boost the overall pace of combat. Despite the tactical advantages of being able to react in real time, especially in asymmetric contested environments, experts warn that massive increases in the speed of combat could result in machines reacting to combat situations at a pace that surpasses human comprehension—so much so that commanders might be unable to control, contain, or terminate events.

Because human commanders would be unable to react quickly enough, the decision to delegate control to an autonomous system would be a challenging ethical and tactical dilemma. Thus, while AI-enabled autonomous early-warning systems would theoretically allow defense planners to identify and
monitor threats faster and more reliably than before, the lack of human judgment and supervision coupled with the inherent brittleness (or lack of real-world common sense to deal with new situations) and black box (or opaque and unexplainable) characteristics of AI machine-learning algorithms mean that the risk of destabilizing accidents and false alarms will likely rise.

As strategist Thomas Schelling argued, “when speed is critical, the victim [in this case a nuclear-armed state] of an accident or a false alarm is under terrible pressure.” This point is especially true of the time pressures associated with AI. Fear created by the unpredictability and uncertainty of an enemy conducting warfare at machine speed may, therefore, tempt nuclear states to automate their nuclear retaliatory capability. That is, all else being equal, a nuclear-armed state less confident in its second-strike capabilities (i.e., China, North Korea, Pakistan, and perhaps Russia) will be more inclined to use automation.

The high speed associated with AI and autonomy can have unexpected and escalatory outcomes. For example, the 2010 stock market flash crash—which reduced the stock market value by one trillion dollars within a matter of minutes—was, according to the US Securities and Exchange Commission (SEC), enabled and amplified by the use (or misuse) of autonomous financial trading algorithms. Unlike the financial markets, however, there is no overarching authority in international relations to enforce pre-programmed fail-safe mechanisms based on a shared set of rules. Thus, the prospect of a flash crash-like event occurring in adversarial, loosely regulated, offense-dominant, and strategically competitive domains, such as cyberspace, missile defense, or anti-satellite weapons (ASATs), is especially alarming.

Military AI systems functioning at machine speed could push the pace of combat to a point where the actions of machines eclipse the ability of human decision-makers to control (or even comprehend) events. In extremis, human commanders might lose control of the outbreak, course, and termination of warfare. Were humans to effectively lose (or pre-delegate) control of warfare to machines, inadvertent escalation pathways and crisis instability would increase, potentially with catastrophic results. Compelled by the speed and precision of AI to make decisions in a compressed timeframe, a state might accept higher risks and escalate a conflict with the belief it was in a use it or lose it situation, or a lack of confidence in its ability to guarantee the safety and control of its nuclear arsenals.
Bias in machine learning

Despite the speed, diverse data pools, and processing power of algorithms compared to humans, machine-learning systems will still depend on the assumptions encoded into them by human engineers, who risk inadvertently sowing their own biases into the systems they design. Implicit or explicit biases baked into military support systems may cause errors from feedback loops (i.e., action and counter-reaction cycles), especially in cluttered and complex battlefield environments. For example, during the Cold War, the Soviets developed a computer program known as VRYAN designed to notify Soviet leaders of a pre-emptive US nuclear strike. However, the data used to feed the system was often biased and thus propelled a feedback loop that heightened the Kremlin’s fear that the United States was pursuing first-strike superiority. These feedback loops generated by AI systems could potentially trap human operators into machines’ bias and flawed assumptions.

Machine learning systems, then, function much like black boxes, making them technically highly unpredictable and potentially unreliable. Moreover, the added complexity of AI systems will likely amplify existing human uncertainties about the value, scope, availability, credibility, and interpretation of information. For the foreseeable future, narrow AI-infused sensing, self-learning, intelligence gathering and analysis, and decision-making support systems will continue to exhibit a similar penchant for cognitive bias and subjectivity (e.g., attribution error, decision-making heuristics, path-dependency, and dissonance) that has long plagued the human foreign policy and national security decision-making process.

Vulnerability to cyber-attacks

AI has the potential to both reduce and increase a military’s vulnerability to cyber attacks. On one hand, AI cyber-defense tools (or counter-AI), which are designed to recognize changes to patterns of behavior in a network and detect anomalies, can automatically recognize software code vulnerabilities, potentially forming a more robust defense against cyber subversions. The US DoD’s Defense Innovation Unit (DIU) is prototyping an application that leverages AI to decipher high-level strategic questions, map probabilistic chains of events, and develop alternative strategies in order to make DoD systems more resilient to AI-augmented cyberattacks and fix errors faster than humans.

On the other hand, an adversary could, for example, use malware to take control, manipulate, or fool the behavior and pattern recognition systems of autonomous systems. In one case, analysts found it extremely challenging to detect a malware that infected systems controlling the cockpits of American UAVs (Predator and Reaper drones) flying missions over Middle Eastern warzones at the Creech US Air Force Base in Nevada in 2011. Offensive attacks such as
this would be relatively easy to execute but difficult to detect, attribute, or effectively counter.

It is now thought to be possible that a cyberattack could infiltrate a nuclear weapons system, threaten the integrity of its communications, and ultimately (possibly unbeknownst to its target) gain control of its nuclear as well as non-nuclear command and control systems. Advances in AI might, therefore, exacerbate this challenge by enabling improvements to the cyber offense, thereby conferring further advantages to first movers in this domain. For example, machine learning and AI could dramatically reduce the high levels of labor intensity and technical skill required to execute advanced persistent threat (APT) operations—or “hunting for weaknesses.” Future AI APT tools could offer a would-be attacker a cheap and easy way to replicate powerful cyber weapons, which would require minimal technical knowledge to employ.

While manipulations and system subversions are possible with existing cyber offense tools, AI and increasing degrees of military autonomy could amplify the potential speed, power, and scale of future attacks in cyberspace. In response to these anticipated vulnerabilities and to achieve the first-mover advantage, China, Russia, and the United States have continued to harden their AI cyber defenses. For example, open sources indicate that Chinese analysts view the vulnerability of China’s nuclear command, control, and communication (NC3) systems to cyber infiltrations as highly escalatory, even if an attacker’s objective was limited to cyber espionage. China’s fear that its early-warning missile systems are inadequate to respond to a disarming first strike by the United States has prompted Beijing to prioritize mitigating false negatives, when early-warning systems fail to warn of an imminent attack (as opposed to a false positive that warns of an attack that does not exist). AI-enhanced cyber capabilities simultaneously bolstering deterrence and incentivizing others to attack could intensify a paradox of enhanced capabilities and increased vulnerabilities in the cyber domain.

Non-human decision-making
As AI systems become more integrated into the strategic decision-making process, the potential vulnerabilities and risks associated with military autonomy will likely increase. According to former DARPA director Arati Prabhakar, despite AI’s potential tactical and operational impact, the technology remains “fundamentally limited” due to its propensity to misidentify objects and be easily spoofed. Prabhakar adds that decision-making errors committed by AI systems are often inexplicable and unlike errors humans would make. In one example, Prabhakar demonstrated how a picture of a baby holding a toothbrush was misidentified by a machine-learning algorithm as a baby with a baseball bat. Unpredictable
errors and decisions made by AI systems deployed at scale and across multiple domains and combat theaters would be inherently destabilizing.\(^{21}\)

The inability of AI to understand context (i.e., the rationale and consequences of actions) or empathize (i.e., determine intent) would likely become a liability during wartime, when a degree of flexibility down the chain of command is generally considered positive. For example, the near-catastrophic ICBM test at the US Vandenberg Air Force Base during the 1962 Cuban Missile Crisis was attributed to officers following pre-defined protocols without questioning these guidelines in the context of new information.\(^{22}\) Human errors typically occur at an individual level and seldom repeat in the same way; by contrast, AI systems may conceivably fail simultaneously and repeat this failure indefinitely. In his vision of an AI future, for example, philosopher Nick Bostrom argues that “general” AI applications could overrun the universe in their single-minded pursuit of the efficient manufacturing of paper clips.\(^{23}\) That is, AI-enhanced systems making decisions in pursuit of pre-determined human goals in fundamentally non-human ways, with uncertain and potentially unintended outcomes.

The notion, albeit speculative, of superintelligent AI systems (i.e., machines that possess intelligence that surpasses human cognitive abilities) pitted against humans in the context of nuclear weapons would raise similar questions to the ones human commanders in modern warfare face:\(^{24}\) what are AI's intentions and motives? How might intelligent machines be deterred, coerced, or manipulated? And how could they be co-opted to de-escalate a situation? General AI systems functioning at machine speed and armed with oracle-like predictive foresight could dominate the strategic decision-making process, outsourcing the role of human cognition and agency in the mechanisms designed to control and manage the outbreak, escalation, and termination of warfare.\(^{25}\)

A broader point here is that under crisis and conflict conditions, the deterrent effect of AI is predicated on the perceived risks associated with a particular capability it enables or enhances. The higher the uncertainty generated by a capacity, deploying AI-augmented capabilities in a crisis might actually encourage an adversary to act more cautiously, and in turn, bolster stability. Thus, the uncertainty caused by the introduction of AI into a situation might incentivize states (especially those facing a superior adversary) to delegate decisions to machines for the perceived deterrence effect.\(^{26}\) Because of the difficulty of demonstrating a posture like this before a crisis or conflict, however, this implicit threat could equally worsen crisis instability. AI systems that are programmed to aggressively pursue tactical and operational advantages, for
example, might misperceive (or ignore) an adversary’s bid to deescalate a situation as prelude to an imminent attack. These dynamics would increase the risks of inadvertent escalation and first-strike instability.27

If commanders decide to delegate greater authority to inherently inflexible AI systems, the dehumanization of future defense planning will undermine stability by significantly inhibiting induction. Human induction (i.e., the ability to form general rules from specific pieces of information) is a crucial aspect of defense planning, primarily to manage situations that require high levels of visual and moral judgment and reasoning.28 Some analysts have warned that if human commanders place too much confidence in AI analysis without fully comprehending how machines reach a particular outcome, machine-generated data could be trusted implicitly and without scrutiny by human decision-makers.29 For example, the Tesla Model 3 crash in 2018, when a driver in autopilot mode plowed into a fire truck on a freeway, clearly demonstrated the risk of placing too much trust in autonomous technology—though not necessarily AI.30

The tendency of humans to use automation as a heuristic replacement for vigilant information seeking, cross-checking, and adequate processing supervision is known as automation bias.31 This tendency could mean that defense planners might become more predisposed to view decisions generated by AI algorithms as analogous (or even superior) to those of humans. Absent human judgment, intuition, and accountability, over-reliance on automation in military decision-making will likely create conditions ripe for strategic instability.32

Conditions for Nuclear Instability

The destabilizing conditions that AI could impose have the potential to undermine global nuclear (or strategic) stability. The concept of nuclear or strategic stability emerged in the latter half of the twentieth century, and despite being theoretically and politically contested to this day, it has proven a useful intellectual tool for analyzing the potential of technically advanced weapons to undermine stability. While strategic stability has no single, universally accepted definition, contributing factors, or even agreed-upon metrics to measure it,33 it can be best thought of as a situation wherein no country has incentives to launch a nuclear first strike—thus reducing the danger of a crisis caused by miscalculation or misperception escalating into all-out war.34 It is ultimately a product of a complex interplay of political, economic, and military dynamics in which technology performs several functions.35 The role of technological change and strategic stability is part of a complex interaction of disruptive forces which, during periods of heightened geopolitical rivalry, great power transitions, and strategic surprise, may erode strategic stability and make conflict more likely.36
Two existing conditions for nuclear instability likely to be exacerbated by AI are nuclear multipolarity and differing escalation risk tolerance between military powers. The coalescing of these conditions in the presence of the inherently destabilizing military AI systems, described earlier, will lead to an interplay of forces ripe for miscalculations and misperceptions between nuclear-armed states, increasing the dangers of escalation and deterrence failure under the nuclear shadow.

Nuclear multipolarity
The emergence of nuclear multipolarity in the Second Nuclear Age has created multifaceted escalation pathways to a nuclear confrontation involving an expanding number of nuclear-armed poles, compared with bipolarity during the Cold War era. This multipolarity is important precisely because each state will choose a different response to the new choices emerging in the digital age. Motivated states could eschew the limitations of AI, compromising safety and verification standards to protect or capture the upper hand (the first-mover advantage) on the future digitized battlefield.

The historical record attests that strategic competition—motivation to control warfare—tends to be ratcheted up because of the complexity of military technology and operations over time. Thus, the pursuit of AI technology by great powers—especially China, the United States, and Russia—will likely compound the destabilizing effects of AI in increasing great-power competition.

Against this inopportune geopolitical backdrop, perceived strategic benefits of AI-enhanced weapons generate risks to nuclear security and stability—the greatest of which today is the premature adoption of unsafe, unverified, and unreliable AI technology in the context of decisions to use nuclear weapons, which could be catastrophic. For example, the proliferation of low-risk and low-cost AI-augmented autonomous weapons with ambiguous rules of engagement, such as drone swarms, will become an increasingly enticing asymmetric option to undermine an adversary’s military readiness, deterrence, and resolve.

Competing states making decisions under the nuclear shadow will be more inclined to assume the worst of others’ intentions, especially in situations where the legitimacy of the status quo is contested (e.g., maritime Asia). According to scholar John Mearsheimer, “as long as the system remains anarchic, states will be tempted to use force to alter an unacceptable status quo.” Thus, efforts by one state to enhance the survivability of its strategic forces with state-of-the-art dual-use technology like AI could easily be perceived by the other side as a
potentially threat to its ability to survive, and respond to, a nuclear first strike or second-strike capability.\textsuperscript{42}

In a world of revisionist and dissatisfied nuclear-armed states, it seems improbable that improvements in intelligence collection and analysis derived from advances in AI would have a stabilizing impact.\textsuperscript{43} For this to happen, equal access to intelligence and shared confidence in the accuracy and credibility of these systems would be required. Furthermore, the intentions of all parties would need to be benign for any reassurances or confidence-building efforts to succeed. Because nuclear interactions increasingly involve the complex interplay of nuclear and non-nuclear (and state and non-state) actors, the leveraging of AI in this multipolar context will increasingly place destabilizing pressures on nuclear states. Taken together, these interactions will likely complicate escalation management efforts during future crises or conflicts—especially involving China and the United States.

\textbf{Differing escalation risk tolerance in the US-China dyad}

While US defense analysts and their Chinese counterparts are aware of the potential escalation risks between nuclear-armed great powers, their respective doctrines do not address how an adversary might respond to escalatory behavior. Instead, these rival strategic communities generally assume that escalation in future conflict can be effectively countered and contained by establishing and sustaining escalation dominance. Chinese doctrinal emphasis on seizing the initiative early and preemptively in conventional warfare to achieve escalation dominance may result in the opposite outcome: triggering rapid and possibly uncontrollable escalation to a nuclear level of conflict.\textsuperscript{44}

However, the evidence does not suggest that China intends to use nuclear missiles for escalation dominance or deliberate escalation. Absent commonly held escalation thresholds and a mutual framework to deter either side from violating them, a Sino-American crisis where either side operates under the assumption that it can effectively control escalation (e.g., in the South China Seas, the Korean Peninsula, or Taiwan Straits) would likely increase the risks of inadvertent escalation—because both sides overestimate their ability to prevent a situation from spiraling out of control.\textsuperscript{45}

According to professor of political science at the University of Pennsylvania Avery Goldstein, Chinese overconfidence in its ability to prevent conventional military confrontation escalating to nuclear war might increase the risk of a conflict or crisis inadvertently or accidentally crossing the nuclear Rubicon.\textsuperscript{46} Moreover, divergent views of Sino-American attitudes about controlling escalation below and above the nuclear threshold could also be detrimental to crisis stability. The US defense community is generally more concerned that a low-level
conventional conflict might escalate to a nuclear level but are, apparently, less concerned about the ability of the United States to control escalation above the nuclear threshold. Paradoxically, during a Sino-American crisis, Washington may overstate the possibility that Beijing would use nuclear weapons and simultaneously understate the scale of a Chinese retaliatory nuclear response.\footnote{47}

Furthermore, in a future Sino-American conflict, the United States would have a strong incentive to pre-emptively attack China’s mobile missiles and attendant (and likely dual-use) command, control, communications, and intelligence (C3I) systems to achieve escalation dominance, which Beijing could misperceive as a conventional counterforce attack, or worse, as a precursor to a first nuclear strike. Chinese analysts generally assume that the United States intends to undermine China’s relatively small nuclear deterrent and attendant support systems with advanced conventional weapons—especially US conventional prompt global strike and missile defenses.\footnote{48} In short, divergent US-China attitudes about the ability to de-escalate a low-intensity conventional or nuclear conflict makes a conventional conflict more likely to escalate to a high-level of intensity.\footnote{49}

Managing military—especially inadvertent—escalation risk has not been a traditional feature of Chinese strategic thinking.\footnote{50} China’s strategic community shares a high level of confidence in the ability of China’s long-standing no-first-use (NFU) nuclear pledge to control escalation.\footnote{51} Because Chinese analysts view China’s NFU commitment as a de facto firebreak between the use of its conventional and nuclear capabilities to de-escalate a situation, the resultant overconfidence might increase inadvertent escalation risks.\footnote{52} This overconfidence could make it less likely that Chinese leaders recognize escalation risks caused by miscalculation or misperception of US intentions.

China’s sanguine attitude to managing escalation can, in part, be attributed to the belief that once the nuclear Rubicon is crossed, the use of nuclear weapons cannot be easily controlled by either side. Because of this thinking, Chinese analysts do not believe that a limited nuclear war would stay limited. Moreover, China’s operational doctrine does not contain plans to wage a limited nuclear war, which China might pursue if it believed nuclear escalation could be controlled.\footnote{53}

Another complicating factor is that new escalation thresholds and operating norms for AI-augmented weapons have yet to emerge. Today’s thresholds in the context of autonomous weapons systems are considered inappropriate and ambiguous.\footnote{54} Without commonly held operational norms and an adversary’s strategic priorities and political objectives,
militaries deploying military AI could inadvertently cross already vague escalation thresholds.\textsuperscript{55}

In 2016, for example, China captured a US underwater drone, asserting it posed a hazard to Chinese maritime navigation. Washington, in response, called China’s behavior “unlawful,” claiming the drone was a “sovereign immune vessel.”\textsuperscript{56} This episode ended with China returning the drone after days of diplomatic contention, demonstrating the potential risk of inadvertent escalation caused by the ambiguities surrounding the deployment of new (and especially dual-use) technology in contested territory between strategic rivals. In sum, the combination of first-strike vulnerability and opportunity enabled by a growing portfolio of technologies such as AI for warfighting, coercion, and influence will have significant implications for escalation dynamics in future warfare.

**When AI Goes Nuclear**

Military AI is likely to exacerbate the destabilizing and escalatory effects of an increasingly complex interplay of advanced military technology in a multipolar nuclear world order. Nuclear-armed states leveraging AI to achieve or sustain first-mover advantages in this multipolar context will likely destabilize this fragile order with uncertain outcomes.

Given the multifaceted interplay AI-augmented enabling capabilities might have with strategic weapons (both nuclear and conventional weapons with strategic effects), it will be particularly important for analysts, academics, and decision-makers alike to a) develop a robust understanding of the confluence between these diverse capabilities and b) understand how competing strategic communities view these dynamics, as well as the implications of these trends for nuclear and conventional strategy and posture, arms races, arms control, escalation management, and cross-domain as well as extended deterrence.

To pre-empt and mitigate, or at least manage, the destabilizing and escalatory risks posed by the military AI, great military powers (especially the United States and China) must closely coordinate their confidence-building measures to pre-empt some of the risks to stability described in this article. Specifically, great powers should establish an international framework for governance, norms, regulation, and transparency in the development and deployment of AI-augmented military capabilities. These frameworks must encompass not only the present but also potential future developments, particularly what is and is not being baked into AI algorithms, and the best method to temper public debate from becoming too fixated on killer robots and machine overlords.

Ultimately, success in these efforts will require all stakeholders to be convinced of the need and the potential mutual benefits of taking steps toward the
establishment of a coherent governance architecture to institutionalize and ensure compliance with the design and deployment of AI technology in the military sphere. Future research should investigate how incentives could be altered to enhance strategic stability and what counter-AI and adversarial AI tools are needed to mitigate these risks.

Notes


27. Yuna Huh Wong et al., *Deterrence in the Age of Thinking Machines*, 60–61.


45. Morgan et al., Dangerous Thresholds, 169–70.

46. Goldstein, “First Things First.”


52. Avery Goldstein, “First Things First.”


