Review of International Studies (2024), page 1 of 19
doi:10.1017/S0260210524000767

RIS

RESEARCH ARTICLE

Revisiting the 'stability–instability paradox' in AI-enabled warfare: A modern-day Promethean tragedy under the nuclear shadow?

James Johnson

Department of Politics and International Relations, King's College, University of Aberdeen, Aberdeen, UK
Email: james.johnson@abdn.ac.uk

(Received 26 February 2024; revised 7 October 2024; accepted 7 October 2024)

Abstract

This article contributes to the empirical and theoretical discourse on the 'stability–instability paradox', the idea that while possessing nuclear weapons deters cataclysmic all-out war, it simultaneously increases the likelihood of low-level conflict between nuclear dyads. It critiques the paradox's dominant interpretation (red-line model), which places undue confidence in the nuclear stalemate – premised on mutually assured destruction – to prevent unintentional nuclear engagement and reduce the perceived risks associated with military actions that fall below the nuclear threshold. Recent scholarship has inadequately examined the unintentional consequences of the paradox in conflicts below the nuclear threshold, particularly those relating to the potential for aggression to escalate uncontrollably. The article employs empirically grounded fictional scenarios to illustrate and critically evaluate, rather than predict, the assumptions underpinning the red-line model of the stability–instability paradox in the context of future artificial intelligence (AI)-enabled warfare. It posits that the strategic cap purportedly offered by a nuclear stalemate is illusory and that low-level military aggression between nuclear-armed states increases the risk of unintentional nuclear detonation.

Keywords: artificial intelligence; brinkmanship; nuclear weapons; science fiction; stability–instability paradox

Introduction

This article revisits Glenn Snyder's seminal theory on the 'stability–instability paradox' to consider the risks associated with the artificial intelligence (AI)–nuclear nexus in future digitised warfare.[1] It critiques the dominant interpretation (or 'red-line'[2]) of Snyder's theory: the presence of restraint among nuclear-armed states under mutually assured destruction (MAD) – where both sides possess nuclear weapons that present a plausible mutual deterrent threat – does not inhibit low-level conflict from occurring under the nuclear shadow.[3] The article uses fictional scenarios to illustrate

[1]Glenn H. Snyder, 'The balance of power and the balance of terror', in Paul Seabury (ed.), *The Balance of Power* (San Francisco, CA: Chandler, 1965), pp. 184–201.

[2]Christopher Watterson, 'Competing interpretations of the stability–instability paradox: The case of the Kargil War', *Nonproliferation Review*, 24:1–2 (2017), pp. 83–99 (pp. 86–8).

[3]The foundational literature on accidental and inadvertent escalation, includes Bruce G. Blair, 'Nuclear inadvertence: Theory and evidence', *Security Studies*, 3:3 (1994), pp. 494–500; Paul Bracken, *The Command and Control of Nuclear Forces* (New Haven, CT: Yale University Press, 1983); Peter D. Feaver, *Guarding the Guardians: Civilian Control of Nuclear Weapons in*

how deploying 'AI-enabled weapon systems' in a low-level conflict between the United States and China in the Taiwan Straits might unintentionally spark a nuclear exchange.[4] The scenarios support the alternative interpretation, known as the 'brinkmanship model', of the paradox conceptualised by Christopher Watterson.[5] In other words, they are incentivised to exploit MAD and engage in military adventurism to change the prevailing political status quo or make territorial gains.[6] The scenarios indicate that using AI-enabled weapon systems accelerates a pre-existing trend of modern warfare associated with speed, precision, and complexity, which reduces human control and thus increases escalation risk.[7] The use of fictional scenarios in this article provides a novel and imaginative conceptual tool for strategic thinking to challenge assumptions and stimulate introspection that can improve policymakers' ability to anticipate and prepare for (but not predict) change.

Although Snyder's 1961 essay was the first to elaborate on the stability–instability paradox in detail, he was not the first scholar to identify this phenomenon. B. H. Liddell Hart, in 1954, for example, noted that 'to the extent that the hydrogen bomb reduces the likelihood of full-scale war, it increases the possibility of limited war pursued by widespread local aggression'.[8] The Eisenhower administration's declaratory policy of massive retaliation (i.e. first strike capacity) has primarily been attributed to Hart's strategic rationale.[9] Building on Snyder's essay, Waltz opined that while nuclear weapons deter nuclear use, they may cause a 'spate of smaller wars'.[10] Similarly, Robert Jervis posited that though the logic of the paradox is sound, in the real world, the stabilising effects of the nuclear stalemate (i.e. a strategic cap) implied by the paradox cannot be taken for granted.[11] In other words, neither Waltz nor Jervis was confident that escalation could be controlled. In addition to Snyder's paradox, two other influential strategic concepts emerged in the 1950s, adding a new layer of sophistication and nuance to those originally propounded by Bernard Brodie and William Borden a decade earlier:[12] Thomas Schelling's game-theoretic notion of 'threats that leave something to chance',[13] and Albert Wohlstetter's first and second strike clarification.[14] In subsequent years, these foundational ideas were revisited

the United States (Ithaca, NY: Cornell University Press, 1992); Peter D. Feaver, 'The politics of inadvertence', *Security Studies*, 3:3 (1994), pp. 501–8; Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton, NJ: Princeton University Press, 1993); Samuel B. Robison, 'Conventional wisdom? The effect of nuclear proliferation on armed conflict, 1945–2001', *International Studies Quarterly*, 56:1 (2012), pp. 149–62; Posen R. Barry, *Inadvertent Escalation* (Ithaca, NY: Cornell University Press, 1990).

[4]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* (Oxford: Oxford University Press, 2023), pp. 4–23.

[5]Watterson, 'Competing interpretations of the stability–instability paradox', pp. 87–8.

[6]See Peter R. Lavoy, 'The strategic consequences of nuclear proliferation: A review essay', *Security Studies*, 4:4 (1995), pp. 695–753 (p. 739); Watterson, 'Competing interpretations of the stability-instability paradox', pp. 87–8.

[7]Keir A. Lieber and Daryl G. Press, 'The end of MAD? The nuclear dimension of U.S. primacy', *International Security*, 30:4 (2006), pp. 7–44.

[8]B. H. Liddell Hart, *Deterrent or Defense: A Fresh Look at The West's Military Position* (London: Kessinger Publishing, 2010), p. 23.

[9]Michael Krepon, 'The stability–instability paradox, misperception, and escalation control in South Asia', Henry L. Stimson Center, Washington, DC (2003), p. 1.

[10]Kenneth N. Waltz, *Man, the State, and War* (New York: Columbia University Press, 1959), p. 259. In his later work on the subject, Waltz more explicitly advocates the (red-line model) of the 'paradox'. Kenneth N. Waltz, 'The spread of nuclear weapons: More may be better. Introduction', *The Adelphi Papers*, 21:171 (1981), pp. 1478–5145.

[11]Robert Jervis, *The Illogic of American Nuclear Strategy* (Ithaca, NY: Cornell University Press, 1984); Robert Jervis, *The Meaning of the Nuclear Revolution* (Ithaca, NY: Cornell University Press, 1989).

[12]Bernard Brodie (ed.), *The Absolute Weapon: Atomic Power and World Order* (New York: Harcourt, 1946); William L. Borden, *There Will Be No Time* (New York: Macmillan, 1946).

[13]Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960).

[14]Albert Wohlstetter, 'Selection and the use of strategic air bases', in Marc Trachtenberg (ed.), *The Development of American Strategic Thought: Writing in strategy 1952–1960, 1* (New York: Garland, 1988), pp. 365–69.

to consider nuclear brinkmanship,[15] the 'nuclear revolution',[16] and the impact of emerging technology.[17]

Recent empirical studies on the paradox, while explaining *how* leaders perceive risk when engaging in low-level conflict in nuclear dyads and the unintentional consequences of the paradox playing out in a conflict below the nuclear threshold, say less about how threats of aggression and military adventurism can inadvertently or accidentally escalate out of control.[18] Less still has been said about how emerging technology might affect these dynamics.[19] The limited evidentiary base can, in part, be explained by: (a) the lack of observable empirical phenomena – nuclear weapons have not been used since 1945, and we have not seen an AI-enabled war between nuclear powers[20] – and (b) bias inference problems caused by case selection (qualitative studies focus heavily on the India–Pakistan and US–Soviet dyads) to *prove* the validity of the paradox retroactively.[21] Combining empirical plausibility, theoretical rigour, and imagination, the article uses fictional scenarios (or 'future counterfactuals'[22]) to illustrate and test (not predict or prove) the underlying assumptions of the dominant red-line interpretation of Snyder's stability–instability paradox in future AI-enabled war. In doing so, it contributes to the discourse on using fictionalised accounts of future conflict, filling a gap in the literature about the unintended escalatory consequences of the paradox in modern warfare.

The article is organised into three sections. The first section examines the core logic, key debates, and underlying assumptions of the stability–instability paradox. The second section describes the main approaches to constructing fictional scenarios to reflect on future war and how we can optimise the design of these scenarios to consider the implications of Snyder's paradox of introducing AI-enabled weapons systems into nuclear dyads. The final section uses two fictional scenarios to illustrate how AI-enabled weapons systems in a future Taiwan crisis between the United States and China might spark inadvertent and accidental nuclear detonation.[23] The scenarios expose the fallacy of the supposed safety offered by nuclear weapons in high-intensity digitised conflict under the nuclear shadow.

[15]The literature on nuclear brinkmanship is expansive; seminal works include Robert Powell, 'Nuclear brinkmanship, limited war, and military power', *International Organization*, 69:3 (2015), pp. 589–626; Reid B. C. Pauly and Rose McDermott, 'The psychology of nuclear brinkmanship', *International Security*, 47:3 (2023), pp. 9–51; Michael Dobbs, *One Minute to Midnight: Kennedy, Khrushchev, and Castro on the Brink of Nuclear War* (New York: Alfred A. Knopf, 2008); Mark S. Bell and Julia Macdonald, 'How to think about nuclear crises', *Texas National Security Review*, 2:2 (2019), pp. 41–64.

[16]Keir Lieber and Daryl G. Press, *The Myth of the Nuclear Revolution: Power Politics in the Atomic Age* (Ithaca, NY: Cornell University Press, 2020).

[17]Matthew Kroenig, 'Will emerging technology cause nuclear war? Bringing geopolitics back in', *Strategic Studies Quarterly*, 15:4 (2021), pp. 59–73.

[18]See Sagan, *The Limits of Safety*; Robert Powell, 'Nuclear deterrence theory, nuclear proliferation, and national missile defense', *International Security*, 27:4 (2003), pp. 86–118; Bruce G. Blair, *The Logic of Accidental Nuclear War* (Washington, DC: Brookings, 1993).

[19]Notable exceptions include Todd S. Sechser, Neil Narang, and Caitlin Talmadge, 'Emerging technologies and strategic stability in peacetime, crisis, and war', *Journal of Strategic Studies*, 42:6 (2019), pp. 727–35; Michal Onderco and Madeline Zutt, 'Emerging technology and nuclear security: What does the wisdom of the crowd tell us?', *Contemporary Security Policy*, 42:3 (2021), pp. 286–31.

[21]Some scholars have used counterfactual reasoning to circumvent these empirical limitations. For example, see Francesco Bailo and Benjamin Goldsmith, 'No paradox here? Improving theory and testing of the nuclear stability–instability paradox with synthetic counterfactuals', *Journal of Peace Research*, 58:6 (2021), pp. 1178–93.

[22]Steven Weber, 'Counterfactuals, past, and future', in Philip Tetlock and Aaron Belkin (eds), *Counterfactual Thought Experiments in World Politics* (Princeton, NJ: Princeton University Press, 1996), pp. 268–91.

[23]For a conceptual discussion of accidental and inadvertent escalation in the context of AI-enabled future warfare, see James Johnson, 'Catalytic nuclear war in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states', *Journal of Strategic Studies* (2021), available at: {https://doi.org/10.1080/01402390.2020.1867541}.

https://doi.org/10.1017/S0260210524000767

Rethinking the stability–instability paradox in future AI-enabled warfare

The core logic of the stability–instability paradox captures the relative relationship between the probability of a conventional crisis or conflict and a nuclear crisis or conflict. Snyder's 1961 essay is generally accepted as the seminal treatment of paradox theorising. Specifically, 'the greater the stability of the "strategic" [nuclear] balance of terror, the lower the stability of the overall balance at its lower levels of violence'.[24] Put differently, when a nuclear stalemate (or MAD) exists,[25] the probability of a low-level crisis or conflict is higher (i.e. crisis instability). According to Snyder: 'If neither side has a "full first-strike capability", and both know it, they will be less inhibited about the limited use of nuclear weapons than if the strategic (nuclear) balance were unstable.'[26] In other words, under the condition of MAD, states are more likely to be deterred from risking all-out nuclear Armageddon. Still, because of the strategic cap this balance is believed to offer, they are more likely to consider limited nuclear warfare less risky.

Similarly, Jervis notes that 'the extent that the military balance is stable at the level of all-out nuclear war, it will become less stable at lower levels of violence'.[27] Despite the sense of foreboding during the Cold War about the perceived risks of nuclear war escalation, which led to US–Soviet arms racing dynamics – casting doubt on the stability equation of the paradox equation – the paradox's larger rationale remains theoretically sound. That is, nuclear-armed states, all things being equal, would exercise caution to avoid major wars and nuclear exchange. As a corollary, the credible threat of nuclear retaliation (or second-strike capability) provides states with the perception of freedom of manoeuvre to engage in brinkmanship, limited wars (including the use of tactical nuclear weapons),[28] proxy wars, and other forms of low-level provocation.[29] This interpretation of the paradox, also known as the red-line model (or strategic cap), remains dominant in the strategic literature and policymaking circles.[30] The link between this interpretation of the paradox and pathways to conflict, specifically intentional versus unintentional and accidental nuclear use, is explained by the less appreciated connection with nuclear brinkmanship described below.

Although Snyder's 1961 essay does not offer a rigorous causal theoretical explanation of low-level aggression in nuclear dyads conditioned on MAD, he does acknowledge the 'links' between the nuclear and conventional balance of power and how they 'impinge on each other'. Conventional (i.e. non-nuclear) instability in nuclear stalemates can lead to strategic instability, thus destabilising the 'overall balance of power'.[31] Snyder describes several casual contingencies (or 'provocations') in the relationship between nuclear and conventional stability that could risk nuclear escalation:

> In a confrontation of threats and counterthreats, becoming involved in a limited conventional war which one is in danger of losing, suffering a limited nuclear strike by the opponent, becoming involved in an escalating nuclear war, and many others. Any of these events may drastically increase the potential *cost and risk of not striking first* and thereby *potentially create a state of disequilibrium in the strategic nuclear balance*.[32]

[24] Snyder, 'The balance of power and the balance of terror', p. 199.

[25] See Lieber and Press, 'The end of MAD', pp. 7–44.

[26] Snyder, 'The balance of power and the balance of terror', p. 199.

[27] Jervis, *The Illogic of American Nuclear Strategy*, p. 31.

[28] Limited nuclear wars can occur due to states engaging in nuclear brinkmanship, but they are treated as distinctive concepts in the nuclear security literature. See, for example, Fiona S. Cunningham and M. Taylor Fravel, 'Dangerous confidence? Chinese views on nuclear escalation', *International Security*, 44:2 (2019), pp. 61–109.

[29] For recent research on the intersection of a credible nuclear deterrence and the risk reduction agenda, see Benoît Pelopidas and Kjølv Egeland, 'The false promise of nuclear risk reduction', *International Affairs*, 100:1 (2024), pp. 345–60.

[30] Bryan R. Early and Victor Asal, 'Nuclear weapons, existential threats, and the stability–instability paradox', *The Nonproliferation Review*, 25:3–4, (2018), pp. 223–47 (p. 229).

[31] Jervis, *The Illogic of American Nuclear Strategy*, p. 191.

[32] Ibid., p. 198, emphasis added.

Snyder's causal links formulation offers scholars a useful theoretical baseline to explore the risk of nuclear escalation in the overall balance, particularly the role of nuclear brinkmanship and the attendant trade-off between the perceived risk of total war and the possibilities for military adventurism (e.g. coercive pressures, escalation dominance, conventional counterforce, and false-flag operations).[33] According to Robert Powell, 'states exert coercive pressure on each other during nuclear brinkmanship by taking steps that *raise the risk that events will go out of control*. This is a *real and shared risk* that the confrontation will end in a catastrophic nuclear exchange' (emphasis added).[34] In other words, brinkmanship is a dangerous game of chicken and a manipulation risk that risks uncontrolled inadvertent escalation in the hope of achieving a political gain.[35]

While some scholars characterise brinkmanship as a distinct corpus of the paradox theorising, the connection made by Snyder (albeit not explicitly or in depth) suggests that brinkmanship behaviour is an intrinsic property of the paradox.[36] Although Thomas Schelling's notion of the 'threat that leaves something to chance' says more about the role of uncertainty in deterrence theory, nonetheless, it helps elucidate how and why nuclear dyads might engage in tests of will and resolve at the precipice of Armageddon, and the escalatory effects of this behaviour.[37] Nuclear brinkmanship, understood as a means to manipulate risk during crises, establishes a consistent theoretical link between the paradox and unintentional (inadvertent and accidental) nuclear use. The first scenario in this article illustrates how such a catalysing chain of 'threats and counter threats' could spark an inadvertent 'limited' nuclear exchange.

This feature of the paradox can be viewed from two casual pathways: the red-line model versus the brinkmanship model. Although there is no consensus on the causal mechanisms underpinning an increase in low-level conflict in a relationship of nuclear stalemate (i.e. MAD),[38] the dominant red-line model view holds that mutual fear of nuclear escalation under the conditions of MAD reduces the risk and uncertainty of states' pursuing their strategic goals (altering the territorial or political status quo), which drives military adventurism.[39] The alternative brinkmanship model interpretation posits that the threat of nuclear use to generate uncontrollable escalation risk may incentivise states to accept the risk of uncontrollable escalation to a nuclear exchange to obtain concessions.[40]

These two interpretations predict divergent empirical outcomes. The red-line model predicts there is a negligible risk of lower-level military adventurism causing an all-out nuclear war.

[33] An alternative causal inference (connecting the possession of nuclear weapons with conflict) could posit that symmetrical nuclear dyads are associated with conventional conflict because it is violence that leads them to pursue nuclear weapons in the first instance, rather than a nuclear stalemate causing low-level conflict. Mark S. Bell and Nicholas L. Miller, 'Questioning the effect of nuclear weapons on conflict', *The Journal of Conflict Resolution*, 59:1 (2015), pp. 74–92.

[34] Powell, 'Nuclear deterrence theory, nuclear proliferation, and national missile defense', p. 90.

[35] Pauly and McDermott, 'The psychology of nuclear brinkmanship'.

[36] Watterson, 'Competing interpretations of the stability–instability paradox', pp. 87–8.

[37] Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960); Robert V. Dodge, *Schelling's Game Theory: How to Make Decisions* (New York: Oxford University Press, 2012); Pauly and McDermott, 'The psychology of nuclear brinkmanship'.

[38] S. Paul Kapur, *Dangerous Deterrent: Nuclear Weapons Proliferation and Conflict in South Asia* (Stanford, CA: Stanford University Press, 2007), pp. 35–6.

[39] For examples of the red-line interpretation, see Kapur, *Dangerous Deterrent*; Robert Rauchhaus, 'Evaluating the nuclear peace hypothesis: A quantitative approach', *The Journal of Conflict Resolution*, 53:2 (2009), pp. 258–77; John Mearsheimer, *The Tragedy of Great Power Politics* (New York: W. W. Norton, 2001); Kenneth N. Waltz, 'Nuclear myths and political realities', *American Political Science Review*, 84 (2014), pp. 731–45.

[40] For examples of the brinkmanship interpretation, see Powell, 'Nuclear deterrence theory, nuclear proliferation, and national missile defense', Peter R. Lavoy, 'The strategic consequences of nuclear proliferation: A review essay', *Security Studies*, 4:4 (1995), pp. 695–753 (pp. 739–40); Austin Long, 'Proliferation and strategic stability in the middle east', in Elbridge A. Colby and Michael S. Gerson (eds), *Strategic Stability: Contending Interpretations* (Carlisle, PA: US Army War College, 2013), pp. 383–433 (p. 385); Sumit Ganguly and Harrison Wagner, 'India and Pakistan: Bargaining in the shadow of nuclear war', *Journal of Strategic Studies*, 27:3 (2004), pp. 479–507; Fiona S. Cunningham, 'Strategic substitution: China's search for coercive leverage in the information age', *International Security*, 47:1 (2022), pp. 46–92.

Meanwhile, the brinkmanship model predicts that lower-level military adventurism inherently carries the irreducible risk of nuclear escalation.[41] Despite general agreement amongst scholars about the paradox's destabilising effects at a lower level, the literature is unclear as to how the causal mechanisms underpinning the paradox cause an increase in instability at lower levels. Francesco Bailo and Benjamin Goldsmith caution that 'causation is attributed to nuclear weapons, although the qualitative evidence is often interpreted through the prism of the [stability–instability paradox] theory' and that the near-exclusive focus on the US–Soviet and India–Pakistan nuclear cases has led to an inference that is biased by the selection on the dependent variable.[42] The fictional scenarios in this article use the paradox as a theoretical lens to construct a causal mechanism to envisage a future US–China nuclear exchange in the Taiwan Straits, thus expanding the dependent variable and reducing the risk of bias selection.

Thomas Schelling argues that the power of these dynamics is the 'danger of a war that *neither wants*, and both will have to pick their way carefully through crisis, *never quite sure that the other knows how to avoid stumbling over the brink*' (emphasis added).[43] According to this view, the dominant interpretation of the paradox exaggerates the stability of MAD. It thus neglects the potency of Schelling's 'threats that leave something to chance' – the risk that military escalation cannot be entirely controlled in a low-level conventional conflict.[44] Under conditions of MAD, states cannot credibly threaten *intentional* nuclear war, but, during crises, they can invoke 'threats that leave something to chance', thereby, counter-intuitively, increasing the risk of inadvertent and accidental nuclear war.[45]

Schelling's 'threats that leave something to chance' describes the manoeuvrability of military force by nuclear dyads short of all-out war and thus helps resolve the credibility problem of states making nuclear threats during a crisis.[46] Schelling's game theoretic 'competition in risk-taking' describes the dynamics of nuclear bargaining in brinkmanship, which is heavily influenced by actors' risk tolerance.[47] The higher the stakes, the more risk a state is willing to take. Suppose one side can raise the costs of victory (i.e. defend or challenge the status quo) more than the value the other side attaches to victory. In that case, the contest can become a 'game of chicken' – one that the initiator cannot be sure their adversary will not fight for reasons such as national honour, reputation, the belief the other side will buckle, or misperception and miscalculation.[48] For instance, during crises, leaders are prone to viewing their actions (e.g. deterrence signalling, troop movements, and using 'reasonable' force to escalate a conflict to compel the other side to back down) as benign and view similar behaviour by the other side as having malign intent. Both sides in an escalating crisis could, under the perception that they were displaying more restraint than the other, inadvertently escalate a situation.[49]

One of the goals of competition in risk-taking is to achieve escalation dominance:[50] possessing the military means to contain or defeat an enemy at every stage of the escalation ladder with the 'possible exception of the highest', i.e. an all-out nuclear war where dominance is less relevant.[51]

[41] Recent empirical studies have lent additional support to the 'escalation risk' interpretation and the view that the paradox encourages low-level aggression and increases existential threats. Early and Asal, 'Nuclear weapons, existential threats, and the stability–instability paradox', pp. 223–47.

[42] Bailo and Goldsmith, 'No paradox here?', p. 1190.

[43] Thomas C. Schelling, *Arms and Influence* (New Haven, CT: Yale University Press, 1966), pp. 89–9, 166–8.

[44] Thomas Schelling, *The Threat That Leaves Something to Chance* (Santa Monica, CA: RAND Corporation, 1959).

[45] Powell, 'Nuclear brinkmanship, limited war, and military power', p. 89.

[46] Schelling, *The Strategy of Conflict*.

[47] Jervis, *The Illogic of American Nuclear Strategy*, p. 130.

[48] Ibid., p. 132.

[49] Ibid.

[50] By contrast, it is plausible that an actor without escalation dominance could 'win' the competition in risk-taking, for instance, if their interests weigh heavy enough compared to the other sides' 'balance of terror' and 'escalation dominance' strategic calculations. The author would like to thank an anonymous reviewer for this counterpoint.

[51] Ibid., p. 130.

The assumptions underpinning escalation dominance include accurately predicting the effects of moving up the escalation ladder, the problem of determining the balance of power and resolve during crises, and divergences between both sides' perceptions of the balance.[52] The first scenario in this article illustrates the oversimplification of the assumptions, and thus the inherent risks, that belie escalation dominance. Jervis argues that in the 'real world', absent perfect information, control, and free from Clausewitzian friction, the stabilising effect of nuclear stalemate associated with the 'nuclear revolution' (i.e. secure, second-strike forces) 'although not trivial, is not so powerful'.[53] In other words, though the logic of MAD has indubitable deterrence value, on the battlefield, its premise is flawed because it neglects Schelling's 'threats that leave something to chance', the uncertainty associated with human behaviour (e.g. emotions, bias, inertia, mission creep, and the war machine) during a crisis and, increasingly, the intersection of human psychology with the more and more digitised battlefield.[54] Advances in AI technology (and the broader digitalisation of the battlefield) have not eliminated and, arguably, have thickened the 'fog of war'.[55] For example, similar to the 1973 intelligence failure that led Israel to be surprised by the Arab attacks that initiated the Yom Kippur War, Israeli intelligence, despite possessing world-class Signals Intelligence (SIGINT) and cyber capabilities, miscalculated both Hamas's capabilities and its intentions in the 2023 terrorist attack.[56]

Integrating AI technology with nuclear command, control, and communication systems (NC3) enhances states' early warning systems, situational awareness, early threat detection, and decision-support capabilities, thus raising the threshold below which a nuclear first strike is successful.[57] However, like previous generations of technological enhancements, AI cannot alter the immutable dangers associated with the unwinnable nature of nuclear war.[58] Jervis writes that 'just as improving the brakes on a car would not make playing [the game of] Chicken safer, but only allow them to drive faster to produce the same level of intimidation', so the integration of AI into NC3 will only lead both sides to recalibrate their actions to produce a requisite level of perceived danger to accomplish their goals.[59] In short, neither side in a nuclear dyad can confidently engage in low-level adventurism without incurring 'very high costs, if not immediately, then as a result of a *chain of actions that cannot be entirely foreseen or controlled*' (emphasis added).[60] Indeed, Schelling opined that the lack of control against the probability of nuclear escalation is the 'essence of a crisis'.[61] Similarly, Alexander George and Richard Smoke found that an important cause of deterrence failure is the aggressor's belief that they can control the risks during a crisis.[62] In sum, low-level adventurism under the nuclear shadow (out of confidence that MAD will prevent escalation) does little to contain the likelihood of sparking a broader conflict neither side favours. The scenarios below consider how these 'costs' and 'chains of action' are affected in AI-enabled warfare.

[52] Ibid., p. 131.

[53] Ibid., p. 148.

[54] Lieber and Press, *The Myth of the Nuclear Revolution*; Jervis, *The Illogic of American Nuclear Strategy*, pp. 130–1.

[55] James Johnson, 'Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age', *European Journal of International Security*, 7:3 (2022), pp. 337–59.

[56] Martin van Creveld, *Command in War* (Cambridge, MA: Harvard University Press, 1985), pp. 185–230; Emily Harding, 'How could Israeli intelligence miss the Hamas invasion plans?', *Center for Strategic & International Studies* (11 October 2023).

[57] For an excellent study on AI's ability to replace humans in command decision-making and the potential effects of this phenomenon, see Cameron Hunter and Bleddyn E. Bowen, 'We'll never have a model of an AI major-general: Artificial intelligence, command decisions, and kitsch visions of war', *Journal of Strategic Studies*, 47:1 (2024), pp. 116–46.

[58] Barry Nalebuff, 'Brinkmanship and deterrence: The neutrality of escalation', *Conflict Management and Peace Science*, 9 (1986), pp. 19–30.

[59] Jervis, *The Meaning of the Nuclear Revolution*, p. 96.

[60] Jervis, *The Illogic of American Nuclear Strategy*, p. 148.

[61] Schelling, *Arms and Influence*, p. 97.

[62] Alexander George and Richard Smoke, *Deterrence in American Foreign Policy: Theory and Practice* (New York: Columbia University Press, 1974), pp. 527–30.

Fictional scenarios to anticipate future war

According to a recent RAND Corporation report, 'history is littered with mistaken predictions of warfare'.[63] With this warning as an empirical premise, the future warfare fictional scenarios in this article explore how future inadvertent and accidental nuclear war might unfold in a future Taiwan Straits conflict. Traditional approaches to anticipating future conflict rely on counterfactual qualitative reasoning, quantitative inference, game theory logic, and historical deduction.[64] By contrast, the scenarios presented here combine present and past trends in technological innovation with imagination and thus contribute to the existing corpus of work that considers hitherto-unknown or underexplored drivers of future wars (e.g. new technologies, environmental change, or ideological/geopolitical shifts) to imagine how future conflict could emerge. Thus, these fictional scenarios challenge existing assumptions, operational concepts, and conventional wisdom associated with the genesis of war, offering foresight into the probability of future AI-enabled war, its onset, how it might unfold, and, importantly, how it might end.

Three broad approaches have been used to anticipate future conflict.[65] The first approach uses quantitative methods to consider internal causes of conflict, or 'push factors', such as socio-economic development, arms control, arms racing, and domestic politics.[66] Rapid progress in AI and big data analytics has significantly empowered the progress of this field – for example, Lockheed Martin's Integrated Crisis Early Warning System, the EU's Conflict Early Warning System, and Uppsala University's ViEWS.[67] These models use statistical probabilistic methods to predict the conditions under which a conflict (mainly civil wars sparked by internal factors) might be triggered, not the causal pathways between states leading to war and *how* a war might play out. In short, this approach is restricted in forecasting interstate conflict, and its reliance on historical datasets means that it cannot consider novelty and challenging assumptions.

The second approach, established during the Cold War, applies qualitative historical deduction and international relations theorising (e.g. hegemonic stability theory, 'anarchy' in international relations, neo-realism, the security dilemma, the democratic-peace theory, deterrence theory, and the related stability–instability paradox, which this study explores)[68] to top-down 'pull factors' to understand 'international politics, grasp the meaning of contemporary events, and foresee and influence the future'.[69] This approach has posited explanations and causes for conflicts ranging from the First World War, the Cold War, ethnic conflict in Africa, the break-up of Yugoslavia, and, more recently, the conflict in Ukraine.[70] This approach has also been used to explain the absence

[63] Raphael S. Cohen, Nathan Chandler, Shira Efron et al., *The Future of Warfare in 2030: Project Overview and Conclusions* (Santa Monica, CA: RAND Corporation, 2020).

[64] For example, see Bailo and Goldsmith, 'No paradox here?'; Bell and Miller, 'Questioning the effect of nuclear weapons on conflict'; Powell, 'Nuclear brinkmanship, limited war, and military power'.

[65] Florence Gaub (ed.), 'Conflicts to Come: 15 Scenarios from 2030', *The European Union Institute for Security Studies, Challiot Paper 161* (December 2020), pp. 2–9.

[66] See Guo Weisi Kristian Gleditsch and Alan Wilson et al., 'Retool AI to forecast and limit wars', *Nature*, 562 (2018), pp. 331–3; Lars-Erik Cederman and Nils B. Weidmann, 'Predicting armed conflict: Time to adjust our expectations?', *Science*, 355 (2017), pp. 474–6.

[67] Tate Ryan Mosley, 'We are finally getting better at predicting organized conflict', *MIT Technology Review*, 122 (2019), available at: {https://cdn.technologyreview.com/s/614568/predicting-organized-conflict-ensemble-modeling-ethiopia-ahmed/}; Stamatia Halkia, Stefano Ferri, Yannick Deepen et al., 'The global conflict risk index: Artificial intelligence for conflict prevention', *JRC Technical Reports* (2020).

[68] See Stephen D. Krasner, 'Hegemonic stability theory: An empirical assessment review of international studies', *Special Issue on the Balance of Power*; Robert Jervis, 'Cooperation under the security dilemma', *World Politics*, 30:2 (1978), pp. 169–214; Patrick M. Morgan, *Deterrence Now* (Cambridge: Cambridge University Press, 2003); Kori Schake, 'What causes war?', *Orbis*, 61:4 (2017), pp. 449–62; Randall L. Schweller, 'Neorealism's status-quo bias: What security dilemma?', *Security Studies*, 5:3 (1996), pp. 90–121; Azar Gat, 'The democratic peace theory reframed: The impact of modernity', *World Politics*, 58:1 (2005), pp. 73–100.

[69] Hans Morgenthau, *Politics among Nations: The Struggle for Power and Peace* (New York: Knopf, 1948), pp. 4–5.

[70] See, for example, John Stoessinger, *Why Nations Go to War* (New York: St. Martin's Press, 1997); Geoffrey Blainey, *The Causes of War* (New York: Free Press, 1988); Barry Posen, 'The security dilemma and ethnic conflict', *Survival*, 35:1 (1993), pp.

of major war between nuclear dyads and the concomitant propensity of nuclear states to engage in low-level aggression and arms competition.[71] Nuclear deterrence theorising is limited empirically by the scarcity of datasets on nuclear exchanges, making this method susceptible to accusations of retroactive reasoning, biased inference-dependent variables, and problematic rational-choice assumptions.[72] Similar to the first approach, war theorising relies too much on historical data and statistics risks; thus, under-utilising novelty is of limited use for policymakers concerned with the ways and means of future AI-enabled war.

The third approach used in this study is neither wedded to generalised theories, limited by data scarcity, nor concerned with predicting conflict. Like the first two approaches, it uses historical analysis, causal inference, and assumptions then combines these with imagination to produce plausible depictions of future conflict. Thus, it is the only approach that incorporates hitherto-unforeseen technological innovation possibilities to deduce the impact of these trends on the ways and means of future war.[73] Like the other two, this approach suffers several shortcomings. First, it uses non-falsifiable hypotheses. For instance, fictionalised accounts of future war rarely consider the strategic logic (or the endgame) once confrontation ensues. In the case of high-intensity AI-enabled hyper-speed warfare (see Scenario 1), while the pathway to war can be intellectualised, the catastrophic outcome of an all-out nuclear exchange between great powers, like recent conflicts in Ukraine and Gaza, is not apparent or straightforward to fathom. This challenge can produce unrealistic and contrived portrayals of future war.[74]

Second, groupthink, collective bias, and an overreliance on predetermined outcomes affect policy change. An oversimplistic view often substantiates this collective mindset that the course of future conflict can be extrapolated from present technological trends. Recent cognitive studies corroborate Scottish philosopher David Hume's hypothesis that individuals tend to favour what is already established. We mirror-project the 'status-quo with an unearned quality of goodness, in the absence of deliberative thought, experience or reason to do so'.[75] For instance, many depictions of future wars between the United States and China reflect the broader US national security discourse on Chinese security issues.[76] For example, the US think-tank community used the techno-centric novel *Ghost Fleet* to advocate accelerating President Obama's 'pivot' to Asia policy.[77]

Finally, it mistakenly views emerging technology (especially AI, autonomy, and cyber) as a deus ex machina possessing a magical quality, resulting in the fallacy that the nature of future wars will be entirely different from past and present ones.[78] This fallacy emerged during the first revolution in military affairs (RMA) in the late 1990s. The RMA was associated with the notion that technology

27–47; Michael C. Webb and Stephen D. Krasner, 'Hegemonic stability theory: An empirical assessment review of international studies', *Review of International Studies*, 15:2 (1989), pp. 183–98; John Lewis Gaddis, 'International Relations theory and the end of the Cold War', *International Security*, 17:3 (1992–3), pp. 5–58; James Goldgeier and Lily Wojtowicz, 'Reassurance and deterrence after Russia's war against Ukraine', *Security Studies*, 31:4 (2022), pp. 736–43.

[71] Aaron Bateman, 'Hunting the red bear: Satellite reconnaissance and the "second offset strategy" in the late Cold War', *The International History Review* (2024), available at: {https://doi.org/10.1080/07075332.2024.2406215}

[72] See Richard N. Lebow and Janice Gross Stein, 'Rational deterrence theory: I think, therefore I deter', *World Politics*, 41:2 (1989), pp. 208–24; Bailo and Goldsmith, 'No paradox here?'; Bradley MacKay and Peter McKiernan, 'The role of hindsight in foresight: Refining strategic reasoning', *Futures*, 36:2 (2004), pp. 161–79; Lieber and Press, *The Myth of the Nuclear Revolution*; Benoit Pelopidas, 'The unbearable lightness of luck: Three sources of overconfidence in the controllability of nuclear crises', *European Journal of International Security*, 2:2 (2017), pp. 240–62.

[73] See 'Proceed with caution: Artificial intelligence in weapon systems', UK House of Lords, AI in Weapon Systems Committee, Report of Session 2023–24, December 2023.

[74] Mark D. Jacobsen, 'The uses and limits of speculative fiction: Three novels about a U.S.–China war', *Journal of Indo-Pacific Affairs* (11 August 2023).

[75] Scott Eidelman, Christian S. Crandall, and Jennifer Pattershall, 'The existence bias', *Journal of Personality and Social Psychology*, 97:5 (2009), pp. 765–75 (p. 765).

[76] Eric Heginbotham, Michael Nixon, Forrest Morgan, et al., *The U.S.–China Military Scorecard: Forces, Geography, and the Evolving Balance of Power, 1996–2017* (Santa Monica, CA: RAND Corporation, 2015).

[77] Eric M. Murphy, '#Reviewing Ghost Fleet: Go back! It's a trap!', *The Strategic Bridge* (29 July 2015).

[78] H. R. McMaster, 'Discussing the continuities of war and the future of warfare', *Small Wars Journal* (14 October 2014).

would make future wars fundamentally different because advances in surveillance, information, communications, and precision munitions made warfare more controllable, predictable, and thus less risky. Several observers have made similar predictions about the promise of AI and autonomy.[79] Clausewitz warns there is no way to 'defeat an enemy without too much bloodshed … it is a fallacy that must be exposed'.[80] This fallacy is equally valid in the emerging AI–nuclear nexus, particularly the potential 'bloodshed' resulting from inadvertent and accidental nuclear use illustrated in the scenarios below.

However, the history of conflict demonstrates that technology is invariably less impactful (or revolutionary) than anticipated or hoped for. Most predictions of technological change have gone awry when they overestimate the revolutionary impact of the latest technical Zeitgeist and underestimate the uncertainties of conflict and the immutable political and human dimensions of war.[81] According to Major General Mick Ryan, Commander of the Australian Army Defence College, 'science fiction reminds us of the enduring nature of war', and 'notwithstanding the technological marvels of science fiction novels, the war ultimately remains a human endeavor'.[82] The scenarios in this article demonstrate that while the character of war continues to evolve, future AI-enabled warfare will continue to be human, arguably more so.[83] Moreover, absent an agenda or advocacy motive, the scenarios in this article offer a less US-centric depiction of future United States–China conflict, thus distinguishing itself from the broader US national security discourse about China.[84] Specifically, the scenarios pay close attention to strategic, tactical, and operational considerations and are mindful of US, Chinese, and Taiwanese domestic constituencies and political clocks (especially Scenario 2).

Fictional Intelligence (FICINT) is a concept coined by authors August Cole and P. W. Singer to describe the anticipatory value of fictionalised storytelling as a change agent. Cole and Singer describe FICINT as 'a deliberate fusion of narrative's power with real-world research utility'.[85] As a complementary approach alongside traditional methods such as wargaming, red-teaming, counterfactual thinking, and other simulations,[86] FICINT has emerged as a novel tool for conflict anticipation, gaining significant traction with international policymakers and strategic communities.[87] A comparative empirical exploration of the competing utility of the various approaches and tools to consider future warfare would benefit from further research.

[79] Owen J. Daniels, 'The AI "revolution in military affairs": What would it really look like?', *Lawfare* (21 December 2022).

[80] Carl von Clausewitz, *On War*, ed. and trans. Michael Howard and Peter Paret (Princeton, NJ: Princeton University Press, 1976), p. 75.

[81] Michael O'Hanlon, 'A retrospective on the so-called revolution in military affairs, 2000–2020', *Brookings* (September 2018).

[82] Mick Ryan and Nathan K. Finney, 'Science fiction and the strategist 2.0', *The Strategic Bridge* (27 August 2018).

[83] Peter L. Hickman, 'The future of warfare will continue to be human', *War on the Rocks* (12 May 2020); Avi Goldfarb and Jon Lindsay, 'Prediction and judgment, why artificial intelligence increases the importance of humans in war', *International Security*, 46:3 (2022), pp. 7–50.

[84] Jacobsen, 'The uses and limits of speculative fiction'.

[85] August Cole and P. W. Singer, 'Thinking the unthinkable with useful fiction', *Journal of Future Conflict*, online journal, no. 2 (2020).

[86] For recent uses of wargaming and other simulations (including the use of AI technology) to view future digitised warfare, see Paul Davis and Paul Bracken, 'Artificial intelligence for wargaming and modeling', *The Journal of Defense Modeling & Simulation* (2022), available at: {https://doi.org/10.1177/15485129211073126}; Jacquelyn Schneider, Benjamin Schechter, and Rachael Shaffer, 'Hacking nuclear stability: Wargaming technology, uncertainty, and escalation', *International Organization*, 77:3 (2023), pp. 633–67; Ivanka Barzashka, 'Wargames and AI: A dangerous mix that needs ethical oversight', *The Bulletin of the Atomic Scientists* (4 December 2023).

[87] Notable examples include the US Army Cyber Institute's adoption of FICIT for professional military education, the UK Ministry of Defence's use of FICIT as a tool to support the drafting of 'Global Britain in a Competitive Age: The Integrated Review of Security, Defence, Development and Foreign Policy'; and NATO's use of FICIT to develop its 'Four Worlds' future scenario model, which is designed to explore the future of war to assess the utility and effectiveness of the alliance's capabilities in the evolving security environment. *Invisible Force*; NATO's Strategic Warfare Development Command, 'Strategic foresight analysis 2023 (NATO Allied Command Transformation, 2023); UK Ministry of Defence, 'Stories from tomorrow: Exploring new technology through useful fiction', *Defence Science and Technology Laboratory* (28 February 2023).

Table 1. Effective scenario construction.

Tests to determine the validity and empirical robustness of fictional scenarios	Benchmarks for the construction of fictional scenarios
• Technologies and their usage are technically, politically, and operationally feasible	• Scenarios must incorporate falsifiable hypotheses that go beyond 'all other things being equal' – which is rarely the case in international affairs
• Considerations of alternative outcomes and wild-card events	• Scenarios must establish clear benchmarks for well-defined, consistent causal influences, mechanisms, and underlying assumptions
• The presence of path dependency (i.e. how past events or decisions can constrain later events or decisions)	
• Acknowledged and adequately accounted for underlying assumptions	

Because of cognitive biases – for example, hindsight and availability bias and heuristics – most conflicts surprise leaders but, in hindsight, appear eminently predictable.[88] Well-imagined fictional prototypes (novels, films, comics, etc.) conjure aesthetic experiences such as discovery, adventure, and escapism that can stimulate introspection, for instance, in the extent to which these visions correspond with or diverge from reality.[89] This thinking can improve policymakers' ability to challenge established norms, to project, and to prepare for change. Psychology studies have shown that stimulating people's imagination can speed up the brain's neural pathways that deal with preparedness in the face of unexpected change.[90] This focus can help policymakers recognise and respond to the trade-offs associated with technological change and thus be better placed to pre-empt future technological inflection points, such as the possibility of AI general intelligence (AGI) – or 'superintelligence'.[91] Political scientist Charles Hill argues that fiction brings us closer to 'how the world works … literature lives in *the realm strategy requires, beyond rational calculation*, in acts of the imagination' (emphasis added).[92]

Fictionalised future war in the Taiwan Straits

Much like historical counterfactuals, effective fictional scenarios must be empirically plausible (i.e. the technology exists or is being developed) and theoretically rigorous, thus avoiding *ex post facto* reasoning and bias extrapolations and inference (see Table 1).[93]

Well-crafted scenarios will allow the reader to consider questions such as under what circumstances might the underlying assumptions be invalid (if all things were not equal) and what would need to change. How might the scenarios play out differently (i.e. falsifiability), and to what end? What are the military and political implications of these scenarios? How might changing the causal influences and pathways change things? Moreover, the effective operationality of these criteria in fictional scenarios requires consideration of the following: the motivations for the

[88] Baruch Fischoff, 'Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty', *Journal of Experimental Psychology*, 1:2 (1975), pp. 288–99; Creveld, *Command in War*.

[89] Current AI approaches (i.e. Bayesian and non-Bayesian) that reason under uncertainty and consider multiple future worlds lack humans' ability for introspection and self-modification. Some observers predict that future artificial general intelligence (AGI) will possess this ability. See Douglas R. Hofstadter, *I Am a Strange Loop* (New York: Basic Books, 2007).

[90] Daniel Kahneman, *Thinking, Fast and Slow* (New York: Penguin, 2012).

[91] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

[92] Charles Hill, *Grand Strategies: Literature, Statecraft, and World Order* (New Haven, CT: Yale University Press, 2011), p. 6.

[93] For literature on the effective construction of counterfactual scenarios and thought experiments, see Philip E. Tetlock and Aaron Belkin (eds), *Counterfactual Thought Experiments in World Politics* (Princeton, NJ: Princeton University Press, 1996); Lebow, 'Counterfactuals and security studies'; James Johnson, 'Counterfactual thinking & nuclear risk in the digital age: The role of uncertainty, complexity, chance, and human psychology', *Journal for Peace and Nuclear Disarmament*, 5:2 (2022), pp. 394–421.

https://doi.org/10.1017/S0260210524000767

assumptions underlying the scenarios;[94] the influence of these assumptions on the outcome; and exploring alternative casual (i.e. escalation vs de-escalation) pathways.[95] AI-enhanced wargaming could also improve the insights derived from traditional wargaming and identify innovative strategies and actions, including comparing and contrasting multiple underlying assumptions, escalation pathways, and possible outcomes.[96]

Scenarios that adhere to these criteria are distinguished from science fiction and thus have anticipatory capabilities (but do not make concrete predictions) and policy relevance. Specifically, it aims to reflect on operational readiness and manage the risks associated with the intersection of nuclear weapons and AI-enabling weapon systems.[97] These benchmarks will need to be modified and refined as the underlying assumptions and causal influences change or competing scenarios emerge. In short, the quality of fictional scenarios intended to shape and influence the national security discourse matters.

These scenarios demonstrate how the use of various interconnected AI-enabling systems (i.e. command and control decision-making support, autonomous drones, information warfare, hypersonic weapons, and smart precision munitions) during high-intensity military operations under the nuclear shadow can inadvertently and rapidly escalate to nuclear detonation.[98] Further, they demonstrate how the coalescence of these AI-enabled conventional weapon systems used in conjunction during crises contributes to the stability–instability paradox by (a) reducing the time available for deliberation and thus the window of opportunity for de-escalation, (b) increasing the fog and friction of complex and fast-moving multi-domain operations,[99] and (c) introducing new and novel risks of misperception (especially caused by human-deductive vs machine-inductive reasoning) and accidents (especially caused by human psychological/cognitive fallibilities) in human–machine tactical teaming operations.[100] In sum, these features increase the risk of inadvertent and accidental escalation and, thus, undermine the red-line model of the paradox and support the brinkmanship-model interpretation.

Scenario 1: The 2030 'flash war' in the Taiwan Straits

Assumptions

- Great power competition intensifies in the Indo-Pacific, and geopolitical tensions increase in the Taiwan Straits.[101]

[94] Recent scholarship on nuclear brinkmanship's psychological and emotional features argues that human choice and agency are critical to nuclear escalation dynamics. See Pauly and McDermott, 'The psychology of nuclear brinkmanship'.

[95] 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*, 6:3 (2021), pp. 257–77.

[96] Anna Knack and Rosamund Powell, 'Artificial intelligence in wargaming: An evidence-based assessment of AI applications', *CETaS Research Reports* (June 2023).

[97] See Lawrence Freedman, *The Future of War: A History* (London: Allen Lane, 2017); Army Cyber Institute at West Point, *Invisible Force: Information Warfare and the Future Conflict* (New York: US Army Cyber Institute, West Point, 2020); Cole and Singer, 'Thinking the unthinkable with useful fiction'; Franz-Stefan Gady, 'The impact of fiction on the future of war', *The Diplomat* (7 December 2019).

[98] An expansive discussion on why and how states implement military AI and how AI can affect the operability of weapon systems is beyond the scope of this article. For recent scholarship on these issues, see 'Proceed with caution: Artificial intelligence in weapon systems'; James Johnson, 'Artificial intelligence & future warfare: Implications for international security', *Defense & Security Analysis*, 35:2 (2019), pp. 147–69; Michael Horowitz, 'Artificial intelligence, international competition, and the balance of power', *Texas National Security Review*, 1:3 (2018), pp. 36–57; Final Report, National Security Commission on Artificial intelligence, March 2021.

[99] See Goldfarb and Lindsay, 'Prediction and judgment'.

[100] James Johnson, *The AI Commander: Centaur Teaming, Command, and Ethical Dilemmas* (London, UK: Oxford University Press, 2024), chapters 2 and 3.

[101] 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://doi.org/10.1017/S0260210524000767

- An intense 'security dilemma' exists between the United States and China.[102]
- China's nuclear forces rapidly modernise and expand, and the United States accepts 'mutual vulnerability' with Beijing.[103]
- The United States and China deploy AI-powered autonomous strategic decision-making technology into their NC3 networks.[104]

Hypothesis

Integrating AI-enabled weapon systems into nuclear weapon systems amplifies existing threat pathways that, all things being equal, *increase the likelihood of unintentional atomic detonation*.

How could AI-driven capabilities exacerbate a crisis between two nuclear-armed nations? On 12 December 2030, leaders in Beijing and Washington authorised a nuclear exchange in the Taiwan Straits. Investigators looking into the 2030 'flash war' found it reassuring that neither side employed AI-powered 'fully autonomous' weapons or deliberately breached the law of armed conflict.

In early 2030, an election marked by the island's tense relations with Communist China saw President Lai achieve a significant victory, securing a fourth term for the pro-independence Democrats and further alienating Beijing. As the late 2020s progressed, tensions in the Straits intensified, with hard-line politicians and aggressive military generals on both sides adopting rigid stances, disregarding diplomatic overtures, and being fuelled by inflammatory rhetoric, misinformation, and disinformation campaigns.[105] These included the latest 'story weapons', autonomous systems designed to create adversarial narratives that influenced decision-making and stirred anti-Chinese and Taiwanese sentiments.[106]

Simultaneously, both China and the United States utilised AI technology to enhance battlefield awareness, intelligence, surveillance, and reconnaissance, functioning as early warning systems and decision-making tools to predict and recommend tactical responses to enemy actions in real time.

By late 2030, advancements in the speed, fidelity, and predictive power of commercially available dual-use AI applications had rapidly progressed.[107] Major military powers were compelled to provide data for machine learning to refine tactical and operational manoeuvres and inform

[102]For recent research on the nature and impact of the US–China security dilemma in the digital era, see James Johnson, 'The end of military-techno pax Americana? Washington's strategic responses to Chinese AI-enabled military technology', *The Pacific Review*, 34:3 (2021), pp. 351–78. For foundational security dilemma theorising, see Robert Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976).

[103]According to the 2023 US DoD Report to Congress on the People's Liberation Army, 'over the next decade, the PRC will continue to modernize, diversify, and expand its nuclear forces rapidly'. The report estimates that by 2030, China will possess over 1,000 nuclear warheads. United States Department of Defense, 'Military and security developments involving the People's Republic of China, 2023', *Annual Report to Congress* (October 2023).

[104]The US Defense Department, 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 electro-magnetic 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.' US Department of Defense, *2022 National Defense Strategy of the USA* (Washington DC: US DoD, 2022), p. 22. Open sources further indicate that China is also leveraging emerging technology, including AI, big data analytics, quantum computing, and 5G to prepare its force for future 'intelligentised' warfare at every level of warfare (including nuclear), and to enhance the People's Liberation Army's dual-use (conventional and nuclear) C2 architecture. Lora Saalman, 'China's integration of neural networks into hypersonic glide vehicles', in Nicholas D. Wright (ed.), *AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives*, White Paper (Washington, DC: US Department of Defense and Joint Chiefs of Staff, 2018), pp. 153–60.

[105]For example, during the recent Russian–Ukrainian conflict, both sides have deployed AI-assisted capabilities (defensively and offensively) in information operations, including deep-fake technology, AI-enhanced cyberattacks, disinformation, and intelligence-gathering operations. Will Knight, 'As Russia plots its next move, an AI listens to the chatter', Wired, (4 April 2022).

[106]'Storyweapons' are a new class of unregulated information warfare, adversarial narratives deployed across networks designed to manipulate decision-making. They can be deployed against entire populations on extended timeframes, trained on hacked personal and behavioural data, and iteratively refined with brute-force computing power. Future iterations will be more autonomous, targeted, and able to sustain campaigns for extended periods. *Invisible Force*, p. 24.

[107]Alyssa Schroer, '34 Artificial intelligence companies building a smarter tomorrow', *BuiltIn* (1 May 2019).

strategic choices. Noticing the successful early implementation of AI tools for autonomous drone swarms by Russia, Turkey, and Israel to counter terrorist incursions effectively, China rushed to integrate the latest versions of dual-use AI, often at the expense of thorough testing and evaluation in the race for first-mover advantage.[108]

With Chinese military activities – including aircraft flyovers, island blockade drills, and drone surveillance operations – representing a notable escalation in tensions, leaders from both China and the United States demanded the urgent deployment of advanced strategic AI to secure a significant asymmetric advantage in terms of scale, speed, and lethality. As incendiary language spread across social media, fuelled by disinformation campaigns and cyber intrusions targeting command-and-control networks, calls for China to enforce the unification of Taiwan grew louder.

Amid the escalating situation in the Pacific, and with testing and evaluation processes still incomplete, the United States decided to expedite the deployment of its prototype autonomous AI-powered 'Strategic Prediction & Recommendation System' (SPRS), which was designed to support decision-making in non-lethal areas such as logistics, cyber operations, space assurance, and energy management. Wary of losing its asymmetric edge, China introduced a similar decision-support system, the 'Strategic & Intelligence Advisory System' (SIAS), to ensure its readiness for any potential crisis.[109]

On 14 June 2030, at 06:30, a Taiwanese Coast Guard patrol boat collided with and sank a Chinese autonomous sea-surface vehicle that was conducting an intelligence reconnaissance mission within Taiwan's territorial waters. The day before, President Lai had hosted a senior delegation of US congressional staff and White House officials in Taipei during a high-profile diplomatic visit. By 06:50, the ensuing sequence of events – exacerbated by AI-enabled bots, deepfakes, and false-flag operations – exceeded Beijing's predetermined threshold for action and overwhelmed its capacity to respond.

By 07:15, these information operations coincided with a surge in cyber intrusions targeting US Indo-Pacific Command and Taiwanese military systems, alongside defensive manoeuvres involving Chinese counter-space assets. Automated logistics systems of the People's Liberation Army (PLA) were triggered, and suspicious movements of the PLA's nuclear road-mobile transporter erector launchers were detected. At 07:20, the US SPRS interpreted this behaviour as a significant national security threat, recommending an increased deterrence posture and a strong show of force. Consequently, the White House authorised an autonomous strategic bomber flyover in the Taiwan Straits at 07:25.

In reaction, at 07:35, China's SIAS alerted Beijing to a rise in communication activity between US Indo-Pacific Command and critical command-and-control nodes at the Pentagon. By 07:40, SIAS escalated the threat level concerning a potential US pre-emptive strike in the Pacific aimed at defending Taiwan and attacking Chinese positions in the South China Sea. At 07:45, SIAS advised Chinese leaders to initiate a limited pre-emptive strike using conventional counterforce weapons (including cyber, anti-satellite, hypersonic weapons, and precision munitions) against critical US assets in the Pacific, such as Guam.

By 07:50, anxious about an impending disarming US strike and increasingly reliant on SIAS assessments, Chinese military leaders authorised the attack, which SIAS had already anticipated and planned for. At 07:55, SPRS notified Washington of the imminent assault and recommended a limited nuclear response to compel Beijing to halt its offensive. Following a limited United

[108] A 2021 UN report on Libya stated that a Turkish Kargu-2 drone – which has both autonomous and manual functionality – was used for the first time to attack humans autonomously utilising the drone's AI capabilities. In the same year, in what is thought to be the world's first drone swarming in combat, the Israeli Defense Force used 'dozens' of drones in coordination with mortars and ground-based missiles to strike at Hamas combatants in the Gaza region. David Hambling, 'Israel used world's first AI-guided combat drone swarm in Gaza attacks', *New Scientist* (30 June 2021).

[109] For recent scholarship on the effects of automaticity of decision on human psychology during high-intensity conflict, see Pauly and McDermott, 'The psychology of nuclear brinkmanship'; Hunter and Bowen, 'We'll never have a model of an AI major-general'; James Johnson, 'Automating the OODA loop in the age of intelligent machines: Reaffirming the role of humans in command-and-control decision-making in the digital age', *Defence Studies*, 23:1 (2022), pp. 43–67.

States–China atomic exchange in the Pacific, which resulted in millions dead and tens of millions injured, both sides eventually agreed to cease hostilities.[110]

In the immediate aftermath of this devastating confrontation, which unfolded within mere hours, leaders from both sides were left perplexed about the origins of the 'flash war'. Efforts were undertaken to reconstruct a detailed analysis of the decisions made by SPRS and SIAS. Still, the designers of the algorithms underlying these systems noted that it was impossible to fully explain the rationale and reasoning of the AI behind every subset decision. Various constraints, including time, quantum encryption, and privacy regulations imposed by military and commercial users, rendered it impossible to maintain retrospective back-testing logs and protocols.[111] Did AI technology trigger the 2030 'flash war'?

Scenario 2: Human–machine warfighting in the Taiwan Straits

Assumptions

- Great power competition intensifies in the Indo-Pacific, and geopolitical tensions increase in the Taiwan Straits.[112]
- An intense security dilemma exists between the United States and China.
- China reaches a military capability sufficient to invade and seize Taiwan.[113]
- The United States and China have mastered AI-enabled 'loyal wingman' technology, and operational concepts exist to support them.[114]

Hypothesis

AI-enabled human–machine teaming operations used in high-intensity conflict between two nuclear-armed states, all else being equal, *increase the likelihood of unintentional nuclear use*.

How could AI-enhanced human–machine teaming influence a crisis between two nuclear-armed adversaries? In 2030, the ailing President Xi Jinping, eager to realise his 'China Dream' and secure his legacy in the annals of the Chinese Communist Party, invades Taiwan.

'Operation Island Freedom'

Chinese Air Force stealth fighters, dubbed 'Mighty Dragon', accompanied by a swarm of semi-autonomous AI-powered drones known as 'Little Dragons', launch cyberattacks and missile strikes to dismantle Taiwanese air defences and their command-and-control systems.[115]

[110] An alternative outcome could be one in which US and/or Chinese human commanders buy time for more deliberation and thus expand the window for crisis management decision-making, thus averting deterrence failure caused by miscalculations and errors made during decision-making compression. Space does not permit an exhaustive exploration of possible outcomes, wild card events, and alternative causal pathways.

[111] Quantum cryptography uses the laws of quantum physics to transmit private information, making undetected eavesdropping impossible. Researchers have demonstrated that quantum cryptography works, but technological bottlenecks prevent a system from transmitting data reliably and rapidly across long distances. 'How will quantum technologies change cryptography?' *Science Exchange, Caltech*.

[112] This scenario is adapted from sections of James Johnson, 'The challenges of AI command and control', *European Leadership Network* (12 April 2023).

[113] Recent reports have claimed that the US intelligence community believes that President Xi Jinping has ordered the military to be ready to annex Taiwan by 2027 (marking the PLA's centenary) and that a strategic window of opportunity exists between 2027 and 2030 when favourable conditions exist for China to unify Taiwan by force if peaceful means are not possible. Amy Hawkins, 'Taiwan foreign minister warns of conflict with China in 2027', *The Guardian* (21 April 2023).

[114] For example, the 'Ghost Bat' loyal wingman operates alongside human-crewed military aircraft to complement and extend airborne missions. Boeing Australia is developing an autonomous drone in collaboration with the Royal Australian Air Force (RAAF). It is expected to enter service with the RAAF by 2025. MQ-28A Ghost Bat Unmanned Aircraft, Australia, Airforce Technology, June 2023

[115] Marielle Descalsota, 'Take a look at the "Mighty Dragon", China's $120 million answer to the Lockheed Martin F-35 fighter jet', *Business Insider* (2 June 2022).

A semi-autonomous loitering system of 'barrage swarms' absorbs and neutralises most of Taiwan's missile defences, leaving Taipei almost defenceless against a military blockade imposed by Beijing.[116]

During this blitzkrieg assault, the 'Little Dragons' receive a distress signal from a group of autonomous underwater vehicles engaged in reconnaissance off Taiwan's coast, alerting them to an imminent threat from a US carrier group.[117] With the remaining drones running low on power and unable to communicate with China's command-and-control due to distance, the decision to engage is left to the 'Little Dragons', made without any human oversight from China's naval ground controllers.[118]

Meanwhile, the USS *Ronald Reagan*, patrolling the South China Seas, detects aggressive manoeuvres from a swarm of Chinese torpedo drones. As a precautionary measure, the carrier deploys torpedo decoys to divert the Chinese drones and then attempts to neutralise the swarm with a 'hard-kill interceptor'.[119] However, the swarm launches a fierce series of kamikaze attacks, overwhelming the carrier's defences and leaving it incapacitated. Despite the carrier group's efforts, they cannot eliminate the entire swarm, making them vulnerable to the remaining drones racing towards the mother ship.

In reaction to this bolt-from-the-blue assault, the Pentagon authorises a B-21 Raider strategic bomber to undertake a deterrent mission, launching a limited conventional counterstrike on China's Yulin Naval Base on Hainan Island, which houses China's submarine nuclear deterrent.[120] The bomber is accompanied by a swarm of 'Little Buddy' uncrewed combat aerial vehicles, equipped with the latest 'Skyborg' AI-powered virtual co-pilot,[121] affectionately nicknamed 'R2-D2'.[122]

Using a prioritised list of pre-approved targets, 'R2-D2' employs advanced AI-driven 'Bugsplat' software to optimise the attack strategy, including weapon selection, timing, and deconfliction measures to avoid friendly fire.[123] Once the targets are identified and the weapons selected, 'R2-D2' directs a pair of 'Little Buddies' to confuse Chinese air defences with electronic decoys and AI-driven infrared jammers and dazzlers.

With each passing moment, escalation intensifies. Beijing interprets the US B-21's actions as an attempt to undermine its sea-based nuclear deterrent in response to 'Operation Island Freedom'. Believing it could not afford to let US forces thwart its initial invasion success, China initiates a conventional pre-emptive strike against US forces and bases in Japan and Guam. To signal deterrence,

[116] David Hambling, 'China releases video of new barrage swarm drone launcher', *Forbes* (14 December 2020).

[117] Franz-Stefan Gady, 'How Chinese unmanned platforms could degrade Taiwan's air defence and disable a US navy carrier', *IISS* (9 June 2021).

[118] An alternative causal pathway is the use of man-in-the-loop semi-autonomous operations in which a human makes the final decision to launch an attack. While this pathway has some ethical and technical merit and support, the near-term trajectory suggests that once the remaining technical bottlenecks are overcome (battery power, swarming communications, operability, functionality in complex and dynamic environments, etc.), the tactical advantages of autonomous drone operations will likely trump any ethical considerations. See Anna Konert and Tomasz Balcerzak, 'Military autonomous drones (UAVs) – from fantasy to reality: Legal and ethical implications', *Transportation Research Procedia*, 59 (2021), pp. 292–9. For discussion on how and why militaries will likely delegate decision-making authority (either intentionally or inadvertently) to machines in future war, see Robert J. Sparrow and Adam Henschke, 'Minotaurs, not centaurs: The future of manned-unmanned teaming', *Parameters*, 53:1 (2023), pp. 115–30.

[119] Joseph Trevithick, 'The navy is ripping out underperforming anti-torpedo torpedoes from its supercarriers', *The Drive* (5 February 2019).

[120] Analysts estimate that the United States will have approximately 20 operational-ready B-21s by 2027, which could yield over 300 1,000-pound precision weapon strikes per day against Chinese maritime targets. Robert Haddick, 'Defeat China's navy, defeat China's war plan', *War on the Rocks* (21 September 2022).

[121] Robert Farley, 'A raider and his "little buddy": Which fighter will accompany the USAF's B-21?', *The Diplomat* (24 September 2016).

[122] Brett Tingley, 'Skyborg AI computer "brain" successfully flew a General Atomics Avenger drone', *The Drive* (30 June 2021).

[123] John Emery, 'Probabilities towards death: Bugsplat, algorithmic assassinations, and ethical due care', *Critical Military Studies*, 8:2 (2022), pp. 179–97.

China simultaneously detonates a high-altitude nuclear device off the coast of Hawaii, resulting in an electromagnetic pulse. Time is now critical.

This attack aims to disrupt and disable unprotected electronics on nearby vessels or aircraft while avoiding damaging Hawaii. It marks the first use of nuclear weapons in conflict since 1945. Due to a lack of understanding regarding each other's deterrence signals, red lines, decision-making processes, or escalation protocols, neither side could convey that their actions were intended to be calibrated, proportional, and aimed at encouraging de-escalation.[124]

Conclusion: A new Promethean paradox?

The article revisits Glenn Snyder's stability–instability paradox concept to explore the risks associated with the AI–nuclear nexus in a future conflict. The article applies the dominant red-line interpretation of the paradox to consider how AI-enabled weapon systems in nuclear dyads engaged in low-level conflict might affect the risk of nuclear use. It finds support for the non-dominant brinkmanship model of the paradox which posits that low-level military adventurism between nuclear-armed states increases the risk of inadvertent and accidental escalation to a nuclear level of war. The article combines empirically robust and innovative fictional scenarios to test and challenge the assumptions underlying the red-line model in a future Taiwan Straits conflict. The scenarios expand the existing evidentiary base that considers the paradox and build on the undertheorised brinkmanship model of the paradox.

Whereas conventional counterfactuals consider imaginary changes to past events (i.e. the past is connected by a chain of causal logic to the present), fictional scenarios, by contrast, confront a finite range of plausible alternative outcomes, wild-card events and the uncertainties that might arise from the possible combinations of them.[125] The finitude of possible futures also means that, unlike traditional empirical and theoretical studies, fictional scenarios are not constrained by *what we know happened*. Thus, well-constructed scenarios can sidestep problems such as over-determinism, path dependency, and hindsight bias. This analytical latitude also means scenarios can be designed as introspective learning tools to highlight policymakers' biases and challenge their assumptions and received wisdom – or Henry Kissinger's 'official future' notion.[126] The fictional prototypes illustrated in this article go beyond mere technological extrapolation, mirror-imaging, group-think, and collective bias and can, therefore, play a valuable role in shaping policymakers' perceptions and thus influence how they reason under uncertainty, grapple with policy trade-offs, and prepare for change.

Several additional policy implications exist for using fictional scenarios to inform and influence decision-making. First, the imagination and novelty of future warfare fiction (especially to reflect on possible low-probability unintended consequences of AI systems deployed in conventional high-intensity conflict) can be used to help design realistic national security wargames for defence, intelligence, and think-tank communities. The centuries-old art of wargaming has recently emerged as a critical thinking interactive teaching tool in higher education.[127] Moreover, using AI software alongside human-centric intelligent fiction might mitigate some limitations and ethical concerns surrounding using AI in strategic-level wargames.[128] Policymakers and military professionals can use well-crafted science fiction to expand the range and imaginative

[124]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', *Center for a New American Security* (15 June 2022).

[125]Johnson, 'Counterfactual thinking & nuclear risk in the digital age'.

[126]Henry Kissinger, *Diplomacy* (New York: Simon and Schuster, 1993).

[127]'Strategic studies wargaming club inaugural game: The Soviets are the winners!', *University of Aberdeen, School of Social Sciences* (2 March 2023).

[128]Knack and Powell, 'Artificial intelligence in wargaming'.

https://doi.org/10.1017/S0260210524000767

scope of wargaming scenarios (i.e. underlying assumptions, causal pathways, and outcomes) and counterfactual thinking about future warfare.

Second, fictional scenarios can be used with science and technology and studies scholarship to research further the extent to which policymakers' perceptions of AI-enabled weapon systems are being shaped by popular cultural influences such as intelligent fiction.[129] Fictional scenarios that meet the criteria outlined in this article might also be used by policymakers to positively shape the public 'AI narrative' to counter disinformation campaigns that manipulate the public discourse to confuse the facts (or 'post-truth'), sowing societal discord and distrust in decision-makers.[130]

Third, fiction is increasingly becoming accepted as an innovative and prominent feature of professional military education (PME).[131] The use of AI large language model (LLM) tools such as ChatGPT by authors to augment and complement human creativity (i.e. resonate with human experience, intuition, and emotion) may also expose officers' biases, encourage authors to think in novel and unconventional (and even non-human) ways, and identify empirical gaps or inconsistencies in their writing.[132] Inspiration from reading these scenarios can increase an officer's ability to appreciate complexity and uncertainty and thus better handle technological change and strategic surprise. Exposure to future warfare fiction can also remind officers that despite the latest technological 'silver bullet', warfare will remain (at least for the foreseeable future) an immutably human enterprise.[133] Developing fictional prototypes of future war can influence how officers think about using emerging technologies and how best to deploy them in new ways. Research on the effects of supplementing machine-learning supervised training with science-fiction prototypes (based on future trends in technology, politics, economics, culture, etc.) would be beneficial.

Finally, fictional future war scenarios may be used to train AI machine-learning algorithms to enhance their data's fidelity and strategic depth, which is currently restricted by historical datasets to identify patterns, relationships, and trends. Training AI on a combination of historical data and realistic future fictional scenarios – and then re-training them on new parameters and data as conditions change – could improve the quality and real-world relevance of machine learning to establish correlations and predictions. Enhanced datasets could, for instance, qualitatively improve the modelling and simulation that supports AI-enabled wargaming, making strategic-level (from current tactical-level wargaming) wargames more feasible and realistic.

The fire that Prometheus stole from the gods brought light and warmth, but it also cursed humankind with destruction. Just as nuclear weapons condemned the world to the perennial fear of Armageddon (as a paradoxical means to prevent its use), so AI, offering those who wield it the allure of outsized technological benefits, is potentially a new harbinger of a modern-day Prometheus. The quest for scientific knowledge and tactical mastery in the possession of AI technology risks Promethean overreach and unintentional consequences playing out under the nuclear shadow. In the age of intelligent machines, skirmishes occurring under the nuclear shadow, as predicted by Snyder, will become increasingly difficult to predict, control, and thus pull back from the brink.

[129] For example, academics at the Leverhulme Centre for the Future of Intelligence (a Leverhulme Trust-funded centre based at the University of Cambridge) have begun exploring such questions to understand better the cultural contexts influencing how AI is perceived and developed.

[130] *Invisible Force*, p. 37.

[131] Recent examples include the US Naval Institute and Center for International Maritime Security 'Fiction Essay Contest'; the US Marine Corps Warfighting Lab's graphic science fiction novel illustrations; the US Marine Corp Warfighting Lab and the Brute Krulak Center for Innovation & Future Warfare's 'Essay Contest'; and NATO Allied Command Transformation's programme future conflict in the 2030s and 2040s research. *Invisible Force*, p. 7.

[132] Robert A. Gonsalves, 'Using ChatGPT as a creative writing partner', *Towards Data Science* (4 January 2023).

[133] Ryan and Finney, 'Science fiction and the strategist 2.0'.

Video Abstract. To view the online video abstract, please visit: https://doi.org/10.1017/S0260210524000767.

Acknowledgments. The author would like to thank the three anonymous reviewers, the editors, and the participants of the 'Understanding Complex Social Behavior through AI Analytical Wargaming' workshop hosted by The Alan Turing Institute, the US Department of Defense Department of Research & Engineering, and the University of Maryland for their helpful comments and feedback.

James Johnson is Senior Lecturer in Strategic Studies in the Department of Politics and International Relations at the University of Aberdeen. He is the author of *The AI Commander: Centaur Teaming, Command, and Ethical Dilemmas; AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*; and *Artificial Intelligence and the Future of Warfare: USA, China & Strategic Stability*.

Cite this article: James Johnson, 'Revisiting the 'stability–instability paradox' in AI-enabled warfare: A modern-day Promethean tragedy under the nuclear shadow?', *Review of International Studies* (2024), pp. 1–19. https://doi.org/10.1017/S0260210524000767