World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

The black box paradox: How AI integration inverts nuclear deterrence and creates universal vulnerability

Donatien Sakubu *

ICT Independent Researcher.

World Journal of Advanced Research and Reviews, 2025, 28(02), 1003-1007

Publication history: Received on 16 September 2025; revised on 08 November 2025; accepted on 11 November 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.2.3818

Abstract

The 21st-century pursuit of strategic advantage has shifted from megatonnage to milliseconds, with great powers investing heavily in Artificial Intelligence (AI) to modernize their nuclear command, control, and communications (NC3) systems. The prevailing logic assumes that AI-driven speed and autonomy will enhance deterrence and solidify superpower status. This paper argues that this assumption is a strategic fallacy. Rather than enhancing security, AI integration inverts the logic of nuclear deterrence by creating The Black Box Paradox. It replaces the slow, rational, and mostly stable logic of "Mutual Assured Destruction (MAD)" with a fast, brittle, and opaque system prone to catastrophic failure. This paper deconstructs the fallacy, arguing that AI integration creates Mutual Assured Vulnerability (MAV) by introducing unmanageable risks from hacking, data-poisoning, and black box errors. This new security dilemma traps nations in a 21st-century prisoner's dilemma, where the rational pursuit of individual security leads to collective, assured ruin. The paper concludes that true 21st-century power is not defined by a zero-sum arms race but by positive-sum cooperation, geoeconomic resilience, and a collective exit from this self-defeating logic.

Keywords: Artificial Intelligence; Nuclear Deterrence; Strategic Instability; Security Dilemma; Prisoner's Dilemma; Geoeconomic Resilience; Arms Control

1. Introduction

The great power competition of the 21st century is defined by a technological arms race to achieve AI-driven strategic advantages [1], [2], [3]. Countries are actively exploring the integration of AI into their nuclear command, control, and communications (NC3) systems, hoping to shorten decision-making timeframes, filter data more efficiently, and enhance the survivability of their nuclear forces [4], [5], [6]. The main goal is to win a future conflict by making a better, more accurate and faster decision than the opponents who rely heavily on humans with less sophisticated AI systems.

This pursuit, however, is based on a fundamental misunderstanding of both AI and strategic stability. The 20th-century logic of Mutual Assured Destruction (MAD), as framed by canonical theorists from Schelling to Jervis was stable precisely because it was slow, cumbersome, and mediated by fallible, but context-aware, human judgment. Leaders had time, sometimes even hours, to think, to assess warnings, communicate with others, and most importantly, avoid launching nuclear weapons [7]. This "human-in-the-loop" was not a bug, but the central feature of strategic stability [8].

This paper utilizes an analytical-conceptual approach to challenge the assumption that AI integration is a simple upgrade. Our logic of inference draws from key historical precedents and analogies such as the Stuxnet breach and the 1983 Stanislav Petrov incident to model the likely failure points of future AI-NC3 interactions, arguing that these are not isolated risks but features of a new, unstable strategic system [9]. Specifically, these precedents serve as concrete

* Corresponding author: Donatien Sakubu

validation points for the core tenets of MAV: that system opacity (the Petrov incident) and digital interconnectedness (Stuxnet) create vulnerabilities that override traditional deterrence logic [10].

Proponents argue that AI will reduce human error, filtering out the noise from vast data streams and overcoming the fallibility of a tired or panicky operator, thus making deterrence more credible [2], [11]. While this is the intended goal, this paper argues that it dangerously swaps one type of risk (human fallibility) for another, far more catastrophic one: opaque, high-speed, and scalable machine error [4], [5]. While existing analyses from RAND and SIPRI identify these risks [1], [5], [6], this paper synthesizes them into a novel theoretical framework: "Mutual Assured Vulnerability" (MAV). Where MAD's stability was built on the certainty of retaliatory destruction, MAV posits a new instability built on the uncertainty of systemic failure. This framework argues that the integration of AI does not just add risk, but fundamentally inverts the logic of deterrence itself.

AI, particularly current-generation machine learning, does not think or understand; it performs complex statistical pattern recognition. This creates a critical flaw: AI systems are "brittle" [5], [12]. They perform exceptionally well within their training data but can fail catastrophically and unpredictably when faced with novel, out-of-distribution events, the very definition of a brewing nuclear crisis. This paper argues that by integrating these brittle systems into our nuclear infrastructure, we are not creating superpowers; we are creating fragile glass cannons that make everyone, including their owners, less safe.

2. Mutual Assured Vulnerability (MAV): The New Attack Surface

The primary threat of AI-integrated NC3 is not a Terminator-style AI deciding to start a war. The more plausible and immediate danger is that these systems create a new, shared, and unmanageable vulnerability for all.

2.1. Hacking, Spoofing, and Data Poisoning: From Theory to Reality

Traditional nuclear systems were secure because they were air-gapped and analog. AI, by contrast, is a digital, data-hungry system, creating a vast new attack surface [4], [5], [12]. The notion that high-security systems are invulnerable is a fallacy, as evidenced by the Stuxnet worm, which successfully infiltrated and sabotaged an "air-gapped" Iranian nuclear facility [10].

In an AI-driven NC3, an adversary no longer needs to penetrate a silo; they can instead:

- **Spoof:** Feed the AI sensor systems, (e.g., satellites, radars) with false data that convincingly mimics an incoming attack. This is an adversarial example, a well-documented vulnerability in AI where, for instance, a self-driving car's image recognition can be fooled by a few strategically placed stickers on a stop sign [12], [13]. An NC3 system could be similarly tricked by a sophisticated adversary.
- **Poison:** Subtly corrupt the AI's training data over months or years, creating a hidden vulnerability or bias that will only manifest at the most critical moment [5], [12], [14].

This creates a state of interdependent vulnerability. A nation's security is no longer just dependent on its own defenses; it is now intrinsically linked to the cybersecurity flaws of its worst adversary.

2.2. The Black Box and the Dangers of Shortened Timeframes

Many advanced AI models are black boxes, meaning even their creators cannot fully explain why they reached a specific conclusion [11], [14], [15]. Proponents of AI integration often point to the emerging field of "Explainable AI" (XAI) as a solution, but this optimism is misplaced in the context of NC3. XAI methods, at their current stage, provide post-hoc approximations or high-level summaries of a model's decision, not a verifiable, logical proof of why it is correct [12], [15]. In a nuclear crisis, an "explanation" that is 99% accurate is insufficient; the 1% error could be catastrophic. Crucially, these limitations are not merely about a lack of interpretability; XAI, as a methodology, is not designed to and cannot mitigate the underlying systemic vulnerabilities of data poisoning or adversarial attacks [5], [12]. An explanation of a decision is useless if the decision itself is based on cleverly corrupted data.

This opacity is antithetical to strategic stability. The 1983 incident involving Soviet officer Stanislav Petrov is the quintessential example. Soviet early-warning satellites registered five incoming U.S. missiles. The system's protocol demanded a full-scale counter-launch. Petrov, however, using human intuition and contextual knowledge, identified it as a false alarm, a "brittle" system error (sunlight glinting off clouds) that a human overrode [9].

Now, consider a shortened decision-timeframe [1], [16] where a hybrid system is in place. A human operator is presented with a "launch" recommendation by an infallible-seeming black box, based on data too vast to be cross-referenced, and given only 90 seconds to confirm. This human-in-the-loop becomes a human-in-the-loophole, a rubber stamp for a potentially catastrophic machine error [2], [6], [11]. The 2010 financial Flash Crash, where high-frequency trading algorithms triggered a trillion-dollar market crash in minutes, serves as a stark, real-world precedent for how high-speed, autonomous systems can create cascading failures faster than any human can react [2], [17].

3. The Philosophical Trap: The 21st-Century Prisoner's Dilemma

The political leaders and military strategists pursuing this path are not irrational. On the contrary, they are acting rationally within a deeply flawed, obsolete system of logic. This is a classic Prisoner's Dilemma, a concept from game theory that explains why two rational actors, acting in their own self-interest, can end up with a mutually disastrous outcome [18].

- **The Game:** The current geopolitical situation mirrors the classic dilemma. In this context, cooperation means a mutual agreement to ban or strictly limit AI integration in NC3 systems, preserving strategic stability. "Defection" means secretly or openly developing these systems to gain a perceived first-strike or defensive advantage.
- **The Rational Choice:** Any single nation fears that its adversary will defect and achieve AI-driven nuclear speed first, leaving it vulnerable. Therefore, the rational and safe choice for that individual nation is to also defect from cooperation and race to build the AI weapon itself.
- **The Trap:** But when all nations make this same rational choice, they all defect. They collectively create the worst possible outcome: a brittle, hackable, automatic doomsday machine that no one controls [1], [5]. They are all less safe than if they had all cooperated.

This paradox is driven by an outdated, zero-sum logic of international relations. This mindset, as described by security dilemma theorists, assumes that one nation's gain in security must come at the expense of another's [7]. The AI arms race is the epitome of this zero-sum thinking.

The core fallacy is applying this logic to a shared existential risk. An AI-driven nuclear catastrophe is not a win for anyone; it is a common fate that makes traditional notions of national sovereignty illusory. This is not a game that can be won, but only lost by everyone. The rational choice to defect is, in reality, a collective and irrational march toward mutual destruction [19].

This paper argues that the only rational path forward is to fundamentally shift the logic from a zero-sum competition to a positive-sum cooperation [18]. In an age of shared existential risk, cooperation is not a concession; it is the only pragmatic survival strategy. Those who believe they are winning this race are, in fact, wrong in the most profound sense. They are participating in a negative-sum game where the only winner is the catastrophic failure of the system itself.

4. The Solution: Redefining Superpower from Geopolitics to Geoeconomics

The belief that a new AI weapon will grant superpower status is a dangerous 20th-century anachronism. It reveals a profound opportunity cost. The trillions of dollars, and the intellectual capital of an entire generation of engineers, being poured into this zero-sum race are resources not being spent on the true drivers of 21st-century power:

- Economic resilience and sustainable infrastructure.
- Public health and pandemic preparedness.
- Scientific and technological innovation for public good.
- Global soft power and cultural influence.

A nation that wins the AI arms race but has a crumbling economy, a failing health system, and no global partners is not a superpower; it is a failed state with a doomsday button.

The only way to win this new game is to refuse to play. True strength lies in recognizing this Mutual Assured Vulnerability (MAV) and leading a new global effort for positive-sum cooperation. This requires:

New International Treaties: A global, verifiable ban on AI in NC3 systems, mandating meaningful human control over all nuclear launch decisions. Existing arms control frameworks, such as New START, are insufficient as they are designed

to count warheads and launchers, not to govern opaque, dual-use software or algorithms. Such a treaty would face significant political and technological hurdles, which this paper acknowledges are substantial. These include the immense difficulty of defining and verifying 'AI' in a software context, monitoring black box algorithms that are opaque by nature, and navigating the dual-use problem, where civilian AI research is often indistinguishable from military application [6], [11]. These verification challenges are far more complex than counting physical warheads and represent a major, unresolved challenge for 21st-century diplomacy.

- **Bilateral Safety Cooperation:** Red lines and transparency agreements between nuclear-armed countries to share information on AI safety protocols and prevent accidental escalation [6], [16].
- **A National Re-investment:** A conscious policy choice to shift resources from the brittle power of autonomous weapons to the resilient power of a strong economy, an educated populace, and technological leadership in non-military domains.

5. Conclusion

The integration of AI into nuclear weapons systems is not the next logical step in deterrence; it is the end of it. It creates the Black Box Paradox: in the pursuit of ultimate security, we are creating absolute, universal vulnerability. We are building systems that no one understands, no one can control, and no one can secure from hackers or errors.

This research paper posits that the nations and leaders currently competing in this race are trapped in an obsolete, zero-sum logic. As the Prisoner's Dilemma framework illustrates, to see oneself as a winner in this race is to be wrong about the very definition of power and security.

While future research into Explainable AI (XAI) aims to solve the black box problem, this technology is far from mature [12], [14], [15]. As argued in Section 2.2, XAI offers interpretation, not proof, and is wholly inadequate for the high-stakes, adversarial environment of nuclear command and control. Relying on a hypothetical future solution to justify a current, existential risk is strategically reckless. The pace of vulnerability discovery and new hacking methods is far outpacing the progress of AI safety, making the problem worse, not better, in the near term.

It is important to acknowledge the methodological scope of this article. As a conceptual-analytical paper, it puts forward the theoretical framework of Mutual Assured Vulnerability, supported by historical analogy and documented technical vulnerabilities. It does not, however, provide quantitative modeling or empirical simulations of this framework. Such work represents a critical next step. Future research should be directed at empirically testing the MAV hypothesis, perhaps through wargaming simulations or technical-forecasting models, to more precisely map the failure points and escalation pathways this paper identifies.

True safety and supremacy in the 21st century will belong not to the nation that builds the fastest weapon, but to the nation wise enough to lead the world in not building it. It will belong to those who understand that, in an age of shared existential risk, cooperation is not weakness; it is the only rational strategy for survival.

Compliance with ethical standards

Disclosure of conflict of interest

The author declares that there are no conflicts of interest regarding the publication of this paper. This research did not receive any specific funding.

References

[1] V. Boulanin, L. Saalman, P. Topychkanov, F. Su, and M. P. Carlsson, “ARTIFICIAL INTELLIGENCE, STRATEGIC STABILITY AND NUCLEAR RISK,” Jun. 2020. Accessed: Nov. 10, 2025. [Online]. Available: https://www.sipri.org/publications/2020/policy-reports/artificial-intelligence-strategic-stability-and-nuclear-risk

[2] P. Scharre, “ARMY OF NONE Autonomous Weapons and the Future of War,” 2018. Accessed: Nov. 10, 2025. [Online]. Available: https://www.paulscharre.com/army-of-none

[3] M. Mennesson and M. Harjani, "AI, NC3 and the Future of Strategic Stability in the Trump 2.0 Era," Oct. 2025. Accessed: Nov. 10, 2025. [Online]. Available: https://rsis.edu.sg/rsis-publication/rsis/ip25101-ai-nc3-and-the-future-of-strategic-stability-in-the-trump-2-0-era/

[4] J. John and N. T. Shanahan, "Artificial Intelligence and Nuclear Command and Control: It's Even More Complicated Than You Think | Arms Control Association," Sep. 2025. Accessed: Nov. 10, 2025. [Online]. Available: https://www.armscontrol.org/act/2025-09/features/artificial-intelligence-and-nuclear-command-and-control-its-even-more

[5] E. Geist and A. J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?," 2018. Accessed: Nov. 10, 2025. [Online]. Available: https://www.rand.org/pubs/perspectives/PE296.html.

[6] Future of Life Institute, "Risks of Artificial Intelligence in Nuclear Command, Control and Communications (NC3). Primer & Policy Options for Risk Mitigation," Jul. 2023. Accessed: Nov. 10, 2025. [Online]. Available: https://futureoflife.org/wp-content/uploads/2023/07/FLI_AI_NC3_Policy_Primer.pdf

[7] R. Jervis, "Cooperation Under the Security Dilemma," 1978. Accessed: Nov. 10, 2025. [Online]. Available: https://www.jstor.org/stable/2009958

[8] O. K. Oladele, "The Ethics of Autonomous Weapons Systems: AI in Warfare and Global Governance," 2024. [Online]. Available: https://www.researchgate.net/publication/390033169

[9] P. Stanislav, "The Unsung Hero Who Saved the World from Nuclear Annihilation," Sep. 1983, Accessed: Nov. 10, 2025. [Online]. Available: https://www.historytools.org/stories/stanislav-petrov-the-unsung-hero-who-saved-the-world-from-nuclear-annihilation

[10] D. Kushner, "The Real Story of Stuxnet - IEEE Spectrum. How Kaspersky Lab tracked down the malware that stymied Iran's nuclear-fuel enrichment program." Accessed: Nov. 10, 2025. [Online]. Available: https://spectrum.ieee.org/the-real-story-of-stuxnet

[11] H. Lin, "Artificial Intelligence and Nuclear Weapons: A Commonsense Approach to Understanding Costs and Benefits," 2025. Accessed: Nov. 10, 2025. [Online]. Available: https://cisac.fsi.stanford.edu/publication/artificial-intelligence-and-nuclear-weapons-commonsense-approach-understanding-costs

[12] M. Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," Feb. 2018, doi: https://doi.org/10.48550/arXiv.1802.07228.

[13] K. Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," 2018. Accessed: Nov. 10, 2025. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00175

[14] A. Olvera, "Why nobody can see inside AI's black box," 2025. [Online]. Available: https://thebulletin.org/2025/01/why-nobody-can-see-inside-ais-black-box/

[15] W. J. von Eschenbach, "Transparency and the Black Box Problem: Why We Do Not Trust AI," Philos Technol, vol. 34, no. 4, pp. 1607–1622, Dec. 2021, doi: 10.1007/s13347-021-00477-0.

[16] D. Melnikov, "SIPRI on artificial intelligence in nuclear systems." Accessed: Nov. 10, 2025. [Online]. Available: https://en.topwar.ru/272893-sipri-ob-iskusstvennom-intellekte-v-jadernyh-sistemah.html

[17] CFTC and SEC, "FINDINGS REGARDING THE MARKET EVENTS OF REPORT OF THE STAFFS OF THE CFTC AND SEC TO THE JOINT ADVISORY COMMITTEE ON EMERGING REGULATORY ISSUES Market Event Findings," 2010. Accessed: Nov. 10, 2025. [Online]. Available: https://www.sec.gov/news/studies/2010/marketevents-report.pdf

[18] R. Axelrod, "The Evolution of Cooperation," 1984. Accessed: Nov. 10, 2025. [Online]. Available: https://monoskop.org/images/b/b8/Axelrod_Robert_The_Evolution_of_Cooperation.pdf

[19] S. Goldstein and P. N. Salib, "Nuclear Deterrence in the Age of AGI," 2025. [Online]. Available: https://www.lawfaremedia.org/article/nuclear-deterrence-in-the-age-of-agi