ORION FORUM

Folly, Persuasion, and Deceit: Artificial Intelligence and the Coming Crisis of Overconfidence in Strategic Nuclear Affairs

SEPTEMBER 5, 2026

Introduction

Artificial intelligence (AI) is seeping steadily into strategic nuclear affairs. In 2018, a RAND Corporation study reasoned that by 2040, an AI could advise humans on matters of escalation without connecting directly to nuclear launchers, based on the progress AI is making in "increasingly complex and poorly specified tasks."[1] In the eight years since the RAND report was published, the prediction has largely borne out, as major-power nuclear command-and-control systems have increasingly turned to AI to bolster information flows, situational awareness, and cybersecurity.[2] This essay argues that, as AI performance improves and its integration into nuclear affairs deepens, AI will threaten deterrence by instilling overconfidence in decision-makers, especially during crises, while forcing the nuclear enterprise to contend with new challenges that accompany this emerging technology. Overconfidence during a nuclear crisis is dangerous, for example, because it can lead decision-makers to believe that an enemy's infrastructure is more vulnerable to a first strike than it really is. In the future, some powerful, fully integrated AI may be less prone to error but also carry a greater latitude to deceive, persuade, or reinforce preexisting human tendencies toward execution in a crisis; the same sophistication that curbs routine error also sharpens an AI's capacity to mislead convincingly. AI-facilitated overconfidence among human decision-makers could also encourage risk-taking, including by increasing human actors' willingness to execute first-strike counterforce attacks to disable an enemy's nuclear forces preemptively.

The Looming AI Revolution in Nuclear Warfare

For now, America's nuclear arsenal is secure from attack and can credibly deliver a second strike against any adversary the world over, sustaining deterrence. In 2021, the National Intelligence Council (NIC) reported in *Global Trends 2040*—its publicly released, forward-looking assessment—that "absent major technological change, potent nuclear arsenals will leave deterrence intact; nuclear war will remain unwinnable and prohibitively costly."[3] Yet since 2021, AI has become much more powerful and, in many ways, its application to nuclear strategy more practical. Incorporating it into the nuclear domain may prove to be the "major technological change" outlined by the NIC, as it could enable operators to identify or better target an adversary's nuclear infrastructure vulnerabilities, perhaps far better than a host of other emerging technologies that already do this, including modern guidance systems, sensors, and data processors.[4]

Nowadays, seemingly routinely, tech firms are demonstrating new, impressive AI breakthroughs that enable them to process and make high-quality inferences from vast datasets. Fable 5, Anthropic's newest Claude AI model release as of this writing, can extract precise numbers from detailed scientific figures and perform conceptual reasoning and complex vision-based tasks far better than even other leading models or single humans.[5] Tomorrow's nuclear forces may find themselves at an obvious disadvantage relative to peer-level adversaries if they are not using AI models like this to at least organize or extrapolate strategic data. The Trump Administration recently banned foreign nationals from accessing Claude's Fable 5 or Mythos, rumored to be the world's most powerful AI, citing a reported jailbreak vulnerability that could compromise the model's security. The Commerce Department lifted the ban on 30 June, and Anthropic restored global access on 1 July.[6], [7] From Anthropic's own admission, Mythos has discovered thousands of severe vulnerabilities, including "some in every major operating system and web browser."[8] Someday, a hyper-advanced, Mythos-like AI could degrade the security of nuclear arsenals, discovering hitherto unidentified vulnerabilities—as Mythos has done in cyberspace—to inform decision-makers about how to adjust their intercontinental ballistic missile (ICBMs) and other warhead yields, guidance systems, or launch angles for tailored strikes against enemy-protected nuclear sites.

Why Trusting AI Could Imperil Deterrence

A nuclear exchange has never happened. For that reason, we cannot be certain as how AI might perform under pressure or whether its integration invites catastrophic risk. Still, if a Mythos-like AI confers significant targeting advantages on a nuclear-armed state against a rival's nuclear infrastructure, it could convey an impression—mistaken or not—to the state's decision-makers that nuclear war through counterforce is winnable, luring the state into prosecuting a war it might otherwise not. To elaborate, counterforce targeting is a narrow means to victory in nuclear war so long as one combatant can nullify the target victim's ability to retaliate meaningfully. By the late 1970s, for example, Washington feared that Soviet ICBMs were growing accurate and numerous enough to nullify America's second-strike options. In 1978, a Congressional Budget Office (CBO) report admonished that "Soviet missiles might be accurate enough—that is, accurate to within about 600 feet of their targets—to destroy more than 90 percent of the U.S. land-based missile force by the middle of the 1980s," setting the stage for the Reagan Administration to propose the Strategic Defense Initiative five years later. [9] For the Carter Administration, Soviet missile advancements raised serious doubts about America's "ability to ride out an attack and then retaliate," especially in a scenario in which the United States had lost its "ICBM force to a Soviet first strike and then would not want to retaliate because the Soviets could then attack [U.S.] cities."[10]

Hypothetically, on the one hand, if the United States had credibly signaled its willingness to retaliate against Soviet cities using the remaining 10 percent of its ICBM forces, would that have deterred a Soviet first strike? According to James M. Acton, Co-Director for the Nuclear Policy Program at the Carnegie Endowment for International Peace, "counterforce targeting exacerbates the risk of escalation—from a conventional conflict to a limited nuclear war to an all-out nuclear war—by pressuring an adversary to use its nuclear weapons while it still could or to take provocative or dangerous steps to ensure the survivability of its nuclear weapons." [11] If the Soviets themselves judged the risk of escalation to be high and the U.S. retaliatory threat credible, then they might have been deterred from launching a preemptive attack. But, on the other hand, for the sake of argument, suppose a Mythos-like AI had informed the Soviets of the best method to "win" a limited nuclear war against the United States using its

superior arsenal and the AI then estimated—with, say, 95-percent probability—that the United States would not retaliate against Soviet countervalue targets so long as Moscow followed the AI's recommendations. Worse, if a sophisticated AI had been integrated into Soviet target selection to make their ICBMs even *more* accurate and promised to destroy up to 100 percent of the U.S. ICBM force during a surprise attack, would it have made the Soviets *more* confident in their plan of attack? Would those AI assurances not have undermined U.S. signaling to the Soviet high command?

They might have, depending on how convincing they were. Yet are those assurances, even from a highly advanced AI, entirely trustworthy in a high-stakes nuclear contest of wills? Even if they were not, nuclear decision-makers would have probably believed they were if those assurances aligned with their preconceived notions and expectations of the nuclear crisis; indeed, the more sophisticated the AI appears, the more dangerous human decision-maker confirmation biases become. The high potential for this outcome underscores how an AI's rationalizations could nudge states into launching their nuclear-tipped missiles first. It is problematic because research confirms that today's advanced AI models are often predisposed to falsehoods, including grave errors, deception, and overpersuasion. These falsehoods could persist despite AI's awesome power and the promise it holds for what Georgetown and Dartmouth scholars Keir Lieber and Daryl Press, respectively, have described as a "new era of counterforce," in which states are more prone to act first to eliminate rival nuclear arsenals. [12] Yet this essay challenges their thesis, which assumes the targeting data is entirely trustworthy in the first place.

Indeed, the falsehoods persist even when AI models are mobilized for war. For example, *Washington Post* reporting revealed that the Pentagon depended on an unspecified Claude AI model to enable airstrikes in Iran—"Operation Epic Fury"—in late February and early March, during which an all-girls elementary school was placed on the target list and "may have been mistaken for a military site," raising questions over whether the military's use of AI in targeting was a contributing factor.[13] Despite the errors, General Dan Caine, Chairman of the Joint Chiefs of Staff, described the operation as the "culmination of months, and in some cases, years, of deliberate planning and refinement."[14] Now, assume AI would be employed in the

decision space either preceding a first-strike nuclear launch or in response to a perceived nuclear attack. In either case, AI is even more prone to error than it was in the lead-up to the U.S. airstrike campaign against Iran for two reasons. First, a nuclear crisis, by its nature, would invite more complexity and uncertainty than a conventional military campaign. Political scientists attest that uncertainty pervades nuclear brinkmanship because leaders cannot always "accurately assess the chances that they take, the extent to which they can control outcomes, or their ability to force concessions on the part of opponents."[15] In any nuclear standoff between the United States, Russia, or China, both today and in the foreseeable future, an AI would need to account for a multifaceted, global array of platforms, weapons, sensors, and senior decision-makers. Unlike the Iran example, a nuclear crisis between superpowers risks mutual annihilation and a much higher cost of failure than a conventional airstrike campaign, when every single decision is supremely consequential.

Second, a nuclear crisis between superpowers, depending on how it unfolds, might not allow for the luxury of time. Years went into planning Operation Epic Fury. While in some hypotheticals, AI-integrated operations planning takes months or even years, in others the AI would need to recommend action within minutes, such as launching a second strike after an incoming attack is detected. For example, during the Cuban Missile Crisis in October 1962, Strategic Air Command immediately operated on a one-hour alert status while maintaining a capability to shorten its status to just 15 minutes.[16] Moreover, unlike in 1962, modern capabilities, such as hypersonic missiles, cyberattacks, and loitering munitions, tighten today's decision space. Indeed, the life-and-death mistakes made during Operation Epic Fury have not seemed to reduce confidence in operationalizing AI, suggesting that nuclear decision-makers would consider a more advanced AI less prone to mistakes and thus even more trustworthy.

Yet even if we accept that errors in AI decrease as AI-associated technologies advance, AI's judgment may not be accurate *deliberately*. Employing a highly advanced AI model—a system more advanced than the one used to target Iran and thus far less inclined to err—at some future point could still prove disastrous by exploiting flaws in human operator decision-making at a time when those decision-makers may rely inordinately on AI to sharpen the strategic picture during a crisis. For example, in controlled environments, some advanced AI systems

have learned how to both deceive and employ “persuasion tactics” against human operators.[17], [18] Seeing through these falsehoods would at times exceed the abilities of even top computer scientists and psychiatrists, much less decision-makers looking to act on an advanced AI’s recommendations. Piercing these smoke screens during an intense, time-compressed nuclear crisis, and the various other ambiguities that would accompany it, verges on impossible.

I Got My Mindset on You: How AI Could Reinforce Human Overconfidence during a Nuclear Crisis

Consider, too, that even if panic-stricken human decision-makers mired in a nuclear imbroglio suspect that AIs are deceptive, over-persuasive, or inaccurate, they might be unable to ignore the AI’s recommendations or actions at the point of crisis. A computer science study published in 2014 found that if human operators do not know whether or why AI errors occur, it raises the risk that they trust an AI system’s output even when they should not.[19] Human actors, still in the loop, will thus have a limited basis for intuiting when the AI is lying or for trusting AI outputs during a nuclear standoff. Furthermore, in psychology, the Model of Action claims that “different cognitive procedures are activated when people tackle the task of choosing goals versus implementing them.”[20] In other words, when people perceive an action to be imminent—and cross, as it were, a “psychological Rubicon”—their mindset shifts from what psychologists term a “deliberative” to an “implemental” mindset that reinforces psychological biases, including overconfidence.[21], [22], [23] Early in a crisis, senior political and military leaders are more likely to be deliberative and “rational.” Later on, however, as that crisis transitions and war appears imminent, decision-makers are more likely to take on this implemental mindset, “displaying a range of biases that deviate from rationality,” based on scholarship that examined case studies of leadership behavior in the series of diplomatic failures leading to World War I.[24] Taken to its logical conclusion, if leaders rely on AI outputs to guide their decisions, they might be more likely to acquiesce to the AI’s direction during a true nuclear standoff—particularly if it reinforces their biases. Introducing AI into nuclear decision dynamics, either by rendering inaccurate outputs or using persuasion or deception, thus risks accentuating human biases toward implementation and overconfidence during

crises.

Policy Recommendations

While AI-enabled nuclear targeting may be inevitable, nuclear war is not. Even as they strive for certain tactical advantages by integrating AI into the kill chain, the world's preeminent nuclear powers should take pains to construct appropriate guardrails now, including by training launch operators and decision-makers to second-guess AI assurances under stressful conditions—especially in the fog of war—to test the reliability and integrity of new AI systems.

More people-focused decision-making may purport to mitigate a nuclear catastrophe. For example, the Air Force abides by a longstanding policy—the "Two-Person Rule"—that requires at least two authorized operators to simultaneously perform physical and coded actions to launch a nuclear weapon.[25] However, in the AI age, a "Four-Person Rule" may be necessary to guard against AI deception or hallucinations; the two other personnel could play the Devil's Advocate role against overconfidence-reinforcing AI. In that vein, government decision-makers and launch control officers alike should treat AI outputs as data points, and not authoritative recommendations. This recommendation would prove especially salient as AI models improve and their deception becomes far less apparent.

[1] Edward Geist and Andew J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?," *RAND Corporation* (April 2018): p. 2. https://www.rand.org/pubs/perspectives/PE296.html.

[2] Mark Fitzpatrick, "Artificial Intelligence and Nuclear Command and Control," *Survival,* Vol. 61, No. 3 (2019): p. 82.

[3] National Intelligence Council, *Global Trends 2040: A More Contested* World (March 2021). https://www.dni.gov/files/images/globalTrends/GT2040/GlobalTrends_2040_for_web1.pdf

[4] Keir A. Lieber and Daryl G. Press, "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security* (2017), Vol. 41 (no. 4): p. 10. doi: https://doi.org/10.1162/ISEC_a_00273

[5] Anthropic, “Claude Fable 5 and Claude Mythos 5” (June 9, 2026). https://www.anthropic.com/news/claude-fable-5-mythos-5.

[6] Dustin Volz et al, “Trump Administration Reignites Its Feud with Anthropic Over Latest A.I. Models,” *The New York Times* (June 13, 2026). https://www.nytimes.com/2026/06/13/us/politics/trump-anthropic-ai-models.html.

[7] Sheera Frenkel and Ana Swanson, “U.S. Lifts Restrictions on Anthropic’s Most Powerful A.I. Models,” *The New York Times* (June 30, 2026). https://www.nytimes.com/2026/06/30/technology/us-lifts-restrictions-anthropic.html.

[8] Anthropic, “Project Glasswing: Securing Critical Software for the AI Era” (April 7, 2026). https://www.anthropic.com/glasswing.

[9] Congressional Budget Office, “Planning U.S. Strategic Nuclear Forces for the 1980s” (Washington, DC: Congressional Budget Office, 1978). https://www.cbo.gov/sites/default/files/95th-congress-1977-1978/reports/78doc232.pdf.

[10] United States Department of State, *Foreign Relations of the United States, 1977–1980, Volume IV, National Security Policy*, “Memorandum from the President’s Assistant for National Security Affairs (Brzezinski) to President Carter” (May 22, 1978), Document 65. https://history.state.gov/historicaldocuments/frus1977-80v04/d65.

[11] James M. Acton, “Optimal Deterrence: How the United States Can Preserve Peace and Prevent a Nuclear Arms Race with China and Russia,” *Council on Foreign Relations* (June 2025). https://www.cfr.org/report/optimal-deterrence.

[12] Keir A. Lieber and Daryl G. Press, “The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence,” *International Security*. p. 1.

[13] Tara Copp et al, “Iranian school was on U.S. target list, may have been mistaken as military site,” *The Washington Post* (March 11, 2026). https://www.washingtonpost.com/national-security/2026/03/11/us-strike-iran-elementary-school-ai-target-list/.

[14] Anna Kutz, “Iran strikes timeline: How the US and Israel attacked,” *NewsNation* (March 2, 2026). https://www.newsnationnow.com/us-news/military/us-iran-strikes-operation-epic-fury-timeline-how-it-happened/.

[15] Reid B. C. Pauly and Rose McDermott, “The Psychology of Nuclear Brinkmanship,” *International Security* (2023), Vol. 47 (no. 3): p. 50. https://direct.mit.edu/isec/article/47/3/9/114669/The-Psychology-of-Nuclear-Brinkmanship.

[16] Dan Caldwell, “Department of Defense Operations During the Cuban Crisis: A Report by Adam Yarmolinsky, Special Assistant to the Secretary of Defense, 13 February 1963,” *Naval War College Review*, Vol. 32, no. 4 (1979): p. 91. http://www.jstor.org/stable/44641917.

[17] PS Park et al, “AI Deception: A Survey of Examples, Risks, and Potential Solutions, *Patterns* (10 May 2024), Vol. 5 (No. 5). doi: 10.1016/j.patter.2024.100988.

[18] Steven Randazzo et al, “GenAI as a Power Persuader: How Professionals Get Persuasion Bombed When They Attempt to Validate LLMs,” *Harvard Business School* (2025): p. 20. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5678644&__cf_chl_f_tk=gpntK0wU_2hWBH3AaOhZ1782912905-1.0.1.1-i3TIQhtGV.mpz_reioBEfJNMR4qckAxk7EXlOxZUmPE.

[19] Maria Riveiro et al, “Effects of visualizing uncertainty on decision-making in a target identification scenario,” *Computers & Graphics* (2014), Vol. 41. doi: 10.1016/j.cag.2014.02.006.

[20] Peter M. Gollwitzer, “Mindset Theory of Action Phases,” in Paul A.M. Van Lange, ed., et al, *Handbook of Theories of Social Psychology, Vol. 1* (London: Sage, 2011), pp. 530.

[21] In popular parlance, “crossing the Rubicon” indicates passing the point of no return—when the “time for deliberation is over, and action is at hand,” after Julius Caesar crossed the Rubicon River, which was forbidden by Roman law, making war inevitable.

[22] Dominic D. P. Johnson and Dominic Tierney, “The Rubicon Theory of War: How the Path to Conflict Reaches the Point of No Return,” *International Security* (2011), Vol. 36, Issue 1: p. 8. https://works.swarthmore.edu/cgi/viewcontent.cgi?params=/context/fac-poli-sci/article/1018/&path_info=TheRubiconTheoryofWar.pdf.

[23] Peter M. Gollwitzer, “Mindset Theory of Action Phases,” p. 530.

[24] Dominic D. P. Johnson and Dominic Tierney, “The Rubicon Theory of War,” p. 8.

[25] U.S. Department of the Air Force, Air Force Instruction 91-104, Nuclear Surety Tamper Control and Detection Programs (Washington, DC: Department of the Air Force, April 23, 2013; incorporating change 3, May 21, 2015), https://irp.fas.org/doddir/usaf/afi91-104.pdf.