Cambridge Forum on AI: Law and Governance (2025), **1**, e56, 1–15
doi:10.1017/cfl.2025.10021

RESEARCH ARTICLE

Waltzing into uncertainty: AI in nuclear decision making and the challenge of divergent deterrence logics

Luba Zatsepina

International Relations and Politics, Faculty of Arts Professional and Social Studies, Liverpool John Moores University, Liverpool, UK
Email: l.zatsepina@ljmu.ac.uk

(Received 20 January 2025; revised 6 June 2025; accepted 24 August 2025)

Abstract

This article critically examines the integration of artificial intelligence (AI) into nuclear decision-making processes and its implications for deterrence strategies in the Third Nuclear Age. While realist deterrence logic assumes that the threat of mutual destruction compels rational actors to act cautiously, AI disrupts this by adding speed, opacity and algorithmic biases to decision-making processes. The article focuses on the case of Russia to explore how different understandings of deterrence among nuclear powers could increase the risk of misperceptions and inadvertent escalation in an AI-influenced strategic environment. I argue that AI does not operate in a conceptual vacuum: the effects of its integration depend on the strategic assumptions guiding its use. As such, divergent interpretations of deterrence may render AI-supported decision making more unpredictable, particularly in high-stakes nuclear contexts. I also consider how these risks intersect with broader arms race dynamics. Specifically, the pursuit of AI-enabled capabilities by global powers is not only accelerating military modernisation but also intensifying the security dilemma, as each side fears falling behind. In light of these challenges, this article calls for greater attention to conceptual divergence in deterrence thinking, alongside transparency protocols and confidence-building measures aimed at mitigating misunderstandings and promoting stability in an increasingly automated military landscape.

Keywords: deterrence theory; AI; nuclear weapons; arms race; Russia

1. Introduction[1]

Among the foundational perspectives on stability and deterrence in a nuclear-armed world, Kenneth Waltz's optimistic take on nuclear proliferation remains one of the most enduring. Waltz (1990, p. 737) argued that nuclear weapons are the ultimate guarantors of a state's survival as they induce caution and restraint. In his view, more nuclear weapons could enhance strategic stability between superpowers. Although nothing is guaranteed, Waltz (1979, p. 185) argued that nuclear weapons make wars less likely as they are associated with deterrent strategies that promise less damage than war-fighting strategies. These strategies call for caution and hence reduce the incidence of war. The fear of escalation and the disadvantages of striking first mean that nuclear weapons reverse the logic of conventional wars. Thus, based on "easy calculations of what one country can do to another," rational

[1] This is one of fourteen articles published as part of the *Cambridge Forum on AI: Law and Governance* Special Issue, *AI and the Decision to Go to War*, guest edited by Toni Erskine and Steven E. Miller.

actors will avoid armed nuclear conflicts at all costs (Waltz, 1990, p. 734). These assumptions underpinned what some scholars have called the First Nuclear Age (Futter & Zala, 2021, p. 260). Beginning in 1945, it brought forth an era of unprecedented strategic calculus, one dominated by great power competition and deterrence based on the condition of mutually assured destruction. For Waltz and others, the immense destructiveness of nuclear weapons and the fear they induced became cornerstones of strategic stability – understood here as the absence of incentives for nuclear first use and the absence of incentives to engage in nuclear arms racing (Acton, 2013) – shaping decades of nuclear politics and international relations: "The probability of major war among states having nuclear weapons approaches zero" (Waltz, 1988, p. 627).

Following the Cold War, the so-called Second Nuclear Age saw new challenges emerge, including additional threats from non-state actors and the rise of Asian military powers (Bracken, 2000, p. 146). Today, as we are said to be entering a Third Nuclear Age, a new set of pressures threatens to upend the strategic stability assumed by earlier deterrence frameworks. These include, *inter alia*, multipolarity, emerging technologies, the development of strategic non-nuclear weapons (SNNW) and precarious conditions for advancing arms control and disarmament (Crilley, 2023; Favaro et al., 2022; Futter & Zala, 2021; Zala, 2024). The rapid advancement and potential integration of artificial intelligence (AI) into the nuclear weapons domain are among these contemporary anxieties and challenges.

As AI technologies increasingly permeate military strategies, they introduce complexities and uncertainties that challenge the assumption that human rationality can ensure control and restraint in nuclear decision making (Depp & Scharre, 2024; Horowitz, 2018; Johnson, 2021, 2022, 2023; Kroenig, 2021; Nadibaidze & Miotto, 2023; Parke, 2023; Zala, 2024). The automation of threat detection, acceleration of decision timelines and potential deployment of autonomous weapons systems amplify the risks of miscalculations, misperceptions and inadvertent escalations (Johnson, 2023, p. 73; Price et al., 2018, pp. 101–102). Within this context, the integration of AI into states' nuclear deterrence architecture, particularly early warning systems; nuclear command, control and communications (NC3); and intelligence, surveillance and reconnaissance (ISR), constitutes a critical yet relatively neglected discourse on AI's influence over resort-to-force decision making (Erskine & Miller, 2024, p. 137). AI's role in shaping state decisions to go to war introduces both significant risks and potential opportunities, which calls for a thorough examination of its impact on deterrence theory and future crisis management (Erskine & Miller, 2024, p. 139). Moreover, a critical question arises: how will the integration of AI affect stability in a world where major nuclear powers do not share the same conceptual foundation for deterrence?

In this article, I retain Waltzian logic to question whether more AI equals more stability, but challenge the notion that technological advancement leads to restraint. I argue that the integration of AI into nuclear decision-making processes risks undermining the rational deterrence logic that has historically underpinned strategic stability. The main contribution of this article is to demonstrate that divergent understandings of deterrence, particularly between Russia and the West, are likely to be amplified rather than resolved by AI-driven systems. When such conceptual gaps are embedded in opaque, rapid and data-driven technologies, the risk of misperception, miscommunication and inadvertent escalation grows. Adopting a realist-informed, but critically reflective, approach and using Russia as a case study, this article shows how conceptual differences in deterrence thinking complicate crisis management and challenge the foundational assumptions of nuclear restraint in the Third Nuclear Age. This article contributes to the ongoing debates on AI and resort-to-force decision making by foregrounding conceptual divergence as a key source of risk and uncertainty. Moreover, it explores how the integration of AI may intensify the security dilemma and fuel arms race dynamics, as states seek to pre-empt, catch up with, or outmatch adversaries.

This article adopts a qualitative, interpretivist methodology grounded in discourse analysis. Russia is selected as a case study because it represents a particularly distinct and influential alternative to Western deterrence logic, grounded in a tradition of coercive signalling, reflexive control, doctrinal ambiguity and early escalation. It draws on a close reading of official Russian security and military doctrine, statements by political and military leaders and secondary literature on Russian strategic thought and deterrence theory. The analysis is conceptual in nature and does not make empirical claims about the operational integration of AI into Russia's nuclear decision-making processes. Instead, it traces how underlying deterrence logics may shape the development and use of AI-enabled decision support tools in the nuclear domain.

By focusing on the epistemological foundations of deterrence, the article seeks to illustrate how divergent strategic assumptions may undermine shared understandings in crisis situations, particularly when automated or mediated through AI systems. Johnson (2020b) cautions that any discussion of emerging technologies, such as AI, necessarily involves a degree of speculation, given the limited empirical data on how AI might influence deterrence, escalation and crisis decision making in real-world nuclear contexts (p. 426). While insights from strategic war gaming, nuclear decision-making simulations and expert analysis offer valuable contributions, much of the current discourse remains conceptual. This article reflects that epistemic uncertainty. It does not aim to predict outcomes with certainty, but rather to illuminate underexplored risks stemming from divergent deterrence logics and to encourage more nuanced thinking about the conditions under which AI integration might destabilise resort-to-force decision making.

To advance my argument, I first examine how AI challenges traditional nuclear deterrence frameworks by introducing new risks such as misperception, acceleration of decision timelines, and inadvertent escalation. I also consider its limited stabilising potential. I then examine Russian deterrence thinking to illustrate how divergent strategic logics, when combined with AI integration, can intensify uncertainty and raise the likelihood of crisis instability. The third section explores how AI is fuelling a new arms race and focuses on how doctrinal asymmetries, the erosion of arms control, and recursive technological competition create a volatile strategic environment and security dilemma. I conclude by stressing the need for renewed transparency, confidence-building measures, and regulatory frameworks to mitigate the risks posed by AI's integration into nuclear decision making.

2. AI in nuclear deterrence

International Relations scholars have spent decades debating the flaws of rational choice theories and neorealist thinking during the Cold War (see, e.g., Hymans, 2006; Jervis, 1989; Knopf, 2010; Kroenig, 2015; Lebow & Stein, 1989; Sagan, 1996; Tannenwald, 2007). Changes to the geopolitical and technological landscape, as well as the emergence of new security threats and domains, make current deterrence theorising an increasingly complex endeavour (Johnson, 2023, p. 76). Yet there is a need to acknowledge the undeniable fact that there has been no nuclear exchange during the Cold War or in the present circumstances, despite the ongoing war in Ukraine and increasingly hostile relations between Vladimir Putin's Russia and the North Atlantic Treaty Organization (NATO) states. Nuclear-armed states continue to rely on the principles of deterrence to maintain strategic stability. This reliance suggests that, flawed as it may be, deterrence theory continues to provide a practical – if contested – logic of restraint in nuclear politics, or at the very least, that states behave as though these principles are still effective.

If stable deterrence is the condition to be preserved, the integration of AI into nuclear deterrence architectures poses significant challenges and uncertainties. Zala (2024) notes that placing AI developments within the context of a larger shift towards a Third Nuclear Age can help generate a more nuanced understanding of the primary drivers of stability and instability in the current environment (p. 155). He identifies the loss of human oversight in nuclear and SNNW decision making, as well as machine-informed human decision making, as two areas that contain multiple risks to the stability of deterrence (pp. 156–158). These risks are especially salient given the compressed decision timelines, algorithmic biases, and possible overconfidence in AI-generated recommendations, which may compound misperceptions and increase the likelihood of escalation (Wong et al., 2020, p. 83).

A key concern lies in the historical opacity of nuclear decision making. While we are aware of high-profile incidents such as Able Archer 83 and the 1983 Petrov incident, these represent only a fraction of the potential close-calls and decision-making scenarios involving nuclear weapons that may remain classified or undisclosed. Lewis, Williams, Pelopidas and Aghlani (2014) examine some of these historical cases and more recent incidents, arguing that the risks of inadvertent nuclear use are higher than previously thought and calling for vigilance and prudent decision making (p. 30). The secrecy surrounding these events and the absence of systematic data hamper our ability to understand the real drivers of restraint and escalation. AI can only be as useful or accurate as the underlying data, and unknowns are difficult to include in its modelled assumptions (Holmes & Wheeler, 2024, p. 170). Its recommendations based on systematically incomplete data around nuclear scenarios might therefore lead to biased decision making, but without the awareness that significant gaps exist. Arguably, allowing AI to partake in these decisions to reduce human error and emotion, while increasing the rational use of (incomplete) data, may set an even more dangerous precedent (Johnson, 2022, p. 359). Ironically, in this case, optimising rationality could leave us in a more irrational and precarious situation.

This article focuses primarily on inadvertent escalation, which occurs when one side's intentional actions are unintentionally perceived as escalatory, often due to misperceptions about an opponent's likely reaction (Johnson, 2022, pp. 340–341). It is important to distinguish this from accidental and deliberate escalation. Accidental escalation is also unintended but results from events that were not planned or foreseen, such as accidents or technical malfunctions (Morgan et al., 2008, p. 26). Deliberate escalation, by contrast, involves one side intentionally crossing an escalatory threshold and viewing such actions as rational or strategically necessary (Hoffman & Kim, 2023, p. 18). AI could plausibly influence all three types of escalation. For instance, systems trained on inadequate data or applied to inappropriate contexts could inject flawed information into decision-making processes and raise the risk of accidental escalation (Hoffman & Kim, 2023, pp. 18–19). Deliberate escalation may also be shaped by AI-generated recommendations that promote coercive or escalatory actions, particularly under time pressure or uncertainty (Hoffman & Kim, 2023, p. 20). However, this article concentrates on inadvertent escalation, which represents a distinct and pressing risk in an AI-integrated environment. Here, the danger does not stem from technical failure but from systems operating as intended while misinterpreting signals (Johnson, 2022, p. 358).

If AI systems increasingly take on roles traditionally managed by human decision makers, the potential for miscommunication and misinterpretation grows. Reducing human oversight diminishes the capacity for ethical judgement and contextual interpretation, while increasing reliance on "black box" systems that may be vulnerable to cyber intrusion or failure (Johnson, 2023, p. 79). In a potential nuclear crisis, the already tight timescales for deciding whether to launch a nuclear strike will become even more compressed, placing additional pressure on leaders and increasing the risk of miscalculation (Parke, 2023). Such vulnerabilities could be exploited by adversaries and potentially trigger unintended confrontations through AI-driven responses, especially in the absence of robust crisis management mechanisms to de-escalate the situation (Hoffman & Kim, 2023, pp. 16–17).

Against this backdrop, some scholars have argued that AI, if properly developed and constrained, could theoretically stabilise deterrence relationships. For example, Black et al. (2024) argue that AI systems, when designed and deployed responsibly, have the potential to improve decision making by offering faster and more accurate analyses of complex situations, thereby reducing the likelihood of misinterpreting an adversary's actions (pp. 49–52). Boulanin (2019) similarly suggests that advanced AI models could avoid past technical failures by improving the reliability of threat detection (p. 54). Others point to the potential use of AI in ISR and manoeuvre planning as a means to deter first strikes by improving detection and defensive posturing (Zala, 2024, p. 159). Moreover, Holmes and Wheeler (2024) propose that AI could enhance strategic communication and high-level diplomacy by stimulating a broader range of scenarios and helping actors better understand adversaries' motives and potential red lines (p. 167). In a similar vein, McDonnell et al. (2023) suggest that AI-based

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

simulation and training tools could be used to train political and military leaders and their staff to make decisions in times of war and crisis and to practice navigating difficult choices (p. 34). In theory, this capability could augment human-to-human communication and reduce ambiguity during moments of heightened tension. AI might also assist arms control verification efforts and contribute to improving compliance and transparency (Schörnig, 2022).

However, all these stabilising scenarios rely on ideal conditions such as transparent algorithms, commonly agreed protocols, and mutual trust, which are seldom met in practice. In the context of the Third Nuclear Age, characterised by deep mistrust, multipolarity and the erosion of arms control and non-proliferation norms, AI-enhanced systems are far more likely to become sources of suspicion than instruments of collaboration. Thus, while AI might offer theoretical stabilising functions, the weight of evidence and contextual analysis points towards its destabilising effects. Without robust oversight, interoperable systems or mutual trust, the same characteristics that promise stability – speed, precision and automation – risk becoming liabilities (McDonnell et al., 2023, pp. 36–37). In other words, we return to the Waltzian logic: does more AI equal more stability? Under current conditions, the answer is no.

This risk is further complicated by divergent conceptual understandings of "deterrence." States do not define or operationalise deterrence in identical ways. These frameworks shape how AI systems are designed and deployed. Johnson (2020a) observes that AI's impact on escalation and deterrence is influenced less by technical capacity than by how its purpose is perceived (p. 16). In other words, what AI is designed to deter or signal depends on the deterrence logic of the state employing it. This has implications for how its recommendations are interpreted and responded to by adversaries.

Work on alternative deterrence traditions has shown that some states adopt more punitive or coercive models of deterrence, while others lean towards defensive or normative forms (Adamsky, 2018, 2024; Charap, 2020; Chase & Chan, 2016; Veebel, 2021; Ven Bruusgaard, 2016). For example, Veebel (2021) argues that while Western conceptions of deterrence often rely on normative frameworks and economic statecraft (e.g., sanctions), Russia's approach is more coercive and militarised, reflecting divergent understandings of how threats and responses should be structured. Adamsky (2018) similarly shows how Russian deterrence discourse is rooted in its strategic culture and historical emphasis on pre-emption and escalation control, noting that strategic cultures are not universally shared or understood (pp. 34–35). When these models interact, especially in a high-tech environment, the epistemological foundations of deterrence and interpretations may diverge.

These conceptual differences raise important questions about how AI systems, shaped by distinct deterrence logics, may misread signals and contribute to unintended escalation. AI systems trained and implemented according to different strategic doctrines may not interact in predictable or stabilising ways. The growing literature on the "conceptual multiplicity" of deterrence (see, e.g., Adamsky, 2024; Johnson, 2020b; Ven Bruusgaard, 2024) suggests that AI's effects must be examined within the interpretive frameworks that guide strategic thinking in each state.

This article contributes to this debate by explicitly foregrounding conceptual divergence as a key source of risk and uncertainty in the AI-nuclear nexus. Rather than treating deterrence as a stable or universal framework, it situates AI integration within a world of epistemological friction, where deterrence is interpreted, operationalised and potentially automated in different ways.

3. Russia's deterrence logic and the AI challenge

Russian approaches to containing, deterring, and inflicting varying levels of damage on adversaries are commonly grouped under the umbrella of "strategic deterrence" (Charap, 2020). This military-theoretical concept, first formally introduced in the 2010 Military Doctrine, refers to a continuous, multidomain activity designed to influence adversary behaviour in both wartime and peacetime through a combination of military and non-military measures (Akimenko, 2021, p. 2; President of

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

Russia, 2010). The Russian Ministry of Defence's Military Encyclopaedia defines "strategic deterrence" as "a coordinated system of use-of-force and non-use-of-force measures taken consecutively or simultaneously by one side in relation to another to keep the latter from any military actions that inflict or may inflict damage on the former on a strategic scale" (cited in Akimenko, 2021, p. 3).

The concept has continued to evolve in subsequent doctrinal texts and in Russian strategic analysis and military thought. The 2014 Military Doctrine broadened strategic deterrence to include not only nuclear weapons but also precision-guided conventional weapons, information operations, and threats emerging in new domains such as cyberspace and outer space (Security Council of the Russian Federation, 2014). The 2020 Fundamentals of State Policy on Nuclear Deterrence added further operational detail by clarifying conditions for nuclear use, highlighting the role of early warning and command structures, and reinforcing Russia's right to respond to non-nuclear threats that endanger state survival (President of Russia, 2020). Most recently, the 2024 Fundamentals reaffirmed the integration of nuclear and non-nuclear means, emphasising uncertainty, situational adaptation, and the role of information confrontation in shaping deterrence credibility (President of Russia, 2024). Reflecting this evolution, Dmitri Trenin, a prominent Russian expert and member of Russia's Foreign and Defence Policy Council, writes, "Strategic deterrence includes a military component (nuclear and conventional), a spatial dimension (geopolitics, geoeconomics, and other functional domains such as the cyber environment, space, etc.), and a coalition component (cooperation with friendly states). The complexity and systemic nature of strategic deterrence are essential conditions for its effectiveness" (Trenin, 2024).

This expanding doctrinal scope points to a deeper conceptual divergence between Russian and Western understandings of deterrence. The Russian term for strategic deterrence – *strategicheskoe sderzhivanie* – is more accurately translated as "strategic restraining" or "holding back," placing emphasis on limiting, restraining, or pre-empting aggressive action by an adversary (Ven Bruusgaard, 2016, p. 8). This emphasis on restraining marks a significant conceptual and etymological departure from the English-language notion of deterrence, which centres on threatening a response to dissuade hostile acts. As Kofman (2024) notes, while Western deterrence emphasises the threat of action, Russian *sderzhivanie* implies the initiation of proactive or preventive measures. Adamsky (2024) reinforces this interpretation:

> The connotation of concentrated effort, proactive endeavor, and preemptive action, which underlies the meaning of Russian *sderzhivanie*, is more straightforward, embedded, and explicit than in the case of English *deterrence*, where fear is the central motif; in the latter the threat of the use of force is implicit rather than explicit, almost spelled out, as it is in *sderzhivanie*. (p. 30)

Russia's concept of strategic deterrence is significantly broader and more comprehensive than traditional Western understandings of deterrence (Fink, 2023, p. 11; Ven Bruusgaard, 2016, p. 8). It integrates military and non-military, offensive and defensive, nuclear and conventional measures into a holistic framework aimed at shaping adversary decision making. This strategy spans a wide spectrum of actions from diplomatic signalling to the potential use of nuclear weapons and is designed to deter or coerce by presenting credible threats across multiple domains (Kofman et al., 2020, pp. 12–16).

A defining feature of Russia's strategic deterrence approach is its emphasis on psychological and informational effects, as well as the cultivation of ambiguity around escalation thresholds (Fink, 2017). The updated 2024 Fundamentals document reiterates the integration of non-nuclear and nuclear capabilities into a unified deterrence framework and highlights the importance of maintaining the adversary's "uncertainty regarding the scale, time and place of retaliatory action" (President of Russia, 2024). The document further underlines the role of conventional precision-strike weapons, information confrontation, and situational adaptation as core components of deterrence strategy.

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

These developments reinforce a deterrence logic that privileges ambiguity and favours proactive escalation management, departing sharply from the Western preference for transparency and mutual vulnerability (Fink, 2017).

Such a framework poses challenges for Western states, which often misread or mischaracterise Russian deterrence behaviour. As Adamsky (2018) observes, "the West engages Moscow with only a vague understanding of the conceptual foundations and perception of Russian strategists" (p. 56). This is compounded by what Wachs (2023) describes as the tendency of Western analysts to "mirror-image" Russian strategic thinking, presuming universality in concepts of security and deterrence that do not align (p. 175). Kofman and Fink (2020) highlight this disconnect by pointing to persistent Western assumptions that Russia exploits a "yield gap" to create escalation dilemmas, despite the absence of such logic in the Russian Military Doctrine. In reality, Russia's use of lower-yield nuclear options is embedded in an escalation management strategy aimed at preserving flexibility in regional theatres, not at provoking unresolvable dilemmas for Washington. Kofman and Fink (2020) explain: "Russian strategy has not been based on the premise that the United States is hamstrung by an asymmetry of yields," but rather on the understanding that U.S. interests in such conflicts are limited and geographically distant, which are factors that constrain its willingness to escalate.

Such doctrinal mismatches are already dangerous in human-to-human dynamics. But I argue that when automated or AI-supported systems begin interpreting adversary signals through mismatched frameworks, the likelihood of misperception and inadvertent escalation increases significantly. AI systems trained on flawed or incomplete assumptions about adversary doctrine may amplify, rather than mitigate, existing tensions.

Russian political and military leaders have repeatedly framed AI not merely as a technological innovation but as a critical enabler of national power and strategic sovereignty. President Putin's frequently cited 2017 statement that whoever leads in AI will become "the ruler of the world" is emblematic of this mindset (The Economist, 2024). Subsequent official rhetoric has gone further, emphasising AI's specific role in national defence, deterrence, and information confrontation (Bendett, 2024, pp. 3–4). For example, at a Defence Board Meeting in December 2022, Putin emphasised the need to integrate AI technologies across all tiers of military decision making (President of Russia, 2022). The former Minister of Defence Sergey Shoigu also remarked that it was "now necessary to ensure the integration of artificial intelligence technologies into the armaments that will define the future form of the Armed Forces" (cited in Petrov, 2021). Similarly, a group of experts from Russia's Defence Ministry's Center for Research of Foreign Countries Capabilities stated in 2021:

> In any case, to ensure the security of the Russian Federation, it is necessary to provide support for decision making regarding the use of strategic nuclear forces, definitely using AI as a tool for analysing the dynamically changing geopolitical and military situation and leaving the final decision-making power to the relevant officials. (cited in Shakirov, 2023, p. 18)

Such remarks reflect more than just technological enthusiasm; they signal an understanding of AI as a force multiplier within Russia's broader strategic deterrence posture (Clapp, 2025, p. 3).

This vision of AI as a strategic asset is shaped by long-standing Russian ideas about control, information dominance, and psychological pressure in warfare. As scholars such as Adamsky (2024, p. 68) and Merriam (2023, p. 8) have noted, Russia's military science embraces the principle of "reflexive control," a process of altering the information environment to shape an adversary's decisions by influencing their perceptions, strategic choices, and operational behaviour. To be successful, reflexive control requires a deep understanding of how an opponent thinks, processes information, and makes decisions (Thomas, 2019, pp. 4–11). To Russian strategists, reflexive control is not simply about deception or manipulation but about constructing a reality in which an opponent's behaviour aligns with Russia's strategic preferences. While it may resemble Western logic of coercion, reflexive control operates through the internal logic of the adversary's own decision-making process rather than

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

through external threats, which highlights a deeper cognitive approach often overlooked in Western frameworks (Adamsky, 2024, p. 69).

Building on this cognitive foundation, Russia's growing interest in applying AI to its nuclear command, control, and decision-making infrastructure reflects a broader effort to increase the speed, accuracy, and resilience of deterrence operations under crisis conditions. As Bendett (2024) notes, despite the classified nature of much of this work, publicly available information suggests that the Russian military is actively pursuing AI applications across a range of nuclear-related domains – from intelligence and surveillance to command support and battle damage assessment (p. 8). These developments reflect an ambition to leverage AI not only for enhanced situational awareness but also for reducing decision timelines, automating key data processing tasks, and supporting escalation management (Stokes et al., 2025, pp. 11–12). In this context, AI may offer tools not only for enhanced situational awareness but also for reducing decision timelines, automating data analysis, and supporting escalation management. Yet it also introduces significant risks. When adversaries lack a shared understanding of the frameworks guiding each other's use of AI, particularly in nuclear contexts, the potential for misreading intentions grows. Rather than stabilising deterrence, AI-enabled reflexive control may increase the likelihood of inadvertent escalation through conceptual misalignment and interpretive asymmetry.

These risks are especially pressing at the command level, where AI is increasingly envisioned as a support mechanism for real-time decision making. According to Bendett (2024), Russian military planners view AI as a means of synthesising political-military information rapidly during fast-evolving situations at tactical, operational, and strategic levels (p. 6). This may involve modelling crisis trajectories, visualising force postures, and generating decision options for senior leaders. AI applications are also reportedly being explored for integrating diverse data streams, including satellite imagery, sensor feeds, and open-source intelligence, into a unified operational picture (Boulanin et al., 2020, pp. 49–51). Such capabilities are seen as crucial for shortening decision timelines and improving the responsiveness of command structures in dynamic or ambiguous scenarios. These tools could become integral to the functioning of Russia's National Defence Coordination Centre (NDCC), which is already tasked with coordinating national security operations across ministries and military branches (Bendett, 2024, p. 5).

Yet, the very qualities that make AI appealing for managing fast-moving crises – speed, automation, and the reduction of complexity – also introduce serious concerns regarding inadvertent escalation. These same features may narrow the window for reflection and increase the likelihood of misinterpreting adversary intentions (Johnson, 2022, pp. 349–351). In a high-stress, time-compressed environment, a machine-generated recommendation could be seen as authoritative, especially if it aligns with existing biases or appears more objective than human judgement (McDonnell et al., 2023, pp. 23–24). This is especially problematic given the conceptual mismatches already identified between Russian and Western notions of deterrence. AI systems that process input according to one framework may fundamentally misinterpret the signals or posture changes made by a rival operating from a different logic.

These conceptual discrepancies become even more destabilising when filtered through the lens of AI-enabled decision support. In an AI-driven strategic environment, the risk of misperception is magnified. As Wong et al. (2020) note, AI systems can exacerbate the difficulty of interpreting adversary intent, particularly in ambiguous or rapidly changing circumstances (pp. 66–67). The "black box" nature of many machine learning systems, meaning their inability to clearly explain how conclusions are reached, further complicates matters. According to Bellaby (2024), AI's internal reasoning is often opaque even to its operators, raising serious concerns about accountability and trust, especially in nuclear decision-making contexts (p. 2536).

The unpredictability of AI-generated behaviour in high-stakes scenarios adds another layer of risk. In a study of AI agents in war game simulations, Rivera et al. (2024) observed a recurrent tendency towards escalation. In some cases, the AI adopted arms-racing behaviours or even initiated nuclear

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

weapon use, which illustrates how autonomous systems might misinterpret uncertainty or strategic ambiguity as a rationale for pre-emption (p. 836). While such experiments are limited and stylised, they highlight the hazards of entrusting escalation-sensitive decisions to systems that lack contextual understanding and operate under hard-coded assumptions.

These dynamics undermine the theoretical foundations of deterrence, which depend not only on rationality, signalling clarity and mutual understanding but also on a deep grasp of the other side's interests, perceptions and strategic priorities (Johnson, 2021, p. 424). AI introduces volatility into a system meant to produce caution and restraint. The integration of AI into nuclear planning, early warning, ISR and NC3 architectures across nuclear-armed states intensifies this problem (Johnson, 2021, p. 431). Several experts note that the modernisation efforts underway in Russia, China and the United States increasingly prioritise AI capabilities in ways that extend across their full deterrence architectures (Boulanin et al., 2020, ch. 3; Chernavskikh, 2024, pp. 4–5). Each of these developments, however rational in isolation, risks triggering reciprocal actions, fuelling a new kind of arms race not solely about warheads or delivery systems, but about cognitive and informational dominance. In this sense, integrating AI into nuclear decision making not only complicates deterrence but also accelerates a broader technological competition that heightens instability. This evolving landscape bears the distinctive hallmarks of the Third Nuclear Age: a multipolar order, disruptive technologies and increasingly unstable feedback loops between perception and action (Zala, 2019, p. 42). Today's nuclear environment is marked by sharp doctrinal and informational asymmetries between global superpowers. This section has shown that when AI is layered onto these conceptual mismatches, the danger is not only technical malfunction or miscommunication, but the potential for misperception and inadvertent escalation. These dynamics are already intensifying a competitive push among major powers to develop and integrate AI across their nuclear domains setting the stage for a new and potentially more dangerous arms race.

4. Is AI fuelling a new arms race?

Major powers such as the United States, Russia and China are heavily investing in AI research and development and seek to outpace each other in creating advanced AI-driven systems for surveillance, threat detection, cyber operations, autonomous weapons and decision support in nuclear strategies (Bendett, 2024, p. 2; McDonnell et al., 2023, pp. 8–13). This intensifying technological competition increasingly resembles an arms race, where fear of relative inferiority pushes states to develop and deploy emerging capabilities pre-emptively. As Andersen (2023) notes: "Any country that inserts AI into its command and control will motivate others to follow suit, if only to maintain a credible deterrent" (p. 15).

Russia's posture is especially illustrative of this dynamic. While Russian political and military leaders often claim that AI systems will remain under human control, they also frame Western developments in AI and autonomous military systems as direct threats to Russian security and strategic stability and emphasise the need to catch up (Nadibaidze & Miotto, 2023, pp. 55–56). Bendett (2024) notes that Russia already fears lagging behind the United States and China in developing AI-enabled warfare (p. 3). In 2021, Putin called for the development of AI-driven decision support systems in the military sphere and asserted that success in decision making directly depends on speed and accuracy: "it is necessary to develop decision support systems for commanders at all levels, especially at the tactical level, [and] to integrate artificial intelligence technologies into these systems" (Ria Novosti, 2021). In February 2024, he approved a new national strategy for developing AI until 2030 (TASS, 2024). The new document contains 40 additional pages compared to the previous 2019 version. Furthermore, Deputy Minister of Defence Ruslan Tsalikov openly stated that Russia already has sufficient potential to become a global leader in the development and use of AI technologies (Zvezda News, 2021). While the reality of such proclamations can be questioned, the strategic mentality behind them is significant and likely to intensify the emerging arms race.

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

I have written elsewhere about the dangers of the Russian "catch up and overtake" mentality prevalent during the Cold War nuclear arms race (Zatsepina, 2025, pp. 12–13). During the Cold War, it enabled and legitimised large-scale Soviet industrial and military build-up along with territorial expansion and vertical nuclear proliferation, with Soviet leaders frequently highlighting the necessity of competing with and necessarily overtaking the United States. As Bendett (2024) notes, this Cold War legacy persists in current Russian rhetoric, which portrays AI development not only as a matter of technological progress but as a strategic imperative to avoid falling behind adversaries (p. 3). Similarly, Kroenig (2024) describes Russia as a revisionist power that places nuclear weapons at the centre of its security doctrine and views limited nuclear use, including first use, as a viable strategy to deter or coerce NATO. In this context, as argued earlier, AI is not a neutral tool but a force multiplier that may be shaped by, and in turn may reinforce, destabilising doctrines. In addition, the "catch up and overtake" mentality reinforces technological insecurity, which in turn fuels escalation. This is the classic security dilemma, where actions taken for defensive reasons, like developing AI to improve command and control, are perceived as an offensive threat by others.

This is consistent with what Buchanan (2017) calls the "cybersecurity dilemma": when states introduce new capabilities for defensive purposes, others perceive them as offensive threats and respond in kind, heightening mutual suspicion (pp. 189–190). In the AI-supported nuclear domain, this dilemma is even more acute. As discussed by Johnson (2022, pp. 350–352), because AI systems are opaque, fast-moving and difficult to verify or constrain, adversaries may assume the worst about their function (e.g., pre-emptive targeting, decision acceleration) or deployment (e.g., integrated into NC3 systems). The recursive nature of these security dynamics, combined with the Clausewitzian notion of the "fog of war," and offensively oriented military doctrine, can increase the risk of crises and act as a catalyst for inadvertent escalation (Johnson, 2022, p. 343). In this case, the Russian case underscores the central argument of this article: the divergent deterrence logics, when combined with competitive technological development, amplify misperception risks.

Despite public commitments to human control, concerns remain over how AI is being integrated into nuclear decision-making processes. In May 2024, U.S. State Department arms control official Paul Dean stated that the United States would never defer a decision to use nuclear weapons to AI and urged Russia and China to make similar proclamations (Hart, 2024). Although there appears to be a consensus in Russian military and political circles that humans should retain full control, debates persist regarding the possibility of automating components of early warning and NC3 systems (Shakirov, 2023, p. 30). Such proclamations may appear reassuring, yet they remain insufficient. The insistence on human oversight does not eliminate the risks posed by AI's influence in the nuclear domain. According to Saltini (2023), while all five nuclear-armed P5 states stress the importance of human control, they differ significantly in how they operationalise AI's permissible role in early warning, ISR and decision support functions (pp. 20–24). In addition, as noted previously, AI can compress decision timelines, accelerate crisis dynamics and inject cognitive biases masked as rational optimisation (Hoffman & Kim, 2023, pp. 16–21). This undermines the conditions for deliberate, context-sensitive judgement.

In high-stakes scenarios, reliance on AI to interpret complex data or generate recommendations could inadvertently escalate conflicts, especially if adversaries misinterpret the intent behind actions informed by AI (Wong et al., 2020, p. 60). Andersen (2023, p. 15) similarly warns that AI with no nuclear-weapons authority could still "pursue a gambit that inadvertently escalates a conflict so far and so fast that a panicked nuclear launch follows."

This emerging arms race is further complicated by the erosion of arms control agreements – one of the defining characteristics of the Third Nuclear Age. With Russia's suspension of compliance with the New START Treaty and no successor treaty in sight, we are entering a period without any formal constraints on nuclear arsenals among the major powers for the first time since the early 1970s (Kroenig, 2024). Meanwhile, China's rapid nuclear expansion and growing interest in AI-enabled capabilities

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

introduce a third major actor into the strategic equation, making arms racing more multipolar and less predictable (Johnson, 2022, pp. 358–359).

Even if AI is not granted launch authority, its use in nuclear decision making may introduce escalatory risks. For instance, AI systems trained on incomplete or biased historical data may recommend modernisation or expansion of nuclear arsenals by misinterpreting restraint or treaty compliance as strategic weakness (Black et al., 2024, p. 49). This is especially problematic when AI absorbs Cold War–era strategic assumptions, such as the primacy of strategic parity or the logic of pre-emption, as part of its analytical baseline. In high-stakes scenarios, such systems may generate recommendations that exacerbate escalation pressures.

The traditional deterrence logic hinges on the assumption that decisions are made by rational human actors capable of calculating risks and weighing consequences. Waltz's "more will be better" argument posits that the spread of nuclear capabilities among rational actors induces caution and restraint. But if AI systems trained on divergent doctrinal frameworks shape threat perceptions and operational decisions, then deterrence logic itself may become unstable. Under these conditions, more AI does not equal more stability. Instead, it risks accelerating us into a more dangerous iteration of the arms race. In this evolving context, deterrence can no longer be assumed as a rational or universal framework.

5. Conclusion

AI introduces unprecedented speed, opacity and bias into strategic decision making, which complicates the traditional assumptions that underpin nuclear deterrence. As AI becomes integrated into nuclear decision-making processes, the "easy calculations" Waltz once identified – those presumed to guide rational actors away from nuclear use – are increasingly muddled by automated processes that may obscure human judgement and reduce the space for reflection. In an environment already marked by doctrinal asymmetries and mutual mistrust, AI risks becoming not a stabilising force, but an accelerant of instability.

This article has focused on the overlooked but urgent problem of inadvertent escalation stemming from conceptual divergence in deterrence thinking, specifically, how Russian approaches to strategic deterrence, grounded in distinct linguistic, doctrinal and cognitive traditions, may interact problematically with AI-supported systems. In doing so, it contributes to the broader challenge outlined by Erskine and Miller (2024): understanding how AI is reshaping resort-to-force decision making in an era of uncertainty. These systems may misinterpret adversary actions, reinforce escalation biases and generate false confidence in AI-derived recommendations, particularly in time-compressed or ambiguous crisis settings. The risk is not merely technical failure, but deep misperception, compounded by the speed and "black box" logic of AI.

This challenge is contributing to emerging arms race dynamics, as major powers invest in AI capabilities to avoid strategic disadvantage. While states may claim defensive intent, their AI developments may be interpreted by rivals through pre-existing strategic assumptions and suspicions. As this article has shown, divergent deterrence logics can complicate these interpretations, heightening the potential for mistrust and misperception. In the Russian case, evolving doctrine, AI rhetoric and an emphasis on information confrontation illustrate how military-technological advances are filtered through specific strategic worldviews, reinforcing the need for greater attention to how conceptual differences shape the risks associated with AI integration.

In light of these risks, there is a pressing need for norm-building, transparency and confidence-building measures, and rigorous safeguards to ensure stability in an AI-enhanced nuclear order. Establishing clear limits on AI's role in nuclear decision making, promoting dialogue on the implications of AI in nuclear strategy and reaffirming the primacy of human judgement are essential steps to prevent the escalation of an AI-driven arms race and the risk of an accidental use of nuclear weapons (Boulanin et al., 2020, pp. 140–143). Encouragingly, the November 2024 agreement between

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

Presidents Joe Biden and Xi Jinping, which affirmed that only humans, and not AI, should control nuclear weapons, offers a rare and necessary moment of alignment (Reuters, 2024). While such progress is unlikely with Russia in the near term, especially amid its ongoing invasion of Ukraine, such bilateral initiatives may offer more immediate and pragmatic avenues for setting red lines and establishing AI-related nuclear norms between rivals. Continuous monitoring and adaptation of these AI systems are crucial to address emerging threats as well as conceptual, geopolitical and technological shifts, ensuring they evolve alongside the changing strategic environment.

Furthermore, as Geist (2016) argues, "major military powers will have to strike difficult compromises to forgo some of the warfighting potential of artificial intelligence in exchange for mutual security" (p. 320). This may include prohibiting fully autonomous systems from influencing nuclear launch decisions and developing verification regimes analogous to arms control for AI-enabled capabilities. Without such compromises, the rationality at the heart of deterrence may erode and leave us with a more volatile, mistrustful and unpredictable global nuclear landscape.

In this unfolding Third Nuclear Age, where AI collides with divergent deterrence logics, stability can no longer be taken for granted. These technologies are not neutral tools but amplifiers of strategic assumptions, and when those assumptions are incompatible, the risks of misperception and escalation grow. Without urgent attention to these dynamics, we risk *waltzing into uncertainty*, as AI reshapes the foundations of nuclear deterrence.

Competing interests. The author declares none.

References

Acton, J. M. (2013, February 5). Reclaiming strategic stability. *Strategic Studies Institute*. https://carnegieendowment.org/posts/2013/02/reclaiming-strategic-stability?lang=en

Adamsky, D. (2018). From Moscow with coercion: Russian deterrence theory and strategic culture. *Journal of Strategic Studies*, *41*(1–2), 33–60. https://doi.org/10.1080/01402390.2017.1347872

Adamsky, D. (2024). *The Russian way of deterrence: Strategic culture, coercion, and war*. Stanford University Press.

Akimenko, V. (2021, August 10). Russia and strategic non-nuclear deterrence. *Chatham House*. https://www.chathamhouse.org/2021/07/russia-and-strategic-non-nuclear-deterrence/russias-strategic-deterrence-concept

Andersen, R. (2023, June 2). Never give artificial intelligence the nuclear codes. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2023/06/ai-warfare-nuclear-weapons-strike/673780/

Bellaby, R. (2024). The ethical problems of "intelligence AI". *International Affairs*, *100*(6), 2525–2542. https://doi.org/10.1093/ia/iiae227

Bendett, S. (2024). Center for a New American Security. Retrieved June 5, 2025, from https://www.cnas.org/publications/reports/the-role-of-ai-in-russias-confrontation-with-the-west

Black, J., Eken, M., Parakilas, J., Dee, S., Ellis, C., Suman-Chauhan, K., Bain R., Fine H., Aquilino M.C., Lebret M, and Palicka, O. (2024). RAND Europe. Retrieved May 20, 2025, from https://www.rand.org/content/dam/rand/pubs/research_reports/RRA3200/RRA3295-1/RAND_RRA3295-1.pdf

Boulanin, V. (2019). The future of machine learning and autonomy in nuclear weapon systems. In V. Boulanin, Ed., *The impact of artificial intelligence on strategic stability and nuclear risk. Volume 1. Euro-Atlantic perspectives* (pp. 53–63). Stockholm International Peace Research Institute. https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf.

Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Peldan Carlsson, M. (2020). *Artificial Intelligence, Strategic Stability and Nuclear Risk*. Stockholm: Stockholm International Peace Research Institute. https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf

Bracken, S. (2000). The second nuclear age. *Foreign Affairs*, *79*(1), 146–156. https://doi.org/10.2307/20049619

Buchanan, B. (2017). *The cybersecurity dilemma: Hacking, trust and fear between nations*. Online Edition. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780190665012.001.0001

Charap, S. (2020). Strategic sderzhivanie: Understanding contemporary Russian approaches to "deterrence.". George C. Marshall European Center for Security Studies. Retrieved June 5, 2025, from https://www.marshallcenter.org/en/publications/security-insights/strategic-sderzhivanie-understanding-contemporary-russian-approaches-deterrence-0

Chase, M. S., & Chan, A. (2016). China's evolving strategic deterrence concepts and capabilities. *The Washington Quarterly*, *39*(1), 117–136. https://doi.org/10.1080/0163660X.2016.1170484

Chernavskikh, V. (2024). *Nuclear weapons and artificial intelligence: Technological promises and practical realities*. Stockholm International Peace Research Institute. https://www.sipri.org/sites/default/files/2024-09/bp_2409_ai-nuclear.pdf

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

Clapp, S. (2025). Defence and artificial intelligence. Retrieved May 20, 2025, from https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/769580/EPRS_BRI(2025)769580_EN.pdf

Crilley, R. (2023). *Unparalleled catastrophe: Life and death in the third nuclear age*. Manchester University Press.

Depp, M., & Scharre, P. (2024, January 16). Artificial intelligence and nuclear stability. *War on the Rocks*. https://warontherocks.com/2024/01/artificial-intelligence-and-nuclear-stability/

The Economist (2024, February 8). Vladimir Putin wants to catch up with the West in AI. *The Economist*. https://www.economist.com/business/2024/02/08/vladimir-putin-wants-to-catch-up-with-the-west-in-ai

Erskine, T., & Miller, S. (2024). AI and the decision to go to war: Future risks and opportunities. *Australian Journal of International Affairs*, *78*(2), 135–147. https://doi.org/10.1080/10357718.2024.2349598

Favaro, M., Renic, N., & Kühn, U. (2022). *Negative multiplicity: Forecasting the future impact of emerging technologies on international stability and human security*. Institute for Peace Research and Security Policy. Retrieved June 5, 2025, from https://ifsh.de/file/publication/Research_Report/010/Research_Report_010.pdf

Fink, A. (2023). The wind rose's directions: Russia's strategic deterrence during the first year of the war in Ukraine. Proliferation Papers, No. 65, Ifri. https://www.ifri.org/sites/default/files/migrated_files/documents/atoms/files/ifri_fink_russian_strategic_deterrence_aout2023.pdf.

Fink, A. L. (2017). The evolving Russian concept of strategic deterrence: Risks and responses. *Arms Control Today*. https://www.armscontrol.org/act/2017-07/features/evolving-russian-concept-strategic-deterrence-risks-and-responses#endnote38

Futter, A., & Zala, B. (2021). Strategic non-nuclear weapons and the onset of a third nuclear age. *European Journal of International Security*, *6*(3), 257–277. https://doi.org/10.1017/eis.2021.2

Geist, E. M. (2016). It's already too late to stop the AI arms race – We must manage it instead. *Bulletin of the Atomic Scientists*, *72*(5), 318–321. https://doi.org/10.1080/00963402.2016.1216672

Hart, R. (2024, May 2). Don't let AI control your nukes, U.S. official urges China and Russia. *Forbes*. https://www.forbes.com/sites/roberthart/2024/05/02/dont-let-ai-control-your-nukes-us-official-urges-china-and-russia/

Hoffman, W., & Kim, H. M. (2023). Reducing the risks of artificial intelligence for military decision advantage, policy brief. Center for Security and Emerging Technology. Retrieved June 5, 2025, from https://cset.georgetown.edu/wp-content/uploads/CSET-Reducing-the-Risks-of-Artificial-Intelligence-for-Military-Decision-Advantage.pdf

Holmes, M., & Wheeler, N. J. (2024). The role of artificial intelligence in nuclear crisis decision making: A complement, not a substitute. *Australian Journal of International Affairs*, *78*(2), 164–174. https://doi.org/10.1080/10357718.2024.2333814

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, *1*(3). https://repositories.lib.utexas.edu/server/api/core/bitstreams/74307125-fc5e-4706-86fc-1b035e4bbfbc/content

Hymans, J. E. C. (2006). Theories of nuclear proliferation. *The Nonproliferation Review*, *13*(3), 455–465. https://doi.org/10.1080/10736700601071397

Jervis, R. (1989). Rational deterrence: Theory and evidence. *World Politics*, *41*(2), 183–207. https://doi.org/10.2307/2010407

Johnson, J. (2020a). Artificial intelligence: A threat to strategic stability. *Strategic Studies Quarterly*, *14*(1), 16–39. https://www.jstor.org/stable/26891882

Johnson, J. (2020b). Deterrence in the age of artificial intelligence & autonomy: A paradigm shift in nuclear deterrence theory and practice? *Defense & Security Analysis*, *36*(4), 422–448. https://doi.org/10.1080/14751798.2020.1857911

Johnson, J. (2021). "Catalytic nuclear war" in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states. *Journal of Strategic Studies*, 1–41. https://doi.org/10.1080/01402390.2020.1867541

Johnson, J. (2022). 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), 337–359. https://doi.org/10.1017/eis.2021.23

Johnson, J. (2023). *AI and the bomb: Nuclear strategy and risk in the digital age*. Oxford University Press.

Knopf, J. W. (2010). The fourth wave in deterrence research. *Contemporary Security Policy*, *31*(1), 1–33. https://doi.org/10.1080/13523261003640819

Kofman, M. (2024, September 10). Book review roundtable: Russian ways of thinking about deterrence. *Texas National Security Review*. https://tnsr.org/roundtable/book-review-roundtable-russian-ways-of-thinking-about-deterrence/#_ftn71

Kofman, M., & Fink, A. (2020, June 23). Escalation management and nuclear employment in Russian military strategy. *War on the Rocks*. https://warontherocks.com/2020/06/escalation-management-and-nuclear-employment-in-russian-military-strategy/

Kofman, M., Fink, A., & Edmonds, J. (2020). Russian strategy for escalation management: Evolution of key concepts. Retrieved June 5, 2025, from https://www.cna.org/reports/2020/04/DRM-2019-U-022455-1Rev.pdf

Kroenig, M. (2015). The history of proliferation optimism: Does it have a future? *Journal of Strategic Studies*, *38*(1–2), 98–125. https://doi.org/10.1080/01402390.2014.893508

Kroenig, M. (2021). Will emerging technology cause nuclear war? Bringing geopolitics back in. *Strategic Studies Quarterly*, *15*(4), 59–73. https://www.airuniversity.af.edu/Portals/10/SSQ/documents/Volume-15_Issue-4/D-Kroenig.pdf

Kroenig, M. (2024, October 7). Strategic stability in the third nuclear age. *Atlantic Council*. Issue Brief. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/strategic-stability-in-the-third-nuclear-age/

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

Lebow, R. N., & Stein, J. G. (1989). Rational deterrence theory: I think, therefore I deter. *World Politics*, *41*(2), 208–224. https://doi.org/10.2307/2010408

Lewis, P., Williams, H., Pelopidas, B., & Aghlani, S. (2014). *Too close for comfort: Cases of near nuclear use and options for policy*. Chatham House. The Royal Institute of International Affairs. https://www.chathamhouse.org/2014/04/too-close-comfort-cases-near-nuclear-use-and-options-policy

McDonnell, T., Chesnut, M., Ditter, T., Fink, A., Lewis, L., & Westerhaug, A. (2023). Artificial intelligence in nuclear operations: Challenges, opportunities, and impacts. Retrieved May 26, 2025, from https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf

Merriam, J. (2023). One move Ahead — Diagnosing and countering Russian reflexive control. *The Journal of Slavic Military Studies*, *36*(1), 1–27. https://doi.org/10.1080/13518046.2023.2201113

Morgan, F. E., Mueller, K. P., Medeiros, E. S., Pollpeter, K. L., & Cliff, R. (2008). *Dangerous thresholds: Managing escalation in the 21st century*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/monographs/MG614.html

Nadibaidze, A., & Miotto, N. (2023). The impact of AI on strategic stability is what states make of it: Comparing US and Russian discourses. *Journal for Peace and Nuclear Disarmament*, *6*(1), 47–67. https://doi.org/10.1080/25751654.2023.2205552

Parke, M. (2023, November 8). Preventing AI nuclear Armageddon. *Project Syndicate*. https://www.project-syndicate.org/commentary/dangers-of-artificial-intelligence-ai-applications-nuclear-weapons-by-melissa-parke-2023-11

Petrov, I. (2021, February 9). Shoigu postavil zadachu po vnedreniyu iskusstvennogo intellekta V oruzhie [shoigu set the task of integrating AI into weapons]. *Rossiyskaya Gazeta*. https://rg.ru/2021/02/09/shojgu-postavil-zadachu-po-vnedreniiu-iskusstvennogo-intellekta-v-oruzhie.html

President of Russia (2010, February 5). Ukaz prezidenta rossiyskoy federatsii ot 05.02.2010 no. 146 o voyennoy doctrine rossiyskoy federatsii. [Decree of the President of the Russian Federation from 05.02.2010 no. 146 on the Military Doctrine of the Russian Federation]. http://www.kremlin.ru/acts/bank/30593

President of Russia (2020, June 2). Ukaz prezidenta rossiyskoy federatsii ot 02.06.2020 no. 355 ob osnovah gosudarstvennoy politiki rossiyskoy federatsii v oblasti yadernogo sderzhivaniya. [Decree of the President of the Russian Federation from 02.06.2020 no. 355 on the Fundamentals of State Policy of the Russian Federation on Nuclear Deterrence]. http://www.kremlin.ru/acts/bank/45562

President of Russia. (2022, December 21). Zasedaniye kollegii ministerstva oboroni [meeting of defense ministry board]. http://kremlin.ru/events/president/news/70159

President of Russia (2024, November 19). Ukaz prezidenta rossiyskoy federatsii ot 09.11.2024 no. 991 ob utverzhdenii osnov gosudarstvennoy politiki rossiyskoy federatsii v oblasti yadernogo sderzhivaniya. [Decree of the President of the Russian Federation from 09.11.2024 no. 991 on the approval of the Fundamentals of State Policy of the Russian Federation on Nuclear Deterrence]. http://www.kremlin.ru/acts/bank/51312

Price, M., Walker, S., & Wiley, W. (2018). The machine beneath: Implications of artificial intelligence in strategic decision making. *Prism*, *7*(4), 92–105. https://ndupress.ndu.edu/Media/News/News-Article-View/Article/1983497/the-machine-beneath-implications-of-artificial-intelligence-in-strategic-decisi/

Reuters (2024, November 17). Biden, Xi agree that humans, not AI, should control nuclear arms. https://www.reuters.com/world/biden-xi-agreed-that-humans-not-ai-should-control-nuclear-weapons-white-house-2024-11-16/

RIA Novosti (2021, December 21). Putin prizval razvivat' sistemy iskusstvennogo intellekta V voyennoy sfere [Putin called for developing AI systems in the military sphere]. https://ria.ru/20211221/putin-1764712935.html

Rivera, J., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & Schneider, J. (2024). Escalation risks from language models in military and diplomatic decision-making. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 836–898. https://doi.org/10.1145/3630106.3658942

Sagan, S. D. (1996). Why do states build nuclear weapons? *International Security*, *21*(3), 54–86. https://doi.org/10.2307/2539273

Saltini, A. (2023). AI and nuclear command, control and communications: P5 perspectives. Retrieved June 1, 2025, from https://europeanleadershipnetwork.org/wp-content/uploads/2023/11/AVC-Final-Report_online-version.pdf

Schörnig, N. (2022, January 19). AI for arms control: How artificial intelligence can foster verification and support arms control. *PRIF Spotlight*. https://www.prif.org/fileadmin/Daten/Publikationen/PRIF_Spotlights/2022/PRIF_Spotlight_1_2022_barrierefrei.pdf

Security Council of the Russian Federation (2014, December 25). Voyennaya doctrina rossiyskoy federatsii. [the military doctrine of the Russian Federation]. http://www.scrf.gov.ru/security/military/document129/

Shakirov, O. (2023, November 13). Russian thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making. *The European Leadership Network*. https://europeanleadershipnetwork.org/report/russian-thinking-on-ai-integration-and-interaction-with-nuclear-command-and-control-force-structure-and-decision-making/

Stokes, J., Khal, C. H., Kendall-Taylor, A., & Lokker, N. (2025). *Averting AI armageddon: U.S-China-Russia rivalry at the nexus of nuclear weapons and artificial intelligence*. Center for a New American Security. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/Averting-AI-Armageddon_TSP-IPS_2025_finalB_021325.pdf

https://doi.org/10.1017/cfl.2025.10021 Published online by Cambridge University Press

Tannenwald, N. (2007). *The nuclear taboo: The United States and the non-use of nuclear weapons since 1945*. Cambridge University Press.

TASS (2024, February 15). Putin obnovil Natsional'nuyu strategiyu razvitiya II do 2030 goda [Putin renewed National strategy of developing AI until 2030]. https://tass.ru/politika/20000627?ysclid=ly4pdje7bt395931940

Thomas, T. L. (2019). Russian military thought: Concepts and elements. The MITRE corporation. Retrieved June 5, 2025, from https://www.mitre.org/sites/default/files/2021-11/prs-19-1004-russian-military-thought-concepts-elements.pdf

Trenin, D. (2024, July 1). Strategicheskoye sderzhivanie: Novie konturi. [Strategic deterrence: New contours]. *Russia in Global Affairs*. https://globalaffairs.ru/articles/sderzhivanie-trenin/?ysclid=mbkkin2p82638294007

Veebel, V. (2021). Russia and Western concepts of deterrence, normative power, and sanctions. *Comparative Strategy*, *40*(3), 268–284. https://doi.org/10.1080/01495933.2021.1912509

Ven Bruusgaard, K. (2016). Russian Strategic Deterrence. *Survival*, *58*(4), 7–26. https://doi.org/10.1080/00396338.2016.1207945

Ven Bruusgaard, K. (2024). Deterrence asymmetry and strategic stability in Europe. *Journal of Strategic Studies*, *47*(3), 334–362. https://doi.org/10.1080/01402390.2024.2354322

Wachs, L. (2023). Russian nuclear roulette? Elites and public debates on nuclear weapons in Moscow after Ukraine. *The Nonproliferation Review*, *30*(4–6), 173–196. https://doi.org/10.1080/10736700.2024.2435706

Waltz, K. N. (1979). *Theory of international politics*. Waveland Press.

Waltz, K. N. (1988). The origins of war in neorealist theory. *The Journal of Interdisciplinary History*, *18*(4), 615–628. https://doi.org/10.2307/204817

Waltz, K. N. (1990). Nuclear myths and political realities. *The American Political Science Review*, *84*(3), 731–745. https://doi.org/10.2307/1962764

Wong, Y. H., Yurchak, J., Button, R. W., Frank, A. B., Laird, B., Osoba, O. A., Steeb, R., Harris, B.N. and Bae, S. J. (2020). *Deterrence in the age of thinking machines*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2797.html

Zala, B. (2019). How the next nuclear arms race will be different from the last one. *Bulletin of the Atomic Scientists*, *75*(1), 36–43. https://doi.org/10.1080/00963402.2019.1555999

Zala, B. (2024). Should AI stay or should AI go? First strike incentives & deterrence stability. *Australian Journal of International Affairs*, *78*(2), 154–163. https://doi.org/10.1080/10357718.2024.2328805

Zatsepina, L. (2025). Transforming discourse, driving change: Gendered nuclear identities and the Soviet Union's shift to disarmament in the 1980s. *Millennium*. https://doi.org/10.1177/03058298241309646

Zvezda News (2021, August 23). Tsalikov zayavil, chto Rossiya mozhet stat' odim iz liderov v sfere iskustvennogo intellekta [Tsalikov stated that Russia could become one of the leaders in the sphere of artificial intelligence]. https://tvzvezda.ru/news/20218231628-4bOIi.html?ysclid=m63h3ft4sj670836339

Dr Luba Zatsepina is a Senior Lecturer in International Relations and Politics at Liverpool John Moores University. She primarily teaches courses on International Relations theory, military history, global security, and Strategic Studies. Previously, she held a lecturing position at the University of Edinburgh and a research position with the Proliferation and Nuclear Policy team at the Royal United Services Institute (RUSI). Luba's research interests centre on nuclear politics, particularly in the UK and the Soviet Union/Russia. Her work follows two main strands. The first examines Soviet nuclear weapons policy during the Cold War, with an emphasis on the discursive construction of nuclear identity. The second explores the role of Artificial Intelligence (AI) in nuclear command and control and its implications for deterrence theory.

Cite this article: Zatsepina, L. (2025). Waltzing into uncertainty: AI in nuclear decision making and the challenge of divergent deterrence logics. *Cambridge Forum on AI: Law and Governance*, *1*, e56, 1–15. https://doi.org/10.1017/cfl.2025.10021