

GC REAIM Expert Policy Note Series

Artificial Intelligence and Nuclear Stability: Understanding AI's Impact on Military Escalation Dynamics and Strategic Deterrence

James Johnson

April 2025

POWERED BY



The Hague Centre
for Strategic Studies

GC REAIM Expert Policy Note Series

Artificial Intelligence and Nuclear Stability: Understanding AI's Impact on Military Escalation Dynamics and Strategic Deterrence

Author: James Johnson

April 2025

Cover photo: [Unsplash](#)

The Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) is an initiative of the Government of the Netherlands that was launched during the 2023 REAIM Summit on Responsible Artificial Intelligence in the Military Domain in The Hague. Upon request of the Dutch Ministry of Foreign Affairs, the Hague Centre for Strategic Studies acts as the Secretariat of the Commission.

The GC REAIM Expert Policy Note Series was funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom. GC REAIM Experts maintained full discretion over the topics covered by the Policy Notes. The contents of the GC REAIM Expert Policy Note series do not represent the views of the Global Commission as a whole. The Policy Notes are intended to highlight key issues related to the governance of AI in the military domain and provide policy recommendations.

© The Hague Centre for Strategic Studies. All rights reserved. No part of this report may be reproduced and/ or published in any form by print, photo print, microfilm or any other means without prior written permission from HCSS. All images are subject to the licenses of their respective owners

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
[@hcssnl](#)

The Hague Centre for Strategic Studies
Email: info@hcss.nl
Website: www.hcss.nl



1. Introduction

The stability-instability paradox has historically illustrated the dual effects of nuclear deterrence—preventing large-scale wars while simultaneously allowing low-level conflicts to persist.¹ Today, integrating artificial intelligence (AI) in military systems has intensified these dynamics. AI technologies, capable of compressing decision-making timelines and automating critical processes, present profound risks in high-stakes geopolitical contexts.² This policy note explores how AI exacerbates escalation risks under the nuclear shadow and outlines a comprehensive set of recommendations to mitigate these threats.³ By combining theoretical insights and plausible scenarios, the policy note highlights the urgent need for new governance and safety frameworks.

AI's potential to transform warfare stems from its ability to enhance operational efficiency, reduce human error, and enable faster responses in dynamic environments. However, these attributes introduce destabilizing factors in strategic contexts, especially when AI is applied to nuclear command, control, and communication (NC3) systems.⁴ The delicate balance of deterrence that underpins international security becomes increasingly precarious as nations integrate opaque and unpredictable AI systems into their military arsenals.⁵ Without appropriate safeguards, the unintended consequences of these integrations could lead to catastrophic escalation, making risk mitigation an essential focus for policymakers.

¹ Glenn Herald Snyder, *The Balance of Power and the Balance of Terror* (Chandler, 1965).

² See: James Johnson, 'Artificial Intelligence in Nuclear Warfare: A Perfect Storm of Instability?', *The Washington Quarterly* 43, no. 2 (2 April 2020): 197–211, <https://doi.org/10.1080/0163660X.2020.1770968>; James Johnson, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford, New York: Oxford University Press, 2023); Kenneth Payne, 'Artificial Intelligence: A Revolution in Strategic Affairs?', *Survival* 60, no. 5 (3 September 2018): 7–32, <https://doi.org/10.1080/00396338.2018.1518374>; Vincent Boulanin, ed., 'The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume I, Euro-Atlantic Perspectives' (SIPRI, May 2019), <https://www.sipri.org/publications/2019/research-reports/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic>; Edward Geist and Andrew J. Lohn, 'How Might Artificial Intelligence Affect the Risk of Nuclear War?' (RAND Corporation, 24 April 2018), <https://www.rand.org/pubs/perspectives/PE296.html>.

³ For a general primer on the type of AI capabilities that could be developed and how AI might influence warfighting, see: James Johnson, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford, New York: Oxford University Press, 2023).

⁴ James Johnson, *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford, New York: Oxford University Press, 2023).

⁵ The US DoD, in its 2022 Nuclear Posture Review, stated that it "will employ an optimized mix of resilience to protect the next-generation NC3 architecture from posed by competitor capabilities. This includes...enhanced protection from cyber, space-based, and electromagnetic pulse threats, enhanced integrated tactical warning and attack assessment, improved command post and communication links, advanced [including AI-enhanced] decision support, and integrated planning and operations" United States Department of Defense. *2022 National Defense Strategy, Nuclear Posture Review, and Missile Defense Review*. 27 October 2022. <https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.pdf>.

2. AI and Escalation Risks in Warfare

The emergence of AI in military and nuclear contexts significantly alters the strategic stability established under traditional deterrence frameworks.⁶ AI-enabled systems can compress operational tempos, reducing the time available for human decision-makers to deliberate during crises.⁷ For instance, autonomous platforms such as drones or AI-assisted targeting systems may misinterpret routine actions as hostile, prompting a cycle of escalating countermeasures. This potential for rapid escalation is compounded by AI algorithms' opacity, making it difficult for operators to assess the reasoning behind their recommendations.⁸

Moreover, AI systems often need more contextual understanding to interpret complex human behaviours accurately. This shortfall becomes critical when adversarial actions are ambiguous or cultural and political nuances play a significant role in decision-making. Human operators may either over-rely on AI outputs or misinterpret their implications, further amplifying the risks of miscalculation.⁹

Fictional scenarios illustrate these risks vividly:

2.1 Scenario 1: The "2030 Flash War" in the Taiwan Strait

The hypothetical "2030 Flash War" scenario illustrates how a small-scale incident can escalate rapidly under the influence of AI-driven decision-making systems.¹⁰ In this scenario, geopolitical tensions between China and Taiwan reach a breaking point when a Taiwanese patrol boat collides with a Chinese autonomous maritime vehicle during a routine navigation exercise. While traditionally, such incidents might lead to diplomatic fallout or limited military posturing, the integration of AI systems on both sides fundamentally alters the trajectory of events.

China's AI-enabled command systems, interpreting the collision as a deliberate provocation, recommend a series of escalating countermeasures. These include deploying unmanned aerial vehicles (UAVs) to survey Taiwanese naval operations and mobilizing missile defence systems. Simultaneously, Taiwan's AI systems perceive these

⁶ James Johnson, 'Artificial Intelligence in Nuclear Warfare: A Perfect Storm of Instability?', *The Washington Quarterly* 43, no. 2 (2 April 2020): 197–211, <https://doi.org/10.1080/0163660X.2020.1770968>.

⁷ Todd S. Sechser, Narang, Neil, and Caitlin and Talmadge, 'Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War', *Journal of Strategic Studies* 42, no. 6 (19 September 2019): 727–35, <https://doi.org/10.1080/01402390.2019.1626725>.

⁸ UK House of Lords, 'Proceed with Caution: Artificial Intelligence in Weapon Systems - AI in Weapon Systems Committee', 2023, <https://publications.parliament.uk/pa/ld5804/ldselect/ldaiwe/16/1602.htm>.

⁹ James Johnson and James Johnson, *The AI Commander: Centaur Teaming, Command, and Ethical Dilemmas* (Oxford, New York: Oxford University Press, 2024).

¹⁰ This scenario is adapted from sections of James Johnson, 'AI, Autonomy, and the Risk of Nuclear War', *War on the Rocks*, 29 July 2022, <https://warontherocks.com/2022/07/ai-autonomy-and-the-risk-of-nuclear-war/>.

actions as aggressive, triggering the automatic deployment of coastal missile batteries. The situation spirals further as AI algorithms on both sides amplify perceived threats, misinterpreting routine military manoeuvres as a prelude to an attack.

The compressed decision-making timelines introduced by AI exacerbate the risks. Human operators, overwhelmed by the speed and volume of AI-generated recommendations, struggle to assert control over the unfolding crisis. Diplomatic channels, which might have otherwise de-escalated the situation, are sidelined as both nations' systems prioritize rapid response over measured deliberation. Within 48 hours, the conflict escalates to a limited nuclear exchange, resulting in catastrophic loss of life and irreversible geopolitical consequences.¹¹

This scenario underscores the critical risks posed by AI in military contexts. It highlights how the opacity of AI systems—their inability to explain the rationale behind their decisions—can lead to misinterpretations and overreactions. Furthermore, it demonstrates the potential for cascading failures, where one misstep triggers a chain reaction of increasingly aggressive responses. The "2030 Flash War" is a cautionary tale, emphasizing the urgent need for robust oversight mechanisms, transparent AI systems, and international collaboration to prevent similar outcomes.

2.2 Scenario 2: Operation Island Freedom

The second scenario, "Operation Island Freedom," envisions a full-scale military conflict over Taiwan between China and the United States.¹² Unlike the "2030 Flash War," which begins with an isolated incident, this scenario unfolds within a premeditated invasion. China, seeking to assert control over Taiwan, deploys semi-autonomous drones and AI-enabled logistics systems to coordinate a rapid and overwhelming assault. These systems are designed to minimize human intervention, enabling high-speed decision-making and execution.

The United States, honouring its defence commitments to Taiwan, responds by mobilizing its AI-enabled assets, including autonomous submarines and cyber-defence systems. As the conflict escalates, both nations rely heavily on their AI systems to interpret the adversary's intentions and formulate counterstrategies. However, the inherent limitations of these systems become glaringly apparent.

One critical moment occurs when a U.S. naval vessel's AI system misinterprets a Chinese drone swarm's reconnaissance mission as an imminent attack. Acting on this assessment, the system launches a pre-emptive strike, destroying several Chinese assets. China's AI systems, interpreting the strike as a deliberate escalation, recommend retaliatory measures, including a limited nuclear demonstration to deter further U.S.

¹¹ In a recent high-level future Taiwan conflict wargame hosted by the Center for a New American Security (CNAS), it was found that unintended escalation could quickly spiral out of control as both sides cross red lines that the other side was unaware of. Moreover, despite Beijing's no-first-use policy, China, in a Taiwan military crisis, may nonetheless conduct limited nuclear demonstrations to deter US involvement or to achieve escalation dominance: Stacie Pettyjohn, Becca Wasser, and Chris Dougherty, 'Dangerous Straits: Wargaming a Future Conflict over Taiwan', CNAS, 2022, <https://www.cnas.org/publications/reports/dangerous-straits-wargaming-a-future-conflict-over-taiwans>.

¹² This scenario is adapted from sections of James Johnson, 'The Challenges of AI Command and Control', 2023, <https://europeanleadershipnetwork.org/commentary/the-challenges-of-ai-command-and-control/>.

aggression. This decision is executed before human operators fully comprehend the situation, pushing the conflict into uncharted territory.

"Operation Island Freedom" highlights the dangers of over-reliance on AI systems in complex, high-stakes environments. The rapid tempo of AI-enabled operations leaves little room for human judgment, increasing the likelihood of errors and miscalculations. Additionally, the opacity of AI algorithms prevents meaningful accountability as operators struggle to determine whether the systems' actions were justified or avoidable. This scenario underscores the need for fail-safe mechanisms and human oversight, ensuring that critical decisions remain within the realm of human judgment.

Both scenarios reveal the profound risks associated with AI in military contexts. They demonstrate how these systems can exacerbate existing tensions, amplify misunderstandings, and accelerate the pace of escalation beyond human control. By examining these hypothetical conflicts, policymakers can better understand the potential pitfalls of AI-enabled warfare and develop strategies to mitigate these risks.

To reduce the risk of AI-driven miscalculations in both scenarios, it is essential to implement robust measures such as human oversight, transparency, and fail-safe protocols. Establishing international protocols for human oversight would ensure that critical decisions, particularly those involving military escalation, are always subject to human verification. Creating transparency in AI decision-making processes can foster trust and accountability, enabling clearer communication during crises. Furthermore, incorporating fail-safe mechanisms, requiring multiple human confirmations before initiating high-stakes military actions, and conducting joint military drills involving AI systems can help mitigate the risks of automatic escalation, ensuring that AI-enhanced decisions are always tempered by responsible human judgment and oversight.

The use of fictional scenarios in policy analysis serves several critical functions. By creating controlled, imaginative simulations grounded in plausible trends, these narratives help illuminate potential pathways to conflict that might not be immediately obvious through traditional analyses.¹³ Fictional scenarios encourage policymakers to confront the interplay of technology, human behaviour, and strategic ambiguity under high-pressure conditions.¹⁴ Furthermore, these narratives explore cascading failures and unintended consequences, highlighting the need for adaptive and robust crisis management frameworks. For example, a well-crafted scenario might reveal how an AI-enabled decision-support system could escalate tensions by misinterpreting a routine troop movement as a hostile action, prompting an aggressive counter-response.

¹³ James Johnson, 'Revisiting the "Stability-Instability Paradox" in AI-Enabled Warfare: A Modern-Day Promethean Tragedy under the Nuclear Shadow?', *Review of International Studies*, 20 November 2024, 1–19, <https://doi.org/10.1017/S0260210524000767>.

¹⁴ See: August Cole and Peter Singer, 'Invisible Force: Information Warfare and the Future of Conflict', 2020, <https://hdl.handle.net/20.500.14216/524>; August Cole and P. W. Singer. "Thinking the Unthinkable with Useful Fiction." *Queen's University Psychology Graduate Online Journal*, no. 2 (September 2020): 1–13; Mark Jacobsen, 'The Uses and Limits of Speculative Fiction: Three Novels about a US–China War', Air University (AU), 11 August 2023, <https://www.airuniversity.af.edu/JIPA/Display/Article/3490715/the-uses-and-limits-of-speculative-fiction-three-novels-about-a-uschina-war/https%3A%2F%2Fwww.airuniversity.af.edu%2FJIPA%2FDisplay%2FArticle%2F3490715%2Fthe-uses-and-limits-of-speculative-fiction-three-novels-about-a-uschina-war%2F>.

3. Mitigating the Risks of AI-Enabled Warfare

Effective risk mitigation requires a multi-pronged approach combining governance, technological safeguards, and international collaboration. At the governance level, governments must establish internationally recognized standards for AI deployment in military systems. Transparency in developing and using AI-enabled capabilities is essential to building mutual trust among nuclear-armed states. Expanding arms control agreements to include provisions for autonomous weapons and AI systems is critical to reducing risks.

Governance efforts should prioritize creating legal and ethical frameworks that ensure human accountability in AI-driven decision-making processes. This includes mandating human oversight in all critical operations, particularly those involving the use of force. By clearly delineating the roles and responsibilities of human operators, such frameworks can prevent the abdication of moral responsibility to machines.

Technological measures are equally important. AI systems should undergo rigorous testing to ensure their reliability under diverse operational conditions. Ensuring the reliability of AI systems is crucial for their safe and effective deployment. Rigorous testing plays a fundamental role in this process, as it helps identify vulnerabilities, biases, and potential failure modes before these systems are integrated into critical applications. Testing should include real-world simulations, stress-testing under extreme conditions, and adversarial evaluations to ensure AI systems can perform consistently across a wide range of scenarios. Moreover, continuous monitoring and iterative improvement must accompany initial testing to account for the dynamic nature of AI applications.

As AI systems interact with new environments and data sources, unexpected behaviours may emerge, requiring updates to training data, model parameters, and system architecture. Implementing robust feedback loops, safety mechanisms, and fail-safes ensures that AI remains reliable, adaptable, and aligned with intended goals even as operational conditions evolve. Incorporating fail-safe mechanisms and ensuring that AI algorithms are interpretable would enhance operator oversight and accountability. This is particularly vital in nuclear command, control, and communication systems, where errors could have devastating consequences. Additionally, the development of AI systems should prioritize adaptability, allowing for real-time adjustments to changing operational environments.

International collaboration must play a central role in mitigating AI risks. States should engage in bilateral and multilateral dialogues to share best practices and establish norms governing AI in military applications. Collaborative wargaming exercises and simulations can help identify vulnerabilities and inform the development of robust crisis management protocols. By fostering trust and communication, these efforts can reduce the likelihood of miscalculation and escalation during crises.

Research and education are foundational to effective risk mitigation. Governments and academic institutions should fund interdisciplinary studies exploring the intersection of AI, nuclear strategy, and human decision-making. Military personnel must be trained to understand AI systems' ethical and operational limitations. Fictionalized scenarios, grounded in empirical trends, can serve as valuable tools for preparing policymakers and military leaders to navigate the complexities of AI-enabled warfare. Such training exercises can simulate high-pressure scenarios, providing participants with firsthand experience managing AI-driven escalation risks.

4. Policy Implications and Recommendations

Integrating artificial intelligence (AI) into military and nuclear systems presents profound challenges for international security. Policymakers must act decisively to mitigate the risks of inadvertent escalation and miscalculation while preserving strategic stability.¹⁵ This section outlines key policy recommendations grounded in this policy note's findings.

First, internationally recognized safety standards for AI-enabled military systems must be established. These standards should ensure that AI technologies are robust, reliable, and interpretable under diverse operational conditions. Transparency in their design and deployment is critical to building trust among nuclear-armed states. Clear protocols must delineate the roles of human operators and AI systems, ensuring that humans retain ultimate decision-making authority, particularly in scenarios involving the use of force.

Second, existing arms control agreements must be expanded to address the proliferation of autonomous weapon systems and AI-enhanced command structures. Such agreements should explicitly restrict the use of fully autonomous systems in nuclear operations and emphasize the importance of human oversight. These measures can prevent destabilizing arms races and foster greater international collaboration in managing the dual-use nature of AI technologies.

Third, international collaboration is essential for addressing the shared risks posed by AI-enabled warfare. Bilateral and multilateral dialogues can help establish norms and best practices for using AI in military contexts. Joint wargaming exercises and simulation-based crisis management training offer opportunities to identify vulnerabilities and enhance communication protocols.¹⁶ These initiatives can also build mutual understanding of how AI systems operate under high-pressure conditions, reducing the likelihood of miscalculation.

Fourth, investment in interdisciplinary research and education must be prioritized. Governments, academic institutions, and research organizations should explore the intersection of AI, human decision-making, and nuclear strategy. Studies addressing AI's ethical, legal, and operational dimensions are crucial for informing policy decisions. Additionally, military personnel should receive comprehensive training on the limitations and risks associated with AI systems, fostering a culture of ethical accountability and critical oversight.

¹⁵ See: Bruce G. Blair, 'Nuclear Inadvertence: Theory and Evidence', *Security Studies* 3, no. 3 (1 March 1994): 494–500, <https://doi.org/10.1080/09636419409347558>; Paul Bracken, 'The Command and Control of Nuclear Forces.', *American Political Science Review*, 1983, <https://doi.org/10.2307/1956732>; Peter Feaver, 'Guarding the Guardians: Civilian Control of Nuclear Weapons in the United States', *American Political Science Review*, 1992, <https://doi.org/10.2307/2938877>; Barry R. Posen, *Inadvertent Escalation: Conventional War and Nuclear Risks* (Cornell University Press, 1991), <https://www.jstor.org/stable/10.7591/j.ctt1xx51d>.

¹⁶ Anna Knack and Rosamund Powell, 'Artificial Intelligence in Wargaming', 2023, <https://cetas.turing.ac.uk/publications/artificial-intelligence-wargaming>.

Finally, using fictional scenarios as a policy tool offers significant potential for preparing policymakers and military leaders. Scenarios grounded in empirical trends can simulate complex, high-pressure situations, providing insights into the cascading effects of AI-driven decision-making. By challenging assumptions and highlighting potential points of failure, these exercises can inform the development of more resilient crisis management strategies.

Implementing these recommendations can help policymakers navigate the challenges of AI-enabled warfare and prevent the catastrophic consequences of miscalculation or unintended escalation.¹⁷ Proactive governance, coupled with international cooperation and a commitment to transparency, will be essential for ensuring that AI technologies contribute to global stability rather than undermining it.

¹⁷ As a counterpoint to the widely held thesis that AI-enabling weapons systems are necessarily a force for instability and escalation in the Third Nuclear Age, see: Andrew Futter and Benjamin Zala, 'Strategic Non-Nuclear Weapons and the Onset of a Third Nuclear Age', *European Journal of International Security* 6, no. 3 (August 2021): 257–77, <https://doi.org/10.1017/eis.2021.2>.

5. Conclusion: Navigating the Promethean Paradox

Integrating AI technology into military systems introduces both profound opportunities and grave risks, especially within nuclear deterrence. While AI promises to enhance operational efficiency and reduce human error, it also exacerbates the destabilizing dynamics of the stability-instability paradox. The rapid escalation of conflicts, compounded by the opacity of AI decision-making processes, presents a new set of challenges that threaten to undermine strategic stability and increase the likelihood of catastrophic miscalculation.

The hypothetical scenarios explored in this policy note—the "2030 Flash War" and "Operation Island Freedom"—demonstrate how AI systems can misinterpret situations, accelerate decision-making, and escalate crises in ways that may be difficult for human operators to control. These scenarios highlight the urgent need for comprehensive risk mitigation strategies encompassing robust governance frameworks, technological safeguards, and international collaboration. Without proactive measures, AI's dual-use nature could push global security to the brink of unprecedented dangers, particularly in nuclear contexts.

To address these risks, the policy note has outlined several key recommendations:

- Establish international standards for AI-enabled military systems.
- Expand arms control agreements to include AI.
- Promote transparency and human oversight.
- Foster global collaboration through joint exercises and research.


When implemented effectively, these measures can help mitigate the dangers posed by AI in warfare and nuclear deterrence, ensuring that AI technologies are used responsibly and ethically.

In conclusion, as we navigate the complexities of AI-enabled warfare, policymakers must recognize the profound implications of these technologies for global security. Through proactive governance, international cooperation, and ongoing research, we can manage the risks of AI in military systems, ensuring that these technologies contribute to stability rather than undermining it. AI's dual-use nature demands our vigilance and foresight as the potential for progress and peril remains ever-present.

About the Author

Dr. James Johnson

Dr. James Johnson is a Senior Lecturer (Associate Professor) and Director of Strategic Studies in the Department of Politics and International Relations at the University of Aberdeen. He is also an Honorary Fellow at the University of Leicester, a Non-Resident Research Associate on the European Research Council funded Towards a Third Nuclear Age Project, and a Mid-Career Cadre Member with the Center for Strategic and International Studies (CSIS) Project on Nuclear Issues.



HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
[@hcssnl](#)

The Hague Centre for Strategic Studies