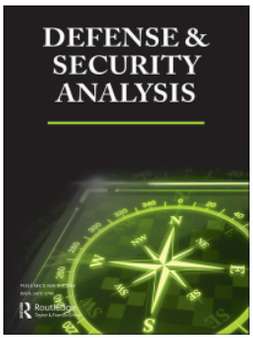

Defense & Security Analysis

ISSN: 1475-1798 (Print) 1475-1801 (Online) Journal homepage: www.tandfonline.com/journals/cdan20

Deterrence in the age of artificial intelligence & autonomy: a paradigm shift in nuclear deterrence theory and practice?

James Johnson

To cite this article: James Johnson (2020) Deterrence in the age of artificial intelligence & autonomy: a paradigm shift in nuclear deterrence theory and practice?, Defense & Security Analysis, 36:4, 422-448, DOI: 10.1080/14751798.2020.1857911

To link to this article: https://doi.org/10.1080/14751798.2020.1857911

Published online: 20 Jan 2021.

Submit your article to this journal

Article views: 4052

View related articles

View Crossmark data

Citing articles: 12 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=cdan20

DEFENSE & SECURITY ANALYSIS
2020, VOL. 36, NO. 4, 422–448
https://doi.org/10.1080/14751798.2020.1857911

Routledge
Taylor & Francis Group

Check for updates

Deterrence in the age of artificial intelligence & autonomy: a paradigm shift in nuclear deterrence theory and practice?

James Johnson

School of Law and Government, Dublin City University, Dublin, Ireland

ABSTRACT

How might nuclear deterrence be affected by the proliferation of artificial intelligence (AI) and autonomous systems? How might the introduction of intelligent machines affect human-to-human (and human-to-machine) deterrence? Are existing theories of deterrence still applicable in the age of AI and autonomy? The article builds on the rich body of work on nuclear deterrence theory and practice and highlights some of the variegated and contradictory – especially human cognitive psychological – effects of AI and autonomy for nuclear deterrence. It argues that existing theories of deterrence are not applicable in the age of AI and autonomy and introducing intelligent machines into the nuclear enterprise will affect nuclear deterrence in unexpected ways with fundamentally destabilising outcomes. The article speaks to a growing consensus calling for conceptual innovation and novel approaches to nuclear deterrence, building on nascent post-classical deterrence theorising that considers the implications of introducing non-human agents into human strategic interactions.

KEYWORDS

Artificial intelligence; autonomous weapons; nuclear deterrence; emerging technology; escalation; strategic stability

Introduction

As a growing number of great military powers invest political capital and financial resources to develop the field of artificial intelligence (AI) technology and AI-enhanced autonomous weapons systems in deriving the maximum potential military benefits – at a tactical, operational, and strategic level – these systems offer.[1] As a result, the ubiquity of these new classes of advanced capabilities – and the incentives for militaries to adopt them – on the future battlefield is fast becoming a foregone certainty.[2] How might the rise of these capabilities weaken or strengthen deterrence? Given the recent genesis of AI and autonomy in a military context, and the rich body of work that describes these trends, this article is premised on the assumption that AI and autonomy technology will continue to be embraced – at different speeds and for different goals – by global militaries.[3] The article argues that AI and autonomy could decrease nuclear stability and increase the tendency for escalation to nuclear use, thereby undermining deterrence.

The article considers the potential implications of AI and autonomy for nuclear deterrence theory and practice. Using the rich deterrence literature as a point of departure, the

CONTACT James Johnson james.johnson@dcu.ie

© 2021 Informa UK Limited, trading as Taylor & Francis Group

article considers how advances in AI and autonomy might affect the manner, and means, by which nuclear-armed states seek to deter adversaries from embarking on a particular course of action through threats, denial, coercion, and compellence. Given the potentially transformative effects of AI and autonomy augmentation on a range of (nuclear and non-nuclear) strategic technologies, re-thinking existing assumptions, theories, and permutations of deterrence – premised on human-rationality, perceptions, and nuanced signalling – is now needed.[4] The focus of this article is to examine the impact of the adoption of these technologies on nuclear deterrence. How might non-human agents' introduction into a crisis or conflict between nuclear powers affect deterrence, escalation, and strategic stability?

Because of the multifaceted possible intersections of AI with the nuclear enterprise – and *before* states reconfigure their nuclear doctrines and postures – research on this topic is important to anticipate (and mitigate) the risk of misperception, miscalculation, and inadvertent escalation. The article considers how increasing complexity, speed, compressed decision-making, and cognitive-psychological associated with AI and autonomy might compound these risks. How might states with different political structures, philosophies of use, and deterrence policies view and likely respond to these dynamics?

This article builds on the extensive body of work that considers nuclear deterrence theory and other theoretical frameworks to consider AI and autonomous systems' potential effects.[5] Classical nuclear deterrence approaches, premised on the assured threat of a retaliatory second strike – or mutually assured destruction (MAD) – is a fundamentally psychological-political-technical phenomenon that has worked to reduce the chances of deliberate use of nuclear weapons.[6] There is an extensive body of literature on the intersection of emerging technology and the nuclear enterprise – and the potential for warfare more broadly defined.[7] The literature says very little about how existing concepts of escalation, nuclear terrorism, and classical deterrence theories might apply (or be tested) in the digital age – increasingly defined by developments in AI and autonomy – where perfect information and rational decision-making cannot be assumed.[8]

This remainder of this article proceeds as follows. First, it identifies and contextualises the core theories and concepts necessary to examine the implications of introducing AI and autonomous systems into the nuclear domain – namely, deterrence, escalation, and strategic stability. It describes the broadened notion of deterrence that emerged after the end of the Cold War, referred to as "fourth wave" in deterrence scholarship – with insights from criminology, terrorism studies, and human psychology, which is a key focus of this article. Are existing theories of deterrence and related concepts still applicable in the era of AI and autonomy?

Second, it applies this conceptual framework to examine the possible ways in which advances in AI and autonomy might undermine the central pillars that, for several decades, have undergirded nuclear deterrence and strategic stability. It unpacks the key technical and non-technical features of this emerging paradigm to elucidate how new AI-powered capabilities could affect the key components of strategic deterrence – especially nuclear command, control, and communication (NC3) systems and second-strike capabilities. Will AI-enabled capabilities strengthen or weaken deterrence?

Third, the article examines the growing multi-polarity in international security competition – radically shifting the geostrategic balance – and considers how the confluence of this trend with the development of AI and autonomy might compound the problem of

uncertainty caused by complexity and the speed of warfare, the entanglement of nuclear and conventional capacities, and managing crisis escalation and signalling premised on the assumption of rationality – between nuclear-armed adversaries. It argues that although the explicit linkages between the structure of the international political system and developments in AI and autonomy are not clear,[9] shifts in the geopolitical balance will significantly influence how deterrence dynamics could play in future crises and conflict between peer and near-peer nuclear-powers.

The final section examines the implications of human and machine interactions for nuclear deterrence. How might these interactions spark inadvertent or accidental nuclear escalation to a strategic level? It highlights the potential for accidents and errors (both technical and human), misperception, and unintended consequences that result from new deterrence models that mix various levels of humans and machines, and new and legacy weapon systems in their interface with nuclear weapons. What trade-offs might arise from this synthesis? This analysis tentatively concludes that *ceteris paribus* greater degrees of automation, coupled with reduced human decision-making, will likely increase inadvertent escalation risk.

Deterrence theory in the digitised age: concepts, assumptions, and paradoxes

How can we best conceptualise AI and autonomy in the context of nuclear weapons? While the definition of deterrence remains contested, at its core, nuclear deterrence is concerned with seeking the means to induce caution in others by applying threats (implicit or explicit) of *punishment* (i.e. retaliation with nuclear weapons) to manipulate an adversary's behaviour, or *deny*ing (i.e. removing the expected benefits of a particular behaviour) the adversary the ability to realise their aggressive objectives in the first place.[10] Therefore, any behaviour that results in caution or apprehension has possible deterrent effects – even when the behaviour to be deterred has to be inferred.[11] Deterrence and atomic weapons are not necessarily synonymous; an adversary can be deterred in various ways *below* the nuclear threshold.[12] A related concept is "compellence," which is about encouraging an adversary to pursue an action that it might have otherwise eschewed, rather than restraining its behaviour with deterrence. However, once a situation escalates, much like the difference between defence and offense, the distinction can disappear – the coercion and counter-coercion mechanisms intrinsic to both deterrence and compellence often occur simultaneously.[13]

Whether nuclear deterrence – and deterrence generally – is effective and credible depends on three factors:[14] (1) rationality on both sides to ensure threats will suffice to shape the others' behaviour;[15] (2) the adversary's perception of the defenders' capacity and resolve (or will) to punish violations (or red-lines) or deny of objectives; and (3) the deterrent threat must be communicated and understood by the adversary – defenders must also *reassure* aggressors that threats (or demands) will *not* be carried out in the event demands are met.[16] In sum, deterrence is about influencing an adversary's sense of risk, cost–benefit assessment, and decision-making outcomes. Deterrence requires a deep understanding of the other sides' interests, priorities, strategic objectives, and perceptions. How adversaries view each other's capabilities and intentions – especially the other sides' philosophy of employment – will crucially influence these systems' deterrent

effect and the potential risk of miscalculation and escalation if these assessments prove wrong.

For deterrence to be effective, both sides need to understand how an adversary perceives its reality (i.e. value attached to its interests), and discovering ways to manipulate these constructions while simultaneously ensuring that allies' interests are addressed. Besides, these interests cannot be considered independent of a particular policy's deterrent effect, which will vary temporally.[17] Thus, determining an adversary's interests, norms, and beliefs about resolve, to craft deterrence policy are intrinsically context-bound and changeable. Deterrence can fail if either side has divergent beliefs about the perceived cost of the punishment a state can inflict and the perceived probability that it will (or is able) to inflict them.[18] Deterrence, and strategic bargaining, can fail between rational actors in asymmetric information situations (i.e. information about an actor's resolve); in particular, when incentives exist to manipulate this information, when credible commitment is problematic, and divisions exist on key policy issues.[19] While the existence of rational thinking – intrinsic to classical rational deterrence theorising – is generally assumed to prove the likelihood of deterrence success; and juxtaposed, non-rational behaviour by actors assumes a high-risk of deterrence failure, rational actors can challenge deterrence assumptions. For instance, as a result of misperception and other human cognitive fallibilities (discussed below), incomplete information (i.e. about a states' reputation costs or perception of the "balance of interests"), uncertainty about an adversary's future capabilities and intentions, time pressures and errors, or grandiose objectives can impel rational actors to fight losing wars to deter aggression.[20] In other words, even in situations when an actor's threats are credible, and the balance of interests and reputational costs are known, fears about the actor's future intentions can create "rational" incentives to engage in – or even preemptively trigger – losing wars.[21]

The political psychology literature reveals that perceptions are heavily encumbered by inferences, reputation, interpretation, and cognitive-psychological theory.[22] Thus, although an actor's behaviour might reveal or signal something important (resolve, deterrence, or reassurance), it is often "not clear exactly what is being revealed, and what others will think is being revealed."[23] For instance, a states' effort to reassure its allies that it will come to its defence may reduce the patron's sense of security; if it infers that the threat is greater than it thought to the degree that the defender considers reassurances necessary.

Grounded in these core tenets and preconditions, nuclear deterrence has undergone several transformations since the Cold War. In light of changes to the geopolitical (from bipolarity to unipolarity and finally multipolarity), the technological landscape (missile defences, hypersonic weapons, and AI and autonomy), and new security threats and domains (non-states, grey-zone conflict, space and cyberspace), political scientists have conceptualised "four waves" of deterrence theorising.[24] The fourth wave followed the end of the Cold War and continues to the present day, coinciding with the Second Nuclear Age's broader features, namely, multipolarity, asymmetric threats, non-state (especially rogue nations and terrorists) actors, and advanced strategic (nuclear and non-nuclear) weapons.[25] These "waves" progressively built upon the first wave's foundation, creating a more complex, nuanced, and applicable theory in light of new research, novel methodologies applied from other disciplines (military and

non-military), and lessons from real-world events.[26] These methodologies were also accompanied by novel concepts that have broadened the scope and practice of deterrence[27] *inter alia*, conventional deterrence, extended deterrence (and its opposite "central deterrence"), inter-war deterrence, and, more recently, cross-domain and cyberspace deterrence.[28]

Similar to research on cyber deterrence, early scholarship on deterrence theory and practice in the digital age has been predominately grounded in classical deterrence approaches – associated with the earlier waves in international relations (IR) deterrence theorising rooted in known hierarchical relationships between actors and the principle of mutually assured destruction.[29] In this way, the article speaks to the nascent "fifth wave" of modern deterrence, representing a conceptual break from previous waves of classical deterrence theorising (or post-classical deterrence) and non-human agents into deterrence.[30] Any discussion surrounding nascent emerging technology such as AI comes with an important caveat. Since we have yet to see how AI might influence deterrence, escalation, strategic stability, and crisis management in the real-world – notwithstanding the valuable insights from experimental wargaming – the discourse is largely a theoretical and speculative endeavour.[31] What are the key concepts important to explore the deterrence implications of the widespread use of AI and autonomous systems? Two other important concepts in the discussion about nuclear deterrence, nuclear-armed states, and AI and autonomy are *escalation* (especially inadvertent) and *strategic stability* (or lack thereof, instability).

Escalation

Escalation in the context of deterrence can be defined as an increase in the intensity or scope of a military situation that crosses a threshold(s) considered significant by one or more actors. Escalation occurs when at least one of the parties involved perceives (or misperceives) a significant qualitative shift in a situation.[32] These mechanisms can escalate a crisis or conflict between two or more nuclear-armed states into nuclear confrontation or cause a low-level conventional conflict to move up the "rungs" of the escalation ladder and cross the nuclear threshold – either intentionally, accidentally, or inadvertently.[33] The Post-Cold War literature is rich in scholarship on how technologically complex nuclear systems can cause technical (and human-related) accidents and false alarms, which are considered particularly escalatory where one side lacks confidence in their retaliatory (or second strike) capacity.[34] During the Cold War, the perennial fear that an action or signal misinterpreted by the other – in the context of uncertainty and incomplete information associated with modern warfare – could trigger nuclear pre-emption is a useful point of departure to consider AI and autonomy.[35]

Three distinct, but not always separate, mechanisms can lead to nuclear escalation – *deliberate* (or intentional), *inadvertent*, and *accidental* escalation (encompassing mistaken or unauthorised usage).[36] These distinctions are not, however, binary or mutually exclusive. An escalation mechanism that leads from a crisis or conflict to its outcome involves more than one of these categories. For example, if an accidental or inadvertent escalation signal or event is triggered by a non-state actor's nefarious actions – such as false flag cyber-operation against a state's NC3 systems – which in turn leads to a deliberate escalatory response.[37] Moreover, the deliberate use of nuclear weapons that

originates from a false, manipulated, or distorted assessment of a situation, or in response to an early-warning system false alarm, can quickly muddy the lines of intentionality.[38] In short, the binary distinction between deliberate and unintentional use of nuclear weapons is inherently problematic.[39] Ultimately, whether the impact of unintended escalation risk is stabilising or destabilising depends on the actor's relative strength, and the fear it instils in its adversary.[40]

Accidental nuclear war – a nuclear confrontation without a deliberate and properly informed decision to use nuclear weapons on the part of the nuclear-armed state(s) involved – could be caused by a variety of accidents, most often encompassing a combination of human –and human-machine interaction failure – system errors, and procedural or organisational factors.[41] Moreover, despite paying lip-service to Machiavelli's Fortuna (role of uncertainty in international affairs), decision-makers underestimate the importance and frequency of accidents and randomness in these interactions.[42] Thus, historical cases where human-machine interactions cause or compound accidents involving complex weapon systems, therefore, AI-enhanced systems operating at higher speeds, levels of sophistication, and compressed decision-making timeframes, will likely further reduce the scope for de-escalating situations, and contribute to future mishaps.[43] Similar to historical cases where human-machine interactions cause or compound accidents involving complex weapon systems, AI-enhanced systems operating at higher speeds, levels of sophistication, and compressed decision-making timeframes, will likely could further reduce the scope for de-escalating situations and contribute to future mishaps. The rapid proliferation and ubiquity of advanced technologies like offensive-cyber, hypersonic weapons, and AI and autonomous weapons, will make it increasingly difficult for states to mitigate this vulnerability without simultaneously improving their ability to strike first, thereby undermining the survivability of others' strategic forces.

Strategic stability

The concept of strategic or nuclear stability emerged in the latter half of the twentieth century, and despite being theoretically and politically contested to this day, it has proven a useful intellectual tool for analysing the potential of technically advanced weapons to undermine stability (i.e. evaluating nuclear force structures, deployment decisions, and a rationale for arms control).[44] Strategic stability is inextricably connected to the strategic thinking and debates that surrounded the "nuclear revolution," including:[45] how a nuclear war might be fought, the requirements and veracity of credible deterrence, the potential risks posed by pre-emptive and accidental strikes,[46] and how to ensure the survivability of retaliatory forces.[47] In short, strategic stability provides an over-arching theoretical framework for understanding the nature of security in the nuclear age.

In the broadest use of the term, strategic stability exists in the absence of a significant incentive for an adversary to engage in provocative behaviour. There was a lack of armed conflict and perceived incentive to use nuclear weapons first between nuclear-armed states.[48] At its core, the concept is a phenomenon that focuses on finding a *modus vivendi* for the complex interactions and incentives of two (or more) actors.[49] The term is often associated with the relative power distribution among great and rising

powers, particularly those in possession of nuclear weapons or the potential to acquire them. Therefore, strategic stability is a product of a complex interplay of political, economic, and military dynamics in which *technology* performs several functions – an equaliser, counterweight, and principal agent of change.[50]

Technology has long been used to augment, automate, and enhance human behaviour and decision-making in a military context; thus far, the "human factor" has trumped technologies' impact upon strategic stability. That is, the underlying forces behind fundamental shifts in strategic stability are generally less concerned with quantitative or qualitative assessments of military capabilities – and other measures of relative power – and instead, more focused on how nuanced institutional, cognitive, and strategic variables impact strategic decision-making and may cause misperceptions of others' intentions.[51] "Stability" is concerned with the relationship amongst these factors. Further, "strategic stability" and "strategic instability" are not necessarily mutually exclusive paradigms; they can both exist in a nuclear multipolar system.[52] The role of technological change and strategic stability can be conceptualised, therefore, as part of a complex interaction of disruptive forces (or agents of change), which during periods of heightened geopolitical rivalry, great power transitions, and strategic surprise, may erode strategic stability and make conflict more likely.

Three broad forms distinguish strategic stability: first strike stability, crisis stability, and arms-race stability. *First-strike stability* exists in situations when no one state can launch a surprise (or pre-emptive) to attack against an opponent without the fear of devastating reprisals from survivable second-strike forces. That is, the lack of *both* incentives or pressures to use nuclear weapons first in a crisis.[53] Thus, fear that the advantage of first-strike capabilities could be eroded or neutralised by an adversary would destabilise and increase incentives to launch a pre-emptive strike.[54] Throughout the Cold War, as today, the vulnerability of command and control (C2) structures to counterforce attack remains high and compounded by the increasing complexity of these systems and the propensity of states to use them to support nuclear and non-nuclear forces.

Crisis stability aims to prevent (or de-escalate) escalation during crises – such as those in Berlin and Cuba in the early 1960s. Therefore, crisis stability depends on reciprocal fear between states; when crises arise, the system does not worsen the situation. Conversely, crisis instability refers to what Thomas Schelling called the "reciprocal fear of surprise attack."[55] The belief that conflict is inevitable, and thus striking first and pre-emptively would create a strategic advantage.[56] Finally, *arms race stability* (during peacetime) can emerge when there are no exploitable inequalities (or asymmetries) separating adversaries' military forces – qualitatively or quantitatively.[57]

Crisis instability and arms race instability can arise when states use strategic capabilities and instil fear into a situation.[58] This fear closely correlates with the incentives these decisions create. Thus, if both sides possess first-strike capabilities, either side's incentive to gain by choosing to strike first depends on the fear that hesitation might allow a rival to gain the upper hand.[59] In other words, *both* the incentive and fear created by strategic weapons can exacerbate escalation risk.[60] Crisis (in)stability is, therefore, fundamentally a psychological problem that during a crisis can raise doubts about the assumptions that decision-makers share a common notion of rationality – such as assessing risk, probability, and how others perceive their actions – with an adversary.[61]

Several psychological factors can compound these dangers and worsening crisis and arms race instability, including: (1) psychological biases that impair the quality of decision-making during high-stress time-pressured situations; (2) a tendency to exaggerate the probability that an adversary is about to launch an attack, and in turn, overestimate the advantages of striking pre-emptively; and (3) the failure to see how their actions are in response to this bias and may cause an adversary to view conflict as inevitable.[62] Advanced AI-augmented autonomous weapons (e.g. cyber-offensive capabilities, hypersonic glide vehicles, and anti-satellite weapons), which blur the distinction between nuclear and conventional warfare, can heighten strategic ambiguity during crises, creating first-mover advantage incentives leading states to overestimate an adversary's capabilities and strike pre-emptively.[63]

When only one side of a competitive dyad possesses AI-augmented autonomous weapons (e.g. unmanned aerial vehicles UAVs, smart munitions, and loitering weapons), they might increase the credibility of the state's deterrent threats.[64] Moreover, even when the perception exists that AI is a force multiplier, if an adversary's AI systems are not, amongst other things, transparent, reliable, or easily verifiable, it would be difficult to determine an adversary's capabilities and objectively assess its credibility of a deterrent threat. This uncertainty could increase these systems' deterrent values if an adversary *overestimates* their capabilities or decreases their deterrent utility if actors *underestimate* them.[65] During peacetime, the proliferation of advanced weapon systems could generate the search for counter-measures that amplify states' fear and uncertainties, leaving them feeling more vulnerable – known as a security dilemma associated with arms-racing and first strike instability.[66] Conversely, where *both* sides possess these capabilities, they could be viewed as a relatively low-risk tactic to launch probing (or "salami-slicing") attacks against an adversary, creating dynamics conducive to crisis instability and unintentional escalation.[67] Autonomous drone systems could, for example, be deployed in low-intensity salami-slicing tactics to chip away at an adversary's will (or resolve), but without crossing a threshold (or psychological red-line) that would provoke escalation.[68] How might AI and autonomy affect strategic deterrence's key components in ways that undermine states' second-strike capabilities?

Unravelling of deterrence in practice

The size, mobility, hardened, and relatively hidden features of the superpowers nuclear arsenals ensured the ability of states to withstand the first strike and deliver a retaliatory second strike, constituting the core pillars of the Cold War-era nuclear deterrence – known as the "nuclear revolution."[69] Like other technologies associated with the "computer revolution" – particularly big-data analytics, robotics, quantum computing, nanotechnology, and cyber-capabilities – advances in AI and autonomy threaten to upend this fragile arrangement *inter alia* in several ways.[70]

Hunting for nuclear weapons in the digital age

How might AI-augmented intelligence gathering, and analysis systems impact the survivability and credibility of states' nuclear-deterrent forces? The integration of AI machine learning and big-data analytics can dramatically improve the ability of militaries to

locate, track, target, and destroy a rival's nuclear-deterrent forces – especially nuclear-armed submarines and mobile missile forces – and without the need to deploy nuclear weapons.[71] AI-enabled capabilities that increase the vulnerability of second-strike capabilities (or are perceived to do so) heightens uncertainty and undermines deterrence – even if the state in possession of these counterforce capabilities did not intend to use them.[72] In short, the capabilities AI might *enhance* (cyber-weapons, drones, precision-strike missiles, and hypersonic weapons), together with the ones it might *enable* (intelligence, surveillance, and reconnaissance ISR, ATR, and autonomous sensor platforms) could make hunting for mobile nuclear arsenals faster, cheaper, and more effective than before.

AI machine learning techniques could significantly improve existing machine vision and other signal processing applications, identify patterns from large data-sets of signals and imagery, and enhance autonomy and sensor fusion applications. Taken together, strengthening ISR functionality, automatic target recognition (ATR), and terminal-guidance systems would have profound implications for strategic stability. Besides, AI used in conjunction with autonomous mobile sensor platforms might compound the threat posed to mobile intercontinental ballistic missiles (ICBM) launchers. Autonomous mobile sensors would only need to locate close to mobile ICBM launchers to be effective, and thus, as the "window of vulnerability" rapidly narrowed and faced with the prospect of an imminent disarming strike, an adversary would be put under immense pressure to escalate. For instance, advances in deep learning can exponentially improve machine vision and other signal processing applications, which may overcome the main technical barriers for tracking and targeting adversaries' nuclear forces (i.e. sensing, image processing, and estimating weapon velocities and kill radius).[73] Some scholars argue that AI and autonomy could enable real-time tracking, shorten decision-cycles, and more accurate targeting – and reduce target selection errors – of an adversary's nuclear assets in ways that make counterforce operations more feasible.[74] Moreover, AI technology could put the defender at a distinct disadvantage, creating additional incentives to strike first (or pre-emptively) technologically superior military rivals. Several technologies under development are designed explicitly for this purpose.[75] The less secure a nation considers its second-strike capabilities to be, therefore, the more likely it is to countenance the use of autonomous systems within its nuclear weapons complex to bolster the survivability of its strategic forces.

Given the tendency of Chinese and Russian strategists to extrapolate from current US capabilities malign intent – and to assume future ones will threaten their security – even modest and incremental improvements in AI techniques to integrate and synthesise data about the location of an adversary's mobile missiles could exacerbate pre-existing fears and distrust.[76] Irrespective of whether future breakthroughs in AI produce irrefutable evidence of a game-changing means of locating, targeting, and destroying mobile missile forces, Chinese and Russian perceptions of US intentions in the pursuit of these capabilities would, therefore, be far more salient.[77] Despite the US's reassurances, its adversaries would be unable to dismiss the possibility that military AI capabilities would not be used in future warfare to erode the survivability of their nuclear forces – a contingency the US has prepared for several decades.[78]

Several observers posit that autonomous systems like US DARPA's *Sea Hunter* by rendering the underwater domain "transparent," might erode the second-strike deterrence

utility of stealthy ballistic-missile submarines (SSBNs), triggering use-them-or-lose-them situations. Today, however, the technical feasibility of this hypothesis remains highly contested.[79] On the one hand, several experts posit that emerging technologies such as AI, quantum communications, and big-data analytics will empower new iterations of highly portable sensing, communications, and signal-processing platforms that could render at-sea nuclear deterrence all but obsolete.[80] On the other hand, others consider this hypothesis technically and operationally premature for several reasons, not least the notoriously complex and dynamic underwater conditions that hampers underwater targeting, and the need for increased power for autonomous systems to operate extended ranges.[81]

AI-cyber threats to nuclear systems

How might AI-infused cyber-capabilities be used to subvert or otherwise compromise states' control over their nuclear systems? Today, it is thought possible that a cyber-attack (i.e. spoofing, hacking, manipulation, and digital jamming) could infiltrate a nuclear weapons system, threaten the integrity of its communications, and ultimately (and possibly unbeknown to its target) gain control of its – possibly dual-use – command and control systems. For instance, a non-state third-party hacker might "break into, interfere with, or sabotage nuclear command and control facilities; spoof or compromise early warning systems or components of the nuclear firing chain; or in a worst-case scenario even cause a nuclear explosion or launch."[82]

Because of the intense time pressures that would loom large with the decision to use nuclear weapons – especially where a state maintains a launch-on-warning posture – AI-enhanced cyber-attacks against nuclear systems would be almost impossible to detect and authenticate. Further, warning signals would be difficult to authenticate, let alone attribute, to initiate a nuclear strike within a short timeframe. A shared concern of China, the United States, and Russia – albeit with varying degrees of sensitivity – are the potential threats posed by AI-augmented cyber-warfare that might impel states to adopt (or rely more heavily upon) a launch-on-warning nuclear posture or a policy of pre-emption during a crisis.[83] In short, conventional strikes against an adversary's nuclear-deterrent forces (especially NC3) combining AI-enabled ISR and cyber-capabilities would likely amplify the potentially destabilising impact of such an operation.

Advances in AI could also exacerbate this cybersecurity challenge by enabling improvements to the cyber-offense. By automating advanced persistent threat (APT) operations, machine-learning and AI could dramatically reduce the extensive resources and skill required to execute APT operations (or "hunting for weaknesses"), especially against hardened nuclear targets. AI-augmented cyber-tools' machine speed could enable an attacker to exploit a narrow window of opportunity to penetrate an adversary's cyber-defences or use APT tools to find new vulnerabilities faster and easier than before. As former US Chairman of the Joint Chiefs, General Joseph Dunford, recently warned, "the *accelerated speed of war* ensures the ability to recover from early missteps is greatly reduced" (emphasis added).[84] For example, when docked for maintenance, air-gapped SSBNs, considered secure when submerged, could become increasingly vulnerable to a new generation of low-cost and highly automated APT cyber-attacks.[85] An attacker might also use machine-learning tools to target autonomous dual-use early-warning

systems with "weaponised software" (i.e. hacking, subverting, spoofing, or tricking), causing random and potentially undetectable errors, malfunctions, and behavioural manipulation to these networks.[86] An attacker may, for instance, poison a data-set to inhibit an algorithm from learning specific patterns, or insert a secret backdoor that can trick the system in the future.[87]

Chinese analysts would view cyber-infiltrations on China's NC3 systems as highly escalatory, even if the perpetrator's goals were limited to collecting information (i.e. espionage) about cyber-threats to prevent a future attack.[88] By contrast, Russian analysts tend to view Russia's NC3 network as relatively isolated and insulated from cyber-attacks.[89] However, both Chinese and Russian analysts worry about the vulnerability of their NC3 systems to fast-paced and stealthy conventional counterforce operations augmented by AI technology. The discovery of an adversary's attempt (successful or otherwise) to degrade a state's nuclear network would heighten mistrust and tension in future nuclear crises.[90]

Uncertainty about the efficacy of AI-augmented cyber-capabilities during a crisis or conflict would likely reduce *both* sides' risk tolerance, increasing the incentive to strike pre-emptively as a hedging strategy.[91] For instance, a state might strike pre-emptively in response to information mined from AI-augmented ISR systems that an adversary was planning a surprise attack. During crisis conditions, an offensive AI cyber-tool that succeeds in compromising an adversary's nuclear weapon systems – resulting in an "asymmetric information" situation – may cause either or both sides to overstate (or understate) its retaliatory capabilities, thus making them more inclined to act in a risky and escalatory fashion.[92] In short, in a competitive strategic environment where states are inclined to assume the worst of others' intentions, one state's efforts to enhance its strategic forces' survivability may be viewed by others as a threat to their nuclear retaliatory capability.[93]

Drone swarming under the nuclear shadow

How might AI-augmented drones swarming and hypersonic weapons complicate missile-defence, undermine states' nuclear deterrent forces, and increase the risk of escalation? Drones (especially micro-drones[94]) used in swarms are conceptually well-suited to conduct pre-emptive attacks and nuclear-ISR missions against an adversary's nuclear mobile missile launchers and SSBNs and their enabling facilities (e.g. early-warning systems, antennas, sensors, and air intakes).[95] In short, the ability of future iterations of AI machine-learning – mining expanded and dispersed data pools – infused drone swarming technology to locate, track, and target strategic missiles (i.e. mobile ICBM launchers in underground silos and onboard stealth aircraft or SSBNs) is set to grow.[96]

The following four scenarios illustrate the possible strategic operations that AI-augmented drone swarms would execute.[97] First, drone swarms could be deployed to conduct ISR operations to locate and track dispersed (nuclear and conventional) mobile missile launchers and their attendant dual-use command, control, communications, and intelligence (C3I) systems.[98] Specifically, swarms incorporating AI-infused ISR, autonomous sensor platforms, ATR, and data analysis systems may enhance sensor drones' effectiveness and speed to locate mobile missiles and evade enemy defenses.

Second, swarming may enhance legacy conventional and nuclear weapon delivery systems (e.g. ICBMs and SLBMs), potentially incorporating hypersonic variants. At least two nuclear-armed states have developed UAV, or unmanned underwater vessels (UUV), prototypes with nuclear delivery optionality.[99] AI applications will likely enhance the delivery system targeting, tracking, and improving drone swarms' survivability against the current generation of missile defenses.

Third, swarming tactics could bolster a states' ability to disable or suppress an adversary's defences and clearing the path for a disarming attack.[100] Drone swarms might be armed with cyber- or EW-capabilities (in addition to anti-ship, anti-radiation, or regular cruise and ballistic-missiles) to interfere with or destroy an adversary's early-warning detection and C3I systems in advance of a broader offensive campaign.[101]

Finally, in the maritime domain, UUVs, UAVs, and unmanned surface vessels (USVs), and supported by AI-enabled intra-swarm communication and ISR systems, may be deployed in *both* offensive and defensive anti-submarine warfare operations to saturate an enemy's defences and to locate, disable, and destroy its nuclear-armed or non-nuclear attack submarines.[102] Perceived as a relatively low-risk *force majeure* with ambiguous rules of engagement and absent a robust normative and legal framework,[103] lethal and non-lethal autonomous weapons will likely become an increasingly attractive asymmetric capability to erode a superior adversary's deterrence and resolve.[104] Notwithstanding the remaining technical challenges – especially the demand for power – swarms of robotic systems fused with AI machine learning techniques may presage a powerful interplay of increased range, accuracy, mass, co-ordination, intelligence, and speed in a future conflict. How might new technologies undermine deterrence between nuclear-armed adversaries in conflict-prone – and possibly asymmetric – situations?

Deterrence in digital multipolarity

The multipolar nuclear world order, characterised by geopolitical tension and security competition, differs substantially from the Cold War-era bipolar dynamics.[105] Nuclear multipolarity will likely challenge several long-held nuclear deterrence assumptions – especially the probability that actors can be coerced to behave in particular ways premised on the assumption of rationality – that significantly impair states' ability to manage escalation and interpret signalling (i.e. deterrence and resolve). Thereby increasing the likelihood of impregnating misperception and miscalculation into fragile strategic dyads – and possibly triads of nuclear-armed states.[106] This multipolarity is important because each state will likely choose different responses to the new choices emerging in the digital age.[107] Motivated states such as China and Russia might eschew the limitations of AI (e.g. AI's brittleness, explainability, interpretability, and vulnerability to adversarial attack[108]), thereby compromising safety and verification standards to secure or capture the first-mover advantages – or to offset a perceived strategic disadvantage at a conventional level – vis-à-vis an adversary on the future digitised battlefield.[109]

During a crisis, when prudent and careful planning can run aground in complex situations with likely an abundance of ambiguous information (about resolve and power), misperceptions and miscalculation may generate temptations for pre-emption.[110] Against this geopolitical backdrop, the use of autonomous weapon systems (perceived as low-risk and low-cost) during tense and complex adversarial environments – with

ambiguous rules of engagement such as anti-access/area denial zones – will become an increasingly enticing asymmetric option to undermine an adversary's military readiness, deterrence, and resolve.

In a world of revisionist and dissatisfied nuclear-armed states, it seems improbable that improvements in intelligence collection and analysis derived from advances in AI would have a stabilising impact.[111] For this to happen, equal access to intelligence and shared confidence in these systems' accuracy and credibility would be required. Further, all parties' intentions would need to be benign for any reassurances or confidence-building efforts to succeed. Because nuclear interactions increasingly involve the complex interplay of nuclear and non-nuclear (and state and non-state) actors, the leveraging of AI in this multipolar context will increasingly place destabilising pressures on nuclear states. These interactions will likely complicate escalation management efforts during future crises or conflict – especially involving China and the United States.[112]

Regime type and deterrence

How might regime type affect these dynamics? Authoritarian states may perceive an adversary's intentions very differently from a democratic one.[113] The belief that a regime's political survival or legitimacy is threatened might cause leaders to consider worst-case scenario judgments and behave in a manner predicted by offensive realist scholars.[114] Conversely, non-democratic leaders operating in closed political systems such as China, Russia, or North Korea may exhibit a higher degree of confidence (or overconfidence) in their ability to respond to perceived threats in world politics.[115] Bias assessments from a non-democratic regime's (or "Stasi" type) intelligence services might reinforce a leader's faith – or a false sense of security – in their diplomatic skill and manoeuverability.[116]

Without institutionalised structures (i.e. a general staff system) connecting the intelligence services with the military and broader political context, decisions will likely be made in vacuums, with minimal checks and balances on the political leadership (or supreme leader), and a reduction in "bottom-up" (or a "fact searching" organisational culture) information flow – because of the fear of contradicting the leadership.[117] This situation can reinforce a distorted (or false) sense of reality, thus compounding the cognitive misperceptions of events that are already present. Social media – and other AI-enhanced information tools – that supply decision-makers with a continuous flow of near-real-time information might also complicate the practice of deterrence and escalation management before and during future crises.[118] Politics and the breakdown of human bargaining ultimately lead to conflict; thus AI and autonomous systems and other advanced technologies are the tools (or dependent variables) that can influence escalation dynamics.[119]

Furthermore, a regime that views its second-strike capabilities – especially its NC3 systems – as vulnerable or insecure (North Korea, Pakistan, or perhaps China) may be more inclined to automate its nuclear forces and launch postures. An insecure nuclear-armed regime in an asymmetric dyad – assuming adequate technical, economic, organisational, and political resources – could do the one or a combination of the following: (1) alter its nuclear force structure or doctrine (i.e. adopt a launch-on-warning posture, or reject a no-first-use commitment);[120] (2) increase the alert status of its

nuclear arsenals;[121] (3) modernise or expand its nuclear capabilities (both strategic and tactical nuclear weapons);[122] (4) develop AI capabilities to bolster deterrence;[123] or (5) use AI technology to automate its second-strike nuclear capability (i.e. launch policy and NC3 systems), *in extremis* pre-delegate launch decisions to machines – to deter a decapitating strike during an existential crisis.[124]

In sum: against the backdrop of geopolitical tensions, the pace of technological diffusion, information and signalling problems (i.e. information asymmetry and signalling resolve and intentions), divergences in regime type, philosophy of employment, and force structures, AI and autonomy will influence nuclear deterrence by decreasing stability and increasing escalation risk. How might machine-human interactions spark inadvertent or accidental escalation at a strategic level?

Automating strategic decision-making: A double-edged sword for deterrence?

Today, the potential tactical and operational impact of AI is qualitatively axiomatic, its effect at a strategic level remains, however, uncertain. AI systems that are programmed to pursue tactical and operational advantages, for example, aggressively, might misperceive (or simply ignore) an adversary's bid to signal to resolve (i.e. to de-escalate a situation) as a prelude to an imminent attack. These dynamics would increase the risks of inadvertent escalation and first-strike instability.[125] If commanders decide to delegate greater authority to inherently inflexible AI systems, the dehumanisation of future defense planning will undermine stability by significantly inhibiting induction. Human induction (i.e. the ability to form general rules from specific pieces of information) is a crucial aspect of defense planning, primarily to manage situations that require high levels of visual and moral judgment and reasoning.[126] Unwarranted confidence and reliance on machines – known as "automation bias" – in the pre-delegation of the use of force during a crisis or conflict, let alone during nuclear brinksmanship, might inadvertently compromise states' ability to control escalation.[127]

Data limitations coupled with constraints on the ability of AI algorithms to capture the nuanced, dynamic, subjective accurately, and changeable nature of human commanders – or theory-of-the-mind functions – will mean that for the foreseeable future strategic decision-making will remain a fundamentally human endeavour – albeit imbued with increasing degrees of interaction with intelligent machines.[128] Thus, AI will continue to include some human agency – especially in collaboration with machines – for managing the attendant issues associated with technological complexity and interdependence, avoiding, for now at least, the risks associated with pre-delegating the use of military force.

While human agency should ensure that the role of AI in the nuclear domain is confined to a predominately ancillary one, through the discharge of its "support role" (data collection and analysis, stockpile management, and decision-making support systems), it may still – and possibly unbeknownst to commanders – influence strategic decisions that involve nuclear weapons. In other words, the distinction between AI's impact at a tactical and strategic level is not a binary one.[129] Technology designed ostensibly to augment autonomous weapon systems (i.e. ISR remote sensing, target recognition, and battlefield situation awareness systems) will likely nonetheless inform and

shape strategic war-faring calculations.[130] In short, escalation at the tactical level could easily have *strategic effect*.

On the one hand, future AI-augmented C2 support tools may overcome many of the shortcomings inherent to human strategic decision-making during wartime (e.g. susceptibility to invest in sunk costs, skewed risk judgment, heuristics, and groupthink) with potentially stabilising effects. Further, faster and more reliable AI applications could also enable commanders to make more informed decisions during a crisis, improve the safety and reliability of nuclear support systems, strengthen the cyber-defenses of C2 networks, enhance battlefield situational awareness, and reduce the risk of human error caused by fatigue and repetitive tasks.[131] On the other hand, AI systems that allow commanders to predict the potential production, commissioning, deployment, and ultimately launch of nuclear weapons by adversaries will likely lead to unpredictable system behaviour and outcomes, which *in extremis* could undermine first-strike stability – the premise of MAD – making nuclear wars winnable.[132]

The US 2018 Nuclear Posture Review (NPR), for example, explicitly states that the DoD would pursue design support technologies such as machine learning to facilitate more effective and faster strategic decision-making.[133] Chinese analysts have also begun to research the use of big-data, and deep-learning AI techniques to enhance the processing speed and intelligence analysis of satellite images, support China's early warning capabilities and enable a "prediction revolution" in future warfare.[134] Besides, China has also applied AI to wargaming and military simulations and researched AI-enabled data retrieval and analysis from remote sensing satellites to generate data and insights that might be used to enhance Chinese early-warning systems, situational awareness, and improve targeting.[135]

Under crisis and conflict conditions, the deterrent effect of AI is predicated on the perceived risks associated with a particular capability it enables or enhances. The higher the uncertainty generated by a capacity, deploying AI-augmented capabilities in a crisis might encourage an adversary to act more cautiously and, in turn, bolster stability. Counter-intuitively, therefore, states may view the expanded automation of their NC3 systems as a way to manage escalation and strengthen deterrence, signalling to an adversary that any attack – or the threat of one – might trigger nuclear escalation. Put differently, if a nuclear-armed state used automation to reduce its flexibility during a crisis, and without the ability to signal this to an adversary, it would be akin to Herman Kahn's notion of "being drunk, blind, and without a steering wheel" in a game of chicken.[136] A *prima facie* argument exists that fusing NC3 systems with AI-augmented automated response mechanisms – akin to an enhanced version of Russia's primitive *Perimeter* or "Dead-Hand" – might resolve the logical paradox inherent with rational-based classical deterrence; predicated on the notion of mutual destruction and the will to retaliate, thereby ensuring mutual deterrence and improving stability.[137]

This perceived deterrent effect, coupled with the uncertainty caused by the introduction of AI into a situation, might incentivise states facing a militarily superior adversary to delegate decisions to machines (i.e. fully autonomous mode) to signal resolve during a crisis.[138] Because of the difficulty of demonstrating a posture like this *before* a crisis or conflict, this implicit threat – akin to the Dr. Strangelove doomsday machine farce (or parable) – may equally worsen crisis instability.[139] Moreover, the confusion and uncertainty that would result from mixing various (and potentially unknown) levels of human-

machine interactions, and AI is reacting to events – such as signalling and low-level conflict – in non-human ways (using force where a human commander would not have), and at machine speed, could dramatically increase inadvertent risk. The recent defeat of a human pilot by an AI system in a DARPA-hosted Alpha Dogfight Challenge demonstrated how AI's performing in complex physics in a dynamic (albeit virtual) environments can compress the observe, orient, decide, and act (OODA) decision-making loop and apply non-conventional tactics in a high-stakes game of human-to-machine chicken.[140]

Deterrence and human-to-machine interaction

How might the introduction of intelligent machines affect human-to-human deterrence? Recent experimental wargaming hosted by the RAND Corporation explored the effects of mixing various levels of humans and machine configurations on escalatory dynamics – signalling, decision-making, and de-escalation – during a crisis revealed some interesting preliminary findings.[141] The wargames' tentative findings demonstrated that where high levels of autonomy coincide with primarily human decision-making (or "humans on the loop"), escalation risk is generally lower. This hypothesis was attributed to the fact that human involvement in decisions allowed more time to de-escalate (e.g. devise off-ramps), and that humans are likely to have a better understanding of signalling (i.e. resolve, deterrent threat, desire to de-escalation, or reassurance) compared to an AI algorithm. In short, AI would likely be worse – or at least less reliable – than humans at understanding signalling involved in deterrence, particularly signalling de-escalation.[142]

Conversely, and most speculative, when decisions are primarily made by machines and combined with high levels of autonomy (or "humans out of the loop"), escalation risk is higher – but because of the lower human-risk, the perceived costs of miscalculation are lower. The potential deterrent effect of this futuristic configuration (i.e. machine vs. machine) is confounding. The removal of human decision-making and judgment from a crisis, and less risk to human life, would reduce the traditional risks associated with accidents in human-machine interactions. Thus, IR deterrence theory – tethered to human perception, intuition, signalling intent, and rationality – would be effectively redundant. The total absence of a normative deterrence framework – in particular, to signal to resolve to an adversary while simultaneously seeking to de-escalate a situation – may compress (or remove entirely) the various rungs of the inherently psychological escalation ladder framework. It may increase inadvertent escalation risks and complicate de-escalation and conflict termination – especially in asymmetric dyads where incentives to strike pre-emptively to achieve escalation dominance exist.[143]

The potentially escalatory effects of AI's tactical optimisation programming would likely be compounded by differences in adversaries' goal setting (an AI's priorities, value alignment, control, and off-ramps),[144] C2 organisation (centralised vs. decentralised), and the configuration of their human-machine interactions. Specifically, machine decision-making – designed to exploit the tactical and operational advantages in a situation – may lack the "theory of the mind" in *a priori* situation with their interaction with humans.[145] Not only would machines need to understand human commanders and human adversaries, but they must also interpret an adversary AI's signalling and behaviour. An AI algorithm optimised to pursue pre-programmed goals might

misinterpret an adversary simultaneously signalling resolve who is seeking to avoid conflict or de-escalate a situation. Absent reliable means to attribute an actors' intentions, AI systems may convey undesirable and unintended (by human commanders) signals to the enemy, thus complicating the delicate balance between an actor's willingness to escalate a situation as a last resort and keeping the option open to step back from the brink.

Conclusion

This article argues that AI and autonomy could affect nuclear deterrence in two ways: (1) decreasing stability in nuclear multipolarity; and (2) increasing the tendency for (especially inadvertent) escalation to nuclear use. The article found that existing classical deterrence theories rooted in rational-based human actor assumptions are no longer applicable in the light of recent developments in AI and autonomy. It builds on the nascent "fifth wave" of modern post-classical deterrence theorising, which considers the potential implications of introducing non-human agents – and the corresponding disengagement or even removal of human agents – from strategic interactions.

The employment of AI and autonomy in the nuclear enterprise entails a multitude of trade-offs and open questions that should persuade designers, policymakers, and operators to explore ways in which new AI-powered capabilities might strengthen or complicate deterrence in the fragile nuclear "balance of terror,"[146] how adversaries think about these developments and the effect of synthesising legacy nuclear systems (especially NC3) with AI technology. Examples of these trade-offs *inter alia* include.[147] First, ensuring adequate control and supervision of AI's to mitigate escalation risk, while exploiting the potential benefits of increased lethality, scale, and speed afforded by this technology. Second, assimilating AI technology into force structure and doctrine, while mitigating the dangers posed by the speed and reduced levels of control of nuclear weapons. Third, instilling AI-augmented C2 systems with sufficient knowledge of the laws of engagement – and other legal and normative war-faring frameworks – without degrading AI's capacity to predict and respond to threats – especially confrontation that arise from introducing new capabilities into a situation (i.e. vertical escalation).[148] The risks of inaction are, therefore, great.

Objectively appraising and reconciling these trade-offs will likely be complicated by people's tendency to avoid making decisions when faced with incommensurable problems containing moral and ethical choices.[149] A reticence to engage in these issues, or worse, down-play the potential risks associated with AI and autonomy, would make it more challenging to alter incentives to enhance strategic stability and shape deterrence and escalation as the technology matures, and the military use of these systems inevitably increases. Cognisant that some states have or plan to deploy AI systems in their nuclear deterrence structures (e.g. high-precision missile systems, missile-defences, cyber-offense, electronic-warfare, and physical security), experts generally agree that AI requires further experimentation, testing, and verification before being integrated into nuclear support systems.[150]

How might militaries develop and deploy AI to steer it towards ensuring mutual deterrence and enhancing strategic stability? To improve strategic stability in an era of rapid technological change, great power strategic competition, and nuclear multipolarity, the formulation of future arms control frameworks must reflect the shifting perspectives

described in this study. How might the existing nuclear arms control regime be reconfigured and broadened to incorporate emerging technologies like AI?[151] To be sure, arms control efforts can no longer be restricted to bilateral engagement. The Nuclear Non-Proliferation Treaty (NPT) provides a successful case study in global governance that minimised the threat posed by the weaponisation of new (i.e. atomic) technologies while enabling the mutual benefits of sharing nuclear technology to strengthen strategic stability. The dual-use and diffused nature of AI compared to nuclear technology will, however, make arms control efforts particularly problematic. Moreover, when nuclear and non-nuclear capabilities and war-faring are blurred, strategic competition and arms racing are more likely to emerge, complicating arms control efforts.[152] In short, legacy arms control frameworks, norms, and even the notion of strategic stability itself will increasingly struggle to assimilate and respond to these fluid and interconnected trends.

Governments should also explore ways to increase transparency and accountability for AI and national security, such as addressing the implications of deepfakes and lethal autonomous weapons.[153] To counter the threat posed by non-state actors using AI-enabled tools such as deepfakes to manipulate, deceive, or otherwise interfere with strategic decision-making systems in misinformation attacks, governments should co-ordinate with both allies' adversaries – continue to harden NC3 systems and processes against cyber-attacks.[154] Other measures that may also improve stability, among other things, include: reducing the number of nuclear weapons; taking arsenals off high-alert (or launch-on-warning) status; separating warheads from delivery systems (or de-matting warheads); shifting to a deterrent-only (or minimum deterrence) force posture, and adopting a no first use declaratory policy – as China and India do today.[155]

Ultimately, success in these efforts will require all stakeholders to be convinced of the need and the potential mutual benefits of taking steps toward the establishment of a coherent governance architecture to institutionalise, internalise new norms, and ensure compliance with the design and deployment of AI and autonomy in the military sphere. Future research would be beneficial to address research puzzles, including: How might autonomous machines be imbued with post-classical deterrence theory's key principles in a comprehensive fashion? What could happen were different states to embed their intelligent machines with different interpretations or deterrence approaches? How might AI and autonomy impact deterrence strategies across multiple domains, regions, and between nuclear and non-nuclear states in extended deterrence scenarios?

Notes

1. Recent progress in AI falls within two distinct fields: (1) "narrow" AI and the machine-learning AI sub-set; and (2) "general" AI, which refers to AI with the scale and fluidity akin to the human brain. Most AI researchers anticipate that "general" AI to be at least several decades away, if at all. See, Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Harlow: Pearson Education, 2014); Nils J. Nilsson, *The Quest for Artificial Intelligence* (New York, NY: Cambridge University Press, 2010). "Autonomy" can be defined as the condition or quality of being self-governing to achieve an assigned task, based on a systems' situational awareness (integrated sensing, perceiving, and analysing), planning, and decision making. A distinction is often made between

automatic, automated, and autonomous systems, while others use these terms interchangeably. See, US Department of Defense, *Directive 3000.09, Autonomy in Weapon Systems*.

2. For example, see Michael C. Horowitz, Sarah E. Kreps, and Matthew Fuhrmann, 'Separating Fact from Fiction in the Debate Over Drone Proliferation', *International Security* 41, no. 2 (Fall 2016): 7–42; Andrew Ilachinski, *Artificial Intelligence & Autonomy: Opportunities and Challenges* (Arlington, VA: CNA, 2017); Mary L. Cummings, *Artificial Intelligence and the Future of Warfare* (London: Chatham House, The Royal Institute of International Affairs, 2017); and Elsa B. Kania, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power* (Washington, DC: Center for a New American Security, 2017).
3. See Michael C. Horowitz, 'Artificial Intelligence, International Competition, and the Balance of Power', *Texas National Security Review* 1, no. 3 (2018): 37–57; and Itai Barsade and Michael C. Horowitz, 'Artificial Intelligence Beyond the Superpowers', *Bulletin of the Atomic Scientists* (2018).
4. Notables studies that consider the implications of AI for deterrence and nuclear stability include: Wong et al., *Deterrence in the Age of Thinking Machines*; Edward Geist and Andrew Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Santa Monica, CA: RAND Corporation, 2018); Michael Horowitz, Paul Scharre, and Alex Velez-Green, *A Stable Nuclear Future? The Impact of Automation, Autonomy, and Artificial Intelligence* (Philadelphia: University of Pennsylvania, 2017); and James Johnson, 'Artificial Intelligence in Nuclear Warfare: A Perfect Storm of Instability?' *The Washington Quarterly*. doi:10.1080/0163660X.2020.1770968.
5. The literature is expansive; notable studies include, *inter alia*, Lawrence Freedman, *Deterrence* (New York, NY: Polity, 2004); Lawrence Freedman, *The Evolution of Nuclear Strategy* (Basingstoke: Palgrave Macmillan 2003); Thomas Schelling, *Arms and Influence* (New Haven and London: Yale University Press, 1966); Glen H. Snyder, *Deterrence and Defense: Towards a Theory of National Security* (Princeton, NJ: Princeton University Press, 1961); and Partick M. Morgan, *Deterrence Now* (Cambridge, UK: Cambridge University Press, 2003); and Michael Quinlan, *Thinking About Nuclear Weapons: Principles, Problems, Prospects* (Oxford: Oxford, University Press 2009).
6. See, Michael D. Intriligator and Dagobert L. Brito, 'Minimizing the Risks of Accidental Nuclear War: An Agenda for Action', in *Inadvertent Nuclear War*, eds. Hakan Wilberg, Ib Damgaard Petersen, and Paul Smoker (New York, NY: Pergammon Press, 1993), 228.
7. For scholarship on the intersection of technology with the potential for war see, James D. Fearon, 'Cooperation, Conflict, and the Costs of Anarchy', *International Organization* 72, no. 3 (2018): 523–59; for escalation dynamics see, Caitlin Talmadge, 'Emerging Technology and Intra-War Escalation Risks: Evidence from the Cold War, Implications for Today', *Journal of Strategic Studies* 42, no. 6 (2019): 864–87; for arms racing and strategic stability see, Todd S. Sechser, Neil Narang, and Caitlin Talmadge, 'Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War', *Journal of Strategic Studies* 42, no. 6 (2019): 727–35; and for deterrence and the offense-defense balance see, Ben Garfinkel and Allan Dafoe, 'How Does the Offense-Defense Balance Scale?' *Journal of Strategic Studies* 42, no. 6 (2019): 736–63; and Joseph S. Nye, 'Deterrence and Dissuasion in Cyberspace', *International Security* 41, no. 3 (2017): 44–71.
8. Notable exceptions include Herbert Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace', *Strategic Studies Quarterly* 6, no. 3 (Fall 2012): 46–70; and Erik Gartzke and Jon R. Lindsay, 'Thermonuclear Cyberwar', *Journal of Cybersecurity* 3, no. 1 (February 2017): 37–48.
9. James Johnson, 'Artificial Intelligence & Future Warfare: Implications for International Security', *Defense & Security Analysis* 35, no. 2 (2019): 147–69.
10. For example, see Bernard Brodie, *Escalation and the Nuclear Option* (Princeton, NJ: Princeton University Press 1965); Snyder, *Deterrence and Defense: Towards a Theory of National Security*; Thomas Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960), and Schelling, *Arms and Influence*.

11. Freedman, *Deterrence*, 122.
12. During the Cold War, deterrence was viewed almost solely in terms of nuclear weapons. However, nuclear-armed states have also sought to deter - both states and non-state - aggression and developed advanced conventional military capabilities to underwrite its 'conventional deterrence' postures. Michael S. Gerson, 'Conventional Deterrence in the Second Nuclear Age', *Parameters* (Autumn 2009): 32–48.
13. Freedman, *Deterrence*, 110–11.
14. Alex S. Wilner, 'US Cyber Deterrence: Practice Guiding Theory', *Journal of Strategic Studies* 43, no. 2 (2020): 245–80.
15. Some have argued that policymakers by confusing "rational" with "reasonableness" can mean that when adversaries fail to act as intended – or as a particular theory predicts – they appear irrational. An actor can be "rational" within their particular belief structure or theoretical framework of understanding. Kenneth Payne, *The Fallacies of Cold War Deterrence and a New Direction* (Lexington, KY: University of Kentucky, 2001), 7–15.
16. For example, President John F. Kennedy's public pledge in 1962 not to invade Cuba – assuring the Soviets that capitulation would not invite future demand - helped resolve the missile crisis.
17. A deterrent effect can result from military threats of force and non-military forces such as economic, diplomatic, and political inducements and assurances. Freedman, *Deterrence*, 59.
18. Robert Jervis, *How Statesmen Think: The Psychology of International Politics* (Princeton, NJ: Princeton University Press, 2017), 192.
19. James Fearon, 'Rationalist Explanations for War', *International Organization* 49, no. 3 (1995): 379–414.
20. Jeffrey W. Knopf, 'The Fourth Wave in Deterrence Research', *Contemporary Security Policy* 31, no. 1 (2010): 26.
21. Todd S. Sechser, 'Goliath's Curse: Coercive Threats and Asymmetric Power', *International Organization* 64, no. 4 (Fall 2010): 627–60; and Knopf, 'The Fourth Wave in Deterrence Research', 646.
22. Jonathan Mercer, *Reputation and International Politics* (Ithaca, NJ: Cornell University Press, 1996); Robert Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976); and Ted Horf, *Peripheral Visions: Deterrence Theory and American Foreign Policy in the Third World* (Ann Arbor: University of Michigan Press, 1994).
23. Ibid., 111.
24. Sechser, 'Goliath's Curse: Coercive Threats and Asymmetric Power', 627–60; and Knopf, 'The Fourth Wave in Deterrence Research', 1–33.
25. Thomas Rid, 'Deterrence Beyond the State: The Israeli Experience', *Contemporary Security Policy* 33, no. 1 (2012): 124–47; and Amir Lupovici, 'The Emerging Fourth Wave of Deterrence Theory', *International Studies Quarterly* 54 (2010): 705–32.
26. Beyond the military sphere, deterrence theory also permeates a broad spectrum of human behavior, including criminology and cognitive-psychological disciplines. For example, see Anthony Braga and David Weisburd, 'The Effects of Focused Deterrence Strategies on Crime', *Journal of Research in Crime and Delinquency* 49 (2012); Rob Guerette and Kate Bowers, 'Assessing the Extent of Crime Displacement and Diffusion of Benefits', *Criminology* 47, no. 4 (2009); Trevor Bennett, Katy Holloway, and David Farrington, 'Does Neighborhood Watch Reduce Crime?' *Journal of Experimental Criminology* 2, no. 4 (2006).
27. Beyond the military sphere, deterrence theory also permeates a broad spectrum of human behavior, including criminology and cognitive-psychological disciplines. For example, see Anthony Braga and David Weisburd, 'The Effects of Focused Deterrence Strategies on Crime', *Journal of Research in Crime and Delinquency* 49 (2012); Rob Guerette and Kate Bowers, 'Assessing the Extent of Crime Displacement and Diffusion of Benefits', *Criminology* 47, no. 4 (2009); Trevor Bennett, Katy Holloway, and David Farrington, 'Does Neighborhood Watch Reduce Crime?' *Journal of Experimental Criminology* 2, no. 4 (2006).

28. See, Gerson, 'Conventional Deterrence in the Second Nuclear Age', 32–48; Daniel Sobelman, 'Learning to Deter', *International Security* 41, no. 3 (2016/17); and Linton Brooks and Mira Rapp-Hooper, 'Extended Deterrence, Assurance, and Reassurance in the Pacific During the Second Nuclear Age', in *Strategic Asia 2013–14: Asia in the Second Nuclear Age*, ed. Ashley J. Tellis, Abraham M. Denmark, and Travis Tanner (Seattle, Washington: National Bureau of Asian Research, 2013), 266–300.
29. Wilner, 'US Cyber Deterrence', 7–8.
30. Tim Prior, 'Resilience: The 'Fifth Wave' in the Evolution of Deterrence', in *Strategic Trends 2018*, ed. Oliver Thränert and Martin Zapfe (Zurich, Switzerland: Center for Security Studies, ETH Zurich, 2018), 63–80; Beyza Unal, Yasmin Afina, and Patricia Lewis, 'Perspectives on Nuclear Deterrence in the 21st Century', *International Security Program Research Paper*, Chatham House (April 2020); and Stephen L. Quackenbush and Frank C. Zagare, 'Modern Deterrence Theory: Research Trends, Policy Debates, and Methodological Controversies', *Oxford Handbooks Online*, Published online, May 2016. doi:10.1093/oxfordhb/9780199935307.013.39.
31. Reid B.C. Pauly, 'Would US Leaders Push the Button? Wargames and the Sources of Nuclear Restraint', *International Security* 43, no. 2 (2018): 151–92; Jacquelyn G. Schneider, 'Cyber Attacks on Critical Infrastructure: Insights from War Gaming', *War on the Rocks* (2017); and Erik Lin-Greenberg, 'Wargame of Drones: Remotely Piloted Aircraft and Crisis Escalation' August 22, 2020; *SSRN*, https://ssrn.com/abstract=3288988.
32. Morgan E. Forrest et al., *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, CA: RAND Corporation, 2008), 8.
33. Herman Kahn, *On Escalation: Metaphors and Scenarios* (New York, NY: Praeger, 1965).
34. Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (New York: Basic Books, 1984); and Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton, NJ: Princeton University Press, 1993); and Thomas C. Schelling, *Arms and Influence* (New Haven, CT: Yale University Press, 1966), 227–28.
35. Stephen J. Cimbala, *The Dead Volcano: The Background and Effects of Nuclear War Complacency*, (Westport, CT: Praeger, 2002), 147.
36. Deterrence theorists have frequently used World War I to study how rapid and inflexible military systems can lead to inadvertent escalation and war. See, Marc Trachtenberg, 'The Meaning of Mobilization in 1914', in *Military Strategy and the Origins of the First World War*, ed. Steven E. Miller, Sean M. Lynn-Jones, and Stephen Van Evera (Princeton, NJ: Princeton University Press, 1991), 195–97.
37. Lin, 'Escalation Dynamics and Conflict Termination in Cyberspace'.
38. Kahn, *On Escalation: Metaphors and Scenarios*, 285.
39. Sico van der Meer, 'Reducing Nuclear Weapons Risks: A Menu of 11 Policy Options', *Policy Brief*, Clingendael Netherlands Institute of International Relations, June 2018.
40. Miles, 'The Dynamics of Strategic Stability and Instability', 429 and 437.
41. Intriligator and Brito, 'Minimizing the Risks of Accidental Nuclear War', 230–30.
42. Jervis, *Perception and Misperception in International Politics*, chapter 3.
43. Military autonomous systems and inadvertent risks associated with these systems is not a new phenomenon. Examples of existing autonomous systems include landmines, torpedoes, the *Aegis* and *Patriot* missile defence systems, and the anti-missile close-in Phalanx system. For analysis of errors involving these systems see, Yuna Huh Wong et al., *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020), 71–72.
44. See Thomas C. Schelling and Morton Halperin, *Strategy and Arms Control* (New York: Twentieth Century Fund, 1961).
45. Robert Jervis, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon* (Ithaca, NY: Cornell University Press, 1989).
46. For example, see Eric Schlosser, *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety* (New York: Penguin, 2013); and Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton, NJ: Princeton University Press, 1993).

47. Michael Gerson, 'The Origins of Strategic Stability', in *Strategic Stability: Contending Interpretations*, ed. Colby Elbridge and Michael Gerson (Carlisle, PA: Army War College, 2013), 1–46.
48. Thomas C. Schelling, *Arms and Influence* (New Haven, CT: Yale University Press, 1966), 234.
49. Michael Gerson, 'The Origins of Strategic Stability', in *Strategic Stability: Contending Interpretations*, ed. Colby Elbridge and Michael Gerson (Carlisle, PA: Army War College, 2013), 26.
50. Ronald F. Lehman, 'Future Technology and Strategic Stability', in *Strategic Stability: Contending Interpretations*, ed. Colby Elbridge and Michael Gerson (Carlisle, PA: Army War College, 2013), 147.
51. Charles Duelfer and Stephen Dyson, 'Chronic Misperception and International Conflict: The U.S.-Iraq Experience', *International Security* 36, no. 1 (Summer 2011): 75–78.
52. "Strategic stability" and "strategic instability" are not necessarily mutually exclusive states, especially in a nuclear multipolar system.
53. Cimbala, *The Dead Volcano*, 66.
54. Jervis, *How Statesmen Think*, chapter 4.
55. Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960).
56. Robert Jervis, *How Statesmen Think*, 222.
57. Similar to 'strategic stability', the concept of 'arms race' is also contested. Barry Buzan and Eric Herring, *The Arms Dynamic in World Politics* (London: Boulder & London Lynne Reinner), 77.
58. Schelling and Halperin, *Strategy and Arms Control*.
59. Jervis, *How Statesmen Think*, 95.
60. Aaron R. Miles, 'The Dynamics of Strategic Stability and Instability', *Comparative Strategy* 35, no. 5 (2016): 423–37.
61. These psychological factors raise questions about the assumption during a crisis commander can depend on a common sense of rationality with the other side. B. A. Thayer, 'Thinking about Nuclear Deterrence Theory: Why Evolutionary Psychology Undermines Its Rational Actor Assumptions', *Comparative Strategy* 26, no. 4 (2007): 311–23.
62. Jervis *How Statesmen Think*, 219.
63. Joseph Johnson, 'MAD in an AI Future?' *Center for Global Security Research, Lawrence Livermore National Laboratory*, June 3, 2019.
64. For example, see Michael J. Boyle, 'The Race for Drones', *Orbis* (November 24, 2014): 76–94.
65. Wong et al., *Deterrence in the Age of Thinking Machines*, 82.
66. Jervis Robert, 'Cooperation Under the Security Dilemma', *World Politics* 30 (1978): 167–214.
67. There is a distinction between the credibility of *capability* (in this case) and the credibility of *resolve* discussed below.
68. For a discussion of the impact of the arms race and effect on arms racing dynamics and military drones, see Michael J. Boyle, 'The Race for Drones', *Orbis* (2014): 76–94.
69. Jervis, *The Meaning of the Nuclear Revolution*.
70. See, Keir A. Lieber and Darryl G. Press, 'The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence', *International Security* 41, no. 4 (Spring 2017): 9–49; and Paul Bracken, 'The Cyber Threat to Nuclear Stability', *Orbis* 60, no. 2 (2016): 188–203.
71. Keir A. Lieber and Daryl G. Press, 'Why States Won't Give Nuclear Weapons to Terrorists', *International Security* 38, no. 1 (2013): 80–104.
72. The technical feasibility of this hypothesis is highly contested, however. See Sebastian Brixey-Williams, *Will the Atlantic Become Transparent?* 2nd ed. *British Pugwash*, November 2016.
73. Defense experts estimate that while AI will solve some of these challenges in the near-term, the many technical problems associated with tracking and targeting mobile missiles are

unlikely to be overcome within the next two decades. Geist and Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* 16.

74. Lieber and Press, 'The New Era of Counterforce', 9–49.
75. US Defense Advanced Research Projects Agency (DARPA), 'ACTUV Sea Hunter prototype transitions to the US Office of Naval Research for further development', January 30, 2018.
76. Tong Zhao and Li Bin, 'The Underappreciated Risks of Entanglement: A Chinese Perspective', in *Entanglement: Russian and Chinese Perspectives on Non-Nuclear Weapons and Nuclear Risks*, ed. James M. Acton (Washington, DC: Carnegie Endowment for International Peace, 2017), 47–75.
77. In an asymmetric dyad, a weaker state's perception about the future (i.e., relative power and resolve) is likely to diverge with more powerful rivals, because stronger states have incentives to understate the future while weaker ones have incentives to overstate it – the existence of uncertainty in asymmetric relationship can lead to the failure of bargaining and deterrence. Erik Gartzke, 'War Is in the Error Term', *International Organization* 53, no. 3 (Summer, 1999): 584.
78. Caitlin Talmadge, 'Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States', *International Security* 41, no. 4 (Spring 2017): 50–92.
79. Bradley Martin et al., *Advancing Autonomous Systems: An Analysis of Current and Future Technology for Unmanned Maritime Vehicles* (Santa Monica, CA: RAND Corporation, 2019).
80. See, Owen R. Cote Jr., 'Invisible Nuclear-Armed Submarines, or Transparent Oceans? Are Ballistic Missile Submarines Still the Best Deterrent for the United States?' *Bulletin of the Atomic Scientists* 75, no. 1 (2019): 30–35.
81. See, Jonathan Gates, 'Is the SSBN Deterrent Vulnerable to Autonomous Drones?' *The RUSI Journal* 161, no. 6 (2016): 28–35.
82. Andrew Futter, *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Washington, DC: Georgetown University Press, 2018), 164.
83. China and Russia have already taken steps towards placing their retaliatory forces on higher alert, and the fear of a US cyber-attack against their nuclear deterrent forces could prompt them to accelerate efforts to place more of their nuclear weapons on hair-trigger alert. Gregory Kulacki, 'China's Military Calls for Putting Its Nuclear Forces on Alert', *Union of Concerned Scientists*, January 2016.
84. Joseph F. Dunford, speech quoted at Jim Garamone, 'Dunford: Speed of Military Decision-Making Must Exceed Speed of War', US Department of Defense, January 31, 2017.
85. Paul Ingram, 'Hacking UK Trident: A Growing Threat', *BASIC*, May 31, 2017.
86. Machine learning systems rely on high-quality data-sets to train their algorithms; poisoning this data could lead these systems to perform unanticipated and potentially undetectable ways.
87. Eugene Bagdasaryan, et al. 'How to Backdoor Federated Learning' (preprint August 6, 2018), arXiv:1807.00459.
88. Acton et al., *Entanglement*, 81.
89. Ibid.
90. Gartzke and Lindsay, 'Thermonuclear Cyberwar', 37–48.
91. For analysis on how AI-augmented-cyber capabilities might affect nuclear deterrence, escalation, and stability, see James Johnson, 'The AI-Cyber Nexus: Implications for Military Escalation, Deterrence, and Strategic Stability', *Journal of Cyber Policy* 4, no. 3 (2019): 442–60.
92. See, Barton Whaley, 'Covert Rearmament in Germany, 1919-1939: Deception and Misperception', *Journal of Strategic Studies* 5, no. 1 (March 1982): 3–39; John J. Mearsheimer, *Why Leaders Lie: The Truth About Lying in International Politics* (Oxford: Oxford University Press, 2013); and Robert L. Jervis, *The Logic of Images in International Relations* (New York: Columbia University Press, 1989).
93. Schelling, and Halperin, *Strategy and Arms Control*, 30.

94. Micro-drone UAVs range from small flying insects to palm-sized devices equivalent to small birds and can carry payloads, including small guns and explosives and unconventional munitions such as poisons and nerve agents. See, Berenice Baker, 'Dogfighting Drones – Swarms of Unmanned Battle-Bots Take to the Skies', *Airforce-technology.com*, July 23, 2013, www.airforcetechnology.com/features/featuredogfight-drones-unmanned-battle-bot-swarms/.
95. The US Office of Naval Research (ONR), for example, envisions currently experimenting with swarms of unmanned surface vehicles to form a defensive perimeter around larger ships and surround enemy ships as part of the counterterrorism effort. Thomas Claburn, 'Navy Tests Swarming Autonomous Boats', *InformationWeek*, November 7, 2014, https://www.informationweek.com/messages.asp?piddl_msgthreadid=21052&piddl_msgorder=thrd.
96. Elias Groll, 'How AI Could Destabilize Nuclear Deterrence', *Foreign Policy*, April 24, 2018.
97. The value of drones in these scenarios does not mean that they are the *only* or necessarily most effective way to fulfill these missions. Gates, 'Is the SSBN Deterrent Vulnerable to Autonomous Drones?' 28–35.
98. The US, Russia, South Korea, and China are also actively pursuing drone swarm technology programs.
99. In 2015, Russia revealed a large nuclear-armed UUV delivery vehicle, *Poseidon* (also known as Status-6). The US is also developing an 'optionally manned' nuclear-capable long-range bomber, the B-21 *Raider*, carrying nuclear payloads. Ria Novosti, 'Russia Could Deploy Unmanned Bomber After 2040 – Air Force', *GlobalSecurity.org*, February 8, 2012, and Robert M. Gates, 'Statement on Department Budget and Efficiencies' (US Department of Defense, January 6, 2011).
100. Mike Pietrucha, 'The Need for SEAD Part 1: The Nature of SEAD', *War on the Rocks*, May 17, 2016.
101. Conversely, drone swarms might enhance states' missile defenses as countervails to this type of offensive threat. Polat Cevik, et al., 'The Small and Silent Force Multiplier: A Swarm UAV-Electronic Attack', *Journal of Intelligent and Robotic Systems* 70 (April 2013): 595–608.
102. For example, DARPA's Anti-Submarine Warfare Continuous Trail Unmanned Vessel (ACTUV) program, designed to track quiet diesel-electric submarines with USVs from the surface.
103. As technology advances – particularly AI – an increasing number of states will likely consider developing and operating Lethal Autonomous Weapons (LAWS). The international community considers the implications of LAWS in conversations held under the auspices of the United Nations Convention on Certain Conventional Weapons (CCW), a multilateral arms control agreement to which the United States became a party in 1982. The United States and Russia (albeit for different reasons) have consistently opposed banning LAWS. China supports a ban on the use, but not development, of LAWS, which it defines as indiscriminate – lethal systems that do not have any human oversight and cannot be terminated. However, some have argued that China is maintaining 'strategic ambiguity' about its position on LAWS. See, Kelly M. Sayler and Michael Moddie, 'International Discussions Concerning Lethal Autonomous Weapon Systems', *Congressional Research Service* (In Focus), October 15, 2020, https://fas.org/sgp/crs/weapons/IF11294.pdf.
104. Paul Scharre, *Autonomous Weapons and Operational Risk: Ethical Autonomy Project* (Washington, DC: Center for a New American Security, February 2016).
105. Stephen E. Miller, Robert Legvold, and Lawrence Freedman, *Meeting the Challenges of the New Nuclear Age: Nuclear Weapons in a Changing Global Order* (Cambridge, MA: American Academy of Arts and Sciences, 2019), 28–61.
106. Robert Einhorn and W. P. S Sidhu, *The Strategic Chain: Linking Pakistan, India, China, and the United States, Arms Control and Non-Proliferation Series Paper 14* (Washington, DC: The Brookings Institution, March 2017).
107. For literature on the "Second Nuclear Age" see Paul Bracken, *The Second Nuclear Age: Strategy, Danger, and the New Power Politics* (New York: Times Books, 2012); Colin S. Gray, *The*

Second Nuclear Age (Boulder, Colorado: Lynne Rienner, 1999); and Keith Payne, *Deterrence in the Second Nuclear Age* (Washington, DC: Georgetown University Press, 1996).

108. Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot, 'Making Machine Learning Robust Against Adversarial Inputs', *Communications of the ACM* 61, no. 7 (2018): 56–66; and Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and Harnessing Adversarial Examples' December 20, 2014, *arXiv* preprint arXiv:1412.6572.
109. For China's approach to military AI see, Elsa B. Kania, 'AI Weapons in China's Military Innovation', *Global China, Brookings*, April 2020, https://www.brookings.edu/research/ai-weapons-in-chinas-military-innovation/; and for Russia's see, 'Artificial Intelligence in Russia', Issue 12, *The Russia Studies Program, Center for Navy Analysis*, October 2020.
110. Game-theoretic analysis suggests that while significant differences exists in crises, they only need to raise the spectre of military conflict to generate reputations for resolve – actors who are able to signal that they place a low value on military costs during a crisis (or cost-tolerance) will likely obtain improved settlement terms. James. D. Fearon, 'Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs', *Journal of Conflict Resolution* 41, no. 1 (1997): 68–90.
111. Donald J. Trump, *National Security Strategy of the United States of America* (Washington, DC: The White House, December 2017).
112. Talmadge, 'Would China Go Nuclear?', 50–92.
113. James Johnson, 'Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?' *Journal of Strategic Studies* (2020). doi:10.1080/01402390.2020.1759038.
114. John J. Mearsheimer, 'The Gathering Storm: China's Challenge to US Power in Asia', *The Chinese Journal of International Politics* 3, no. 4 (Winter 2010): 381–96.
115. For example, see Yang Yaohui, 'A Vision of a New Kind of Combat Systems', *PLA Daily*, June 20, 2020; and Yang Feilong and Li Shijiang 'Cognitive Warfare: Dominating the Era of Intelligence', *PLA Daily*, March 19, 2020.
116. Keren Yarhi-Milo, *Knowing the Adversary* (Princeton NY: Princeton University Press, 2014), 250.
117. David E. Apter, *The Politics of Modernization* (Chicago, IL: Chicago University Press, 1967).
118. See, Michael J. Mazarr, et al., *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment* (Santa Monica, CA: RAND Corporation, 2019).
119. Adam P. Liff, 'Cyberwar: A New 'Absolute Weapon'? The Proliferation of Cyberwarfare Capabilities and Interstate War', *Journal of Strategic Studies* 35, no. 3 (2012): 401–28.
120. Reports indicate that China and India have considered making changes to their respective no-first-use (NFU) pledges. Zhenqing Pan, 'A Study of China's No-First-Use Policy on Nuclear Weapons', *Journal of Peace and Nuclear Disarmament* 1, no. 1 (2018): 115–36; and Harsh V. Pant Yogesh Joshi, 'Nuclear Re-think: A Change in India's Nuclear Doctrine has Implications on Cost & War Strategy', *Economic Times* (New Delhi), August 17, 2019.
121. China, India and Pakistan have reportedly shifted their policy of keeping nuclear warheads separate from delivery systems – in the context of expanding their nuclear submarine deterrence capacity.
122. Lauren J. Borja and M. V. Ramana, 'Command and Control of India's Nuclear Arsenal', *Journal for Peace and Nuclear Disarmament* 3, no. 1 (2020): 1–20; and Tim Craig and Karen De Young, 'Pakistan is Eyeing Sea-Based and Short-Range Nuclear Weapons, Analysts Say', *Washington Post*, September 21, 2014, https://www.washingtonpost.com/world/asia_pacific/pakistan-is-eyeing-sea-based-and-short-range-nuclear-weapons-analysts-say/2014/09/20/1bd9436a-11bb-11e4-8936-26932bcfd6ed_story.html.
123. Nuclear powers, including Russia, the United Kingdom, China, France, India, and the United States, have all released official policy statements that explicitly underscore the importance of AI technology as key military enablers in the multipolar world order. Vincent Boulanin, ed., *Artificial Intelligence, Strategic Stability and Nuclear* (SIPRI Publications, Stockholm, June 2020), 110.

124. For example, reports suggest that President Putin has considered reactivating Russia's controversial *Perimeter* (or "Dead-Hand") system. President of Russia, 'Meeting of the Valdai International Discussion Club', October 18, 2018, http://en.kremlin.ru/events/president/news/58848.
125. Ibid., 60–61.
126. Cummings, *Artificial Intelligence*, 7.
127. See, Parasuraman et al., 'Complacency and Bias in Human Use of Automation', 381–410; and Cummings, 'Automation Bias in Intelligent Time-Critical Decision Support Systems', 557–62.
128. Kenneth Payne, 'Fighting on Emotion and Conflict Termination', *Cambridge Review of International Affairs* 28, no. 3 (August 2015): 480–97.
129. Kenneth Payne, *Strategy from Apes to Evolution Artificial Intelligence and War* (Washington, DC: Georgetown University Press), 183.
130. Ibid.
131. Rebecca Hersman et al., 'Under the Nuclear Shadow: Situational Awareness Technology and Crisis Decision-Making', *Center for Strategic and International Studies*, March 18, 2020, https://ontheradar.csis.org/analysis/final-report/.
132. Keir A. Lieber and Daryl G. Press, 'The End of MAD: The Nuclear Dimension of US Primacy', *International Security* 30, no. 4 (Spring 2006): 7–44.
133. US Office of the Secretary of Defense, *Nuclear Posture Review* (Washington DC: Department of Defense, February 2018), 57–58.
134. Jia Daojin and Zhou Hongmei, 'The Future 20–30 Years Will Initiate Military Transformation', *China Military Online*, June 2, 2016.
135. From the open sources, no unambiguous evidence has emerged to suggest that China plans to use AI to augment its NC3 systems. Fiona S. Cunningham, and Taylor M. Fravel, 'Dangerous Confidence? Chinese Views on Nuclear Escalation', *International Security* 44, no. 2 (2019): 106–8.
136. Herman Kahn, *On Escalation: Metaphors and Scenarios* (New York: Prager, 1965), 11.
137. During the Cold War, the Soviet Union was so worried about this that they took a page from Dr. Strangelove, creating a very primitive version of an automated response mechanism capable of retaliating whether humans wanted to or not. In 2011, Commander-in-Chief of the Russian SRF, General S. Karakayev, confirmed in an interview with one of the prominent Russian newspapers that *Perimeter* – also known as Russia's "Dead hand Doomsday Weapon" – exists and continues to be on combat duty. The system's specific capabilities are unknown, however. See, Ryabikhin Leonid, 'Russia's NC3 and Early Warning Systems', *Tech4GS* July 11, 2019, https://www.tech4gs.org/nc3-systems-and-strategic-stability-a-global-overview.html; Eric Schlosser, *Command and Control* (New York, NY: Penguin Group, 2014); and Richard Rhodes, *Arsenals of Folly: The Making of the Nuclear Arms Race* (Simon & Schuster: London, 2008).
138. Yuna et al., *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020).
139. Johnson, 'Delegating Strategic Decision-Making to Machines'. doi:10.1080/01402390.2020.1759038.
140. In this simulated challenge, Maryland-based Heron Systems designed the AI system used both conventional and unconventional moves, including aggressive gunfire and flying toward their adversary at a close range without letting up until the last possible moment. Oriana Pawlyk, 'Rise of the Machines: AI Algorithm Beats F-16 Pilot in Dogfight', *Military.com*, August 24, 2020, https://www.military.com/daily-news/2020/08/24/f-16-pilot-just-lost-algorithm-dogfight.html.
141. Wong, *Deterrence in the Age of Thinking Machines*.
142. Ibid., 39–59.
143. "Escalation dominance" is a situation in which one side can escalate a conflict in ways that will be disadvantageous (or costly) to the adversary. The other side cannot respond in kind,

either because it has no escalation options or the options it has does not improve its predicament. Kahn, *On Escalation*.

144. Stuart Russell, *Human Compatible* (New York, NY: Viking Press, 2019).
145. "Theory of Mind" is the innate ability of most humans to perceive others' intentions and beliefs, allowing them to make predictions about others' behavior. See, Brittany N. Thompson, 'Theory of Mind: Understanding Others in a Social World', *Psychology Today*, July 3, 2017.
146. John D. Williams, *The Compleat Strategyst: Being a Primer on the Theory of Games of Strategy* (New York, NY: Dover Publications, 1986).
147. Yuna et al., *Deterrence in the Age of Thinking Machines*, 81–82.
148. Morgan et al., *Dangerous Thresholds*, 18–19.
149. For example, this explains why people who are opposed to the use of torture on moral grounds are generally resistant to arguments that its use may save lives. See, Alan Fiske and Philip Tetlock, 'Taboo Trade-offs: Reactions to Transactions that Transgress the Spheres of Justice', *Political Psychology* 18 (June 1997): 255–97.
150. The United States, for example, is known to have deployed AI technology ISR remote-sensing to support its autonomous Perdix UAV swarm operations and its Aegis ballistic missile defense systems' radar seekers. Boulanin, ed., *Artificial Intelligence, Strategic Stability and Nuclear*, 42–43.
151. Matthijs M. Maas, 'How Viable is International Arms Control for Artificial Military Intelligence? Three Lessons from Nuclear Weapons', *Contemporary Security Policy* 40, no. 3 (2019): 285–311.
152. Heather Williams, 'Asymmetric Arms Control and Strategic Stability: Scenarios for Limiting Hypersonic Glide Vehicles', *Journal of Strategic Studies* 42, no. 6 (2019): 789–813.
153. Joshua New, 'Why the United States Needs a National Artificial Intelligence Strategy and What It Should Look Like', *ITIF* December 4, 2018.
154. For example, Matt Turek, 'Semantic Forensics (SemaFor) Proposers Day', *Defense Advanced Research Projects Agency*, August 28, 2019.
155. National Security Commission on Artificial Intelligence (NSCAI) Interim Report to Congress, November 2019.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

James Johnson is an Assistant Professor in the School of Law and Government at Dublin City University and a Non-Resident Fellow with the Modern War Institute at West Point. Dr Johnson was previously a Postdoctoral Research Fellow at the James Martin Center for Nonproliferation Studies in Monterey, California. He the author of *The US–China Military & Defense Relationship during the Obama Presidency*. His latest book is titled, *Artificial Intelligence & the Future of Warfare: USA, China, and Strategic Stability*.