European Journal of International Security (2021), page 1 of 23
doi:10.1017/eis.2021.23

EJIS

RESEARCH ARTICLE

Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age

James Johnson*

Department of Politics and International Relations, University of Aberdeen, King's College, Aberdeen, United Kingdom
*Corresponding author. Email: james.johnson@abdn.ac.uk

(Received 9 April 2021; revised 6 September 2021; accepted 27 September 2021)

Abstract

Will AI-enabled capabilities increase inadvertent escalation risk? This article revisits Cold War-era thinking about inadvertent escalation to consider how Artificial Intelligence (AI) technology (especially AI augmentation of advanced conventional weapons) through various mechanisms and pathways could affect inadvertent escalation risk between nuclear-armed adversaries during a conventional crisis or conflict. How might AI be incorporated into nuclear and conventional operations in ways that affect escalation risk? It unpacks the psychological and cognitive features of escalation theorising (the security dilemma, the 'fog of war', and military doctrine and strategy) to examine whether and how the characteristics of AI technology, against the backdrop of a broader political-societal dynamic of the digital information ecosystem, might increase inadvertent escalation risk. Are existing notions of inadvertent escalation still relevant in the digital age? The article speaks to the broader scholarship in International Relations – notably 'bargaining theories of war' – that argues that the impact of technology on the cause of war occurs through its political effects, rather than tactical or operational battlefield alterations. In this way, it addresses a gap in the literature about the strategic and theoretical implications of the AI-nuclear dilemma.

Keywords: Inadvertent Escalation; Artificial Intelligence; Conventional Weapons; Nuclear Weapons; Emerging Technology; Information Ecosystem

Introduction

How might AI-enabled capabilities increase inadvertent escalation risk? This article revisits Cold War-era thinking about inadvertent escalation to consider how artificial intelligence (AI) technology[1] (especially AI augmentation of advanced conventional counterforce weapons) through

[1]Artificial Intelligence (AI) is an umbrella concept that describes a broad portfolio of applications that enable machines to emulate human intelligence capabilities, such as language, reasoning, learning, heuristics, and observation. Recent progress in AI falls into two distinct fields: (1) 'narrow' AI, which refers to involves statistical algorithms that learn procedures by analysing large training datasets designed to approximate and replicate human cognitive tasks; and (2) 'general' AI, which refers to AI with the scale and fluidity akin to the human brain. Narrow AI is already used in the private sector, particularly in data-rich research fields and applied sciences (for example, predictive analytics for market research, consumer behaviour, logistics, and quality control systems). Other emerging security-related technologies that AI machine-learning techniques might enable or be enhanced include autonomous weapons, robotics, 3D additive printing, quantum computing, 5G networks, semiconductors, and cyberspace. On AI, see Stuart Armstrong, Kaj Sotala, and Seán S. ÓhÉigeartaigh, 'The errors, insights, and lessons of famous AI predictions – and what they mean for the future', *Journal of Experimental and Theoretical Artificial Intelligence*, 26:3 (2014), pp. 317–42. On related technologies, see Reuben Steff, Simone Soare, and Joe Burton (eds), *Emerging Technologies and International Security: Machines, the State and War* (London, UK:

different mechanisms and pathways might influence inadvertent escalation risk between nuclear-armed adversaries during a conventional crisis or conflict. Are existing notions of inadvertent escalation still relevant in the digital age? We are now in an era of rapid disruptive technological change, especially in AI technology.[2] AI technology is already being infused into military machines, and global armed forces are well advanced in their planning, research and development, and in some cases, deployment of AI-enabled capabilities.[3] Therefore, the embryonic journey to reorient military forces to prepare for the future digitised battlefield is no longer merely speculation or science fiction.

While much of the recent discussion has focused on specific technical issues and uncertainties involved as militaries developed and diffuse AI applications at the tactical and operational level of war, the strategic and theoretical treatment of these developments (or the 'AI-nuclear strategic nexus') has received far less attention.[4] This article addresses this gap. It examines the psychological and cognitive features of escalation theorising to consider whether and how AI technology 'characteristics',[5] contextualised with the broader political dynamics associated with today's digital information ecosystem, may increase inadvertent escalation risk. It explains how AI technology could be incorporated into nuclear and conventional operations in ways that affect inadvertent escalation risks during a crisis or subnuclear conflict in strategically competitive dyads – US-China, India-Pakistan, and US-Russia. How might AI be incorporated into nuclear and conventional operations in ways that affect escalation risk?

The article speaks to the broader scholarship in International Relations – notably 'bargaining theories of war' (shifts in the balance of power, uncertainty, asymmetric information, and commitment problems),[6] deterrence theorising,[7] and political psychology[8] – that argues that the

Routledge, 2020); and Todd S. Sechser, Neil Narang, and Caitlin Talmadge, 'Emerging technologies and strategic stability in peacetime, crisis, and war', *Journal of Strategic Studies*, 42:6 (2019), pp. 727–35.

[2]See James Johnson, 'Artificial intelligence and future warfare: Implications for international security', *Defense & Security Analysis*, 35:2 (2019), pp. 147–69.

[3]See Vincent Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI Report, June 2020).

[4]Notable exceptions include: Kenneth Payne, 'Artificial intelligence: A revolution in strategic affairs?', *Survival*, 60:5 (2018), pp. 7–32; Michael C. Horowitz, 'When speed kills: Lethal autonomous weapon systems, deterrence and stability', *Journal of Strategic Studies*, 42:6 (2019), pp. 764–88; Mark Fitzpatrick, 'Artificial Intelligence and nuclear command and control', *Survival*, 61:3 (2019), pp. 81–92; Michael C. Horowitz et al., 'Strategic Competition in an Era of Artificial Intelligence', Artificial Intelligence and International Security (Center for New American Security, July 2018); and James Johnson, *Artificial Intelligence & the Future of Warfare: USA, China, and Strategic Stability* (Manchester, UK: Manchester University Press, 2021).

[5]AI technology 'characteristics' refers to specific technical attributes of artificial intelligence that directly impacts escalation dynamics in a military context, including: brittleness, 'explanability' (or 'black box' features), bias, machine-speed, and vulnerability. See, 'No. 6: The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics, and Definitional Approaches: A Primer', United Nations Institute for Disarmament Research (2017), available at: {http://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomoustechnologies-concerns-characteristics-and-definitionalapproaches-en-689.pdf}.

[6]See, for example, James D. Fearon, 'Rationalist explanations for war', *International Organization*, 49:3 (1995), pp. 379–414; Robert Powell, 'War as a commitment problem', *International Organization*, 60:1 (2006), pp. 169–203; Daniel R. Headrick, *The Tools of Empire: Technology and European Imperialism in the Nineteenth Century* (Oxford, UK: Oxford University Press, 1981); and Robert Powell, *In the Shadow of Power: States and Strategies in International Politics* (Princeton, NJ: Princeton University Press, 1999).

[7]See, for example, Lawrence Freedman, *Deterrence* (New York, NY: Polity, 2004); Lawrence Freedman, *The Evolution of Nuclear Strategy* (Basingstoke, UK: Palgrave Macmillan 2003); Thomas Schelling, *Arms and Influence* (New Haven, CT and London, UK: Yale University Press, 1966); Glen H. Snyder, *Deterrence and Defense: Towards a Theory of National Security* (Princeton, NJ: Princeton University Press, 1961); and Patrick M. Morgan, *Deterrence Now* (Cambridge, UK: Cambridge University Press, 2003); and Michael Quinlan, *Thinking About Nuclear Weapons: Principles, Problems, Prospects* (Oxford, UK: Oxford University Press 2009).

[8]See, for example, Anthony C. Lopez, Rose McDermott, and Michael Bang Petersen, 'States in mind: Evolution, coalitional psychology, and international politics', *International Security*, 36:2 (2011), pp. 48–83; Robert Jervis *How Statesmen Think: The Psychology of International Politics* (Princeton, NJ: Princeton University Press, 2017); Jacques Hymans, *The Psychology of Nuclear Proliferation: Identity, Emotions and Foreign Policy* (Cambridge, UK: Cambridge University Press,

impact of technology on the cause of war occurs through its political and psychological effects, rather than tactical or operational battlefield shifts caused by technological innovation. Specifically, the political consequences that flow from changes to the balance of power and its impact on the redistribution of resources and perceived (by beneficiaries and their rivals) strategic advantage of utilising a particular capability.[9] A new asymmetric capability, doctrine, or strategy that decreases (or increases) the perceived cost, risk, and lethality of warfare, *ceteris paribus*, should affect the observable mechanisms of conflict such as escalation, only to the extent that it changes actors' *perceptions* about how adversaries might perform in battle. Because of the centrality of information (that is, capabilities, interests, and intentions), the critical factor is to what degree a particular technology disproportionality affects states' perception of the balance of power.[10]

The remainder of this article is organised as follows. First, the article offers a conceptual overview of escalation theorising. It defines the various terms, analogies, mechanisms, and metaphors associated with escalation, which describes at its core is a fundamentally psychological and perceptual one. Second, it applies Barry Posen's inadvertent escalation analytical framework to examine the effects of AI technology on the causes of inadvertent escalation – that is, the 'security dilemma', the Clausewitzian notion of the 'fog of war', and offence military strategy and doctrine.[11] This section revisits this model to conceptualise the psychological underpinnings of the novel ways AI technology and the emerging digital information ecosystem may increase inadvertent escalation risk. Next, the article considers how state or non-state actors might use AI technology to conduct mis/disinformation (that is, information that is inaccurate or misleading) asymmetric operations in ways that might increase inadvertent risk.[12] Finally, the article highlights the critical features of inadvertent escalation risk in the emerging AI-nuclear strategic nexus, concludes with policy implications of the AI-nuclear strategic nexus, and suggests possible ways to mitigate inadvertent escalation risk and improve strategic stability.

Conceptualising inadvertent escalation: Escalation ladders, dominance, and other metaphors

The concept of escalation is at its core a fundamentally psychological and perceptual one. Like other related concepts such as deterrence and strategic stability, escalation relies upon the actor's unique understanding of context, motives, and intentions – especially in the use of capabilities.[13] How actors resolve these complex psychological variables associated with the cause, means, and

2006); and Janice G. Stein, 'Building politics into psychology: The misperception of threat', *Political Psychology*, 9:2 (1988), pp. 45–71.

[9]See William H. McNeill, *The Pursuit of Power: Technology, Armed Force, and Society since AD 1000* (Chicago, IL: University of Chicago Press, 1984); and Kier A. Lieber, *War and the Engineers: The Primacy of Politics over Technology* (Ithaca, NY: Cornell University Press, 1995).

[10]Bryan R. Early and Erik Gartzke, 'Spying from space: Reconnaissance satellites and interstate disputes', *Journal of Conflict Resolution* (March 2021), p. 4, available at: {https://doi.org/10.1177/0022002721995894}.

[11]Barry R. Posen, *Inadvertent Escalation* (Ithaca, NY: Cornell University Press, 1991).

[12]For a recent exploration of the effects of misinformation and disinformation on attributes of the digital ecosystem, see Rachel Armitage and Cristian Vaccari, 'Misinformation and disinformation', in Howard Tumber and Silvio Waisbord (eds), *The Routledge Companion to Media Disinformation and Populism* (London, UK: Routledge, 2021), ch. 3; and Cristian Vaccari and Andrew Chadwick, 'Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news', *Social Media + Society* (19 February 2020), available at: {https://journals.sagepub.com/doi/full/10.1177/2056305120903408}.

[13]See Lawrence Freedman, *Deterrence* (New York, NY: Polity, 2004); Lawrence Freedman, *The Evolution of Nuclear Strategy* (Basingstoke, UK: Palgrave Macmillan 2003); Thomas Schelling, *Arms and Influence* (New Haven, CT and London, UK: Yale University Press, 1966); Glen H. Snyder, *Deterrence and Defense: Towards a Theory of National Security* (Princeton, NJ: Princeton University Press, 1961); and Patrick M. Morgan, *Deterrence Now* (Cambridge, UK: Cambridge University Press, 2003); and Michael Quinlan, *Thinking About Nuclear Weapons: Principles, Problems, Prospects* (Oxford, UK: Oxford University Press 2009); and Colby Elbridge and Michael Gerson (eds), *Strategic Stability:*

effects of a military attack (both kinetic and non-kinetic) remains a perplexing and invariably elusive endeavour.[14] Furthermore, deterring escalation is generally achieved as a result of the fear and uncertainty (or 'fear of eruption') of how an adversary might assess capabilities, threats, and respond (or overreact) to a situation, rather than the perceived costs or military advantages of escalation, per se.[15] How might uncertainty about digital vulnerabilities affect inadvertent escalation dynamics?

Escalation theorising came into prominence during the Cold War era with the development of nuclear weapons, particularly the need to conceptualise and control conflict below the level of all-out total war. Nuclear theories of escalation continue to provide the theoretical basis for escalatory strategies and undergird debates about nuclear deterrence,[16] strategic planning, and how a conventional skirmish could become a nuclear war. *On escalation*, Herman Kahn's seminal work conceptualises a 44-rung escalation ladder metaphor, which moves from low-scale violence to localised nuclear war (or counter value warfare) to conventional and nuclear attacks against civilian populations (or strategic counter value warfare).[17] Kahn's 'escalation ladder' metaphor hinges on the idea of psychological obstacles, thresholds, or stages of the escalation process that would impose a threshold (or firebreak) to the next rung or step up the ladder in ascending order of intensity.[18] Escalation theory's emphasis on the importance of firebreaks between the rungs underscored the qualitative, psychological (both rational and irrational), and normative difference (or 'normative stigma' and taboo) between nuclear and non-nuclear domains.[19]

A seminal study by the RAND Corporation defined escalation as 'an increase in the intensity or scope of conflict that crosses a threshold(s) *considered significant by one or more of the participants*'.[20] An 'unintentional' increased intensity or scope of a situation can be inadvertent, catalytic, or accidental – encompassing incorrect or unauthorised usage (see Figure 1).[21] Intentional escalators knowingly take actions that cross thresholds (or firebreaks) for strategic gain, to send a signal of intent, obtain information about an adversary (that is, resolve, credibility commitment, or risk acceptance), or avert military defeat (that is, through pre-emption, a 'bolt-from-the-blue' attack, or grey zone tactics).[22] In contrast, inadvertent escalation occurs when an actor crosses a threshold that it considers benign, but the other side considers significant.[23] The escalator may,

Contending Interpretations (Carlisle, PA: Army War College, 2013); and Barry Buzan and Eric Herring, *The Arms Dynamic in World Politics* (London, UK: Boulder & London Lynne Reinner, 1998).

[14]See Robert Powell, *In the Shadow of Power: States and Strategies in International Politics* (Princeton, NJ: Princeton University Press, 1999).

[15]On uncertainty as a cause of war, see Geoffrey Blainey, *The Causes of War* (New York, NY: Free Press, 1988); Powell, *In the Shadow of Power: States and Strategies in International Politics*; and Rose McDermott, 'Decision-making under uncertainty', in National Research Council (ed.), *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for US Policy* (Washington DC: National Academies Press, 2010), pp. 227–42.

[16]Classical deterrence theorising in the context of nuclear weapons can be categorised as deterrence by denial, deterrence by retaliation, and deterrence by punishment. See Glenn H. Snyder, *Deterrence by Denial and Punishment* (Princeton, NJ: Center of International Studies, January 1959).

[17]Kahn subdivides these 44 rungs into seven units with seven thresholds (or firebreaks), which denote important inflection points along the escalation continuum. Kahn, *On Escalation*, p. 40.

[18]Ibid., pp. 39–40.

[19]See Nina Tannenwald, *The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use* (New York, NY: Cambridge University Press, 2007).

[20]Forrest E. Morgan, Karl P. Mueller et al., *Dangerous Thresholds: Managing Escalation in the 21st Century* (Santa Monica, CA: Rand Corporation, 2008), p. 8.

[21]Escalation, in a broader sense, can be both violent (kinetic) and non-violent (non-kinetic) in nature and occur during a military exchange ('vertical escalation') or represent an expansion in the scope and range of a conflict ('horizontal escalation'). In contrast, 'political escalation' refers to non-military changes in the scope or intensity of a situation – that is, rhetorical, an articulation of expansive objectives, or changes to the accepted rules of engagement). See Morgan et al., *Dangerous Thresholds*, ch. 2.

[22]For discussion on why a state might contemplate intentional escalation, see Kelly M. Greenhill and Peter Krause (eds), *Coercion: The Power to Hurt in International Politics* (New York, NY: Oxford University Press, 2018), pp. 11–12.

[23]Posen, *Inadvertent Escalation*.

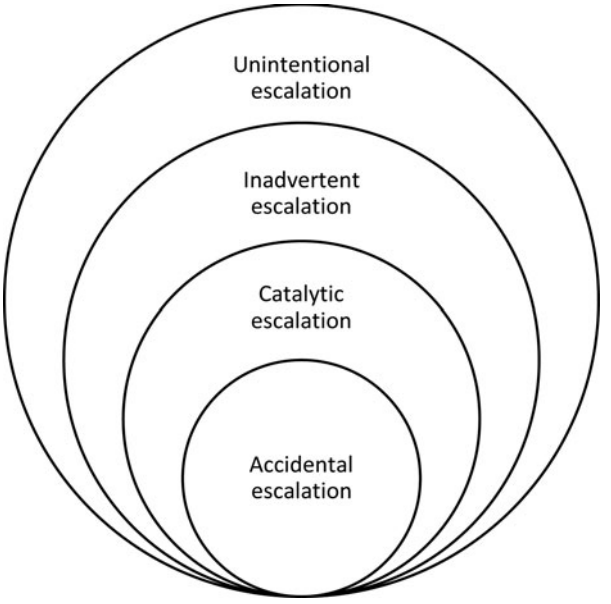

Figure 1. 'Unintentional escalation', designed by the author.

for example, send a signal to an adversary that it does not *intend* to cross a threshold but is perceived to do so by the other.[24]

These distinctions are not, however, binary, or mutually exclusive. An escalation mechanism that leads from a crisis or conflict to its outcome can involve more than one of these categories. For example, a 'false flag cyber-operations'[25] by a third-party actor targeting a state's nuclear command, control, and communication (NC3) systems accidentally sets in train escalatory mechanisms because the victim perceives the attack as a precursor to a pre-emptive strike by an adversary.[26] In this example, the actor accidentally breaches another's psychological (real or illusory) threshold,[27] triggering counter-escalation dynamics that could be in response to misinformation, fear, or misperception – also known as 'catalytic escalation'.[28] Moreover, within the broader digital information ecosystem associated with the 'Third Nuclear Age',[29] the deliberate

[24]For example, amid heightened tensions in the Baltics, NATO's actions to bolster its deterrence posture in Eastern Europe might be perceived by Moscow as preparation for a pre-emptive military offensive, thus risking inadvertent counter-escalation by Russia. Ulrich Kühn, *Preventing Escalation in the Baltics: A NATO Playbook* (New York, NY: Carnegie Endowment for International Peace, 2018).

[25]A false flag cyber-operation is designed to deflect attribution to a neutral party, and the actor behind the attack took steps to impersonate or use the distinctive infrastructure, tactics, techniques, or procedures to appear as if it had been the work of another. For example, the Olympic Destroyer cyberattack against the 2018 PyeongChang Winter Olympic Games is regarded as having been a false flag operation in which Russia's GRU designed its attack to appear as if it had been the work of North Korea. Andy Greenberg, 'The untold story of the 2018 Olympics cyberattack, the most deceptive hack in history', *Wired* (17 October 2019), available at: {https://www.wired.com/story/untold-story-2018-olympics-destroyer-cyberattack/}.

[26]Herbert Lin, 'Escalation dynamics and conflict termination in cyberspace', *Strategic Studies Quarterly*, 6:3 (2012), pp. 46–70.

[27]However, not all threats in the use of force are escalatory. Escalation occurs only when at least one actor views this action (that is, rhetoric or signalling) as shifting the scope or intensity of a situation. See Michael Brecher, 'Crisise escalation: Model and findings', *International Political Science*, 17:2 (1996), pp. 215–30.

[28]James Johnson, '"Catalytic nuclear war" in the age of artificial intelligence & autonomy: Emerging military technology and escalation risk between nuclear-armed states', *Journal of Strategic Studies* (2021), available at: {https://doi:10.1080/01402390.2020.1867541}.

[29]The concept of 'nuclear ages' is a contested one. Broadly speaking, the first nuclear age refers to the years between 1945 and the end of the Cold War, while a second between 1989–91 to the present. On the 'Third Nuclear Age', see Rebecca

use of nuclear weapons that originates from a false, manipulated, or distorted assessment of a situation (for example, in response to an early warning system false alarm), can quickly muddy intentionality lines.[30]

In short, the binary distinction between deliberate and inadvertent use of nuclear weapons is inherently problematic.[31] Escalation can, therefore, be a strategic bargaining tool (that is, for deterrence and coercion) and a risk to be controlled and potentially mitigated.[32] Thus, actions that are interpreted as escalatory by almost all actors (for example, the use of nuclear weapons to respond to a low-level conventional conflict) while others are considerably less clear-cut – for instance, a cyber espionage operation against a states' dual-use command and control systems.[33] Consequently, escalation situations typically involve 'competition in risk-taking' and resolve.[34] Either side can intensity a situation providing the other side does not match that escalation in kind – if this escalation was not matched and victory achieved, the 'cost of the increased effort would be below in relation to the benefits of victory'.[35]

As Bernard Brodie noted in 1965, 'since the beginning of the nuclear era there has been *in the minds of men* a strong tendency to distinguish between nuclear and non-nuclear weapons combined with a *fear of and aversion to the former*', and this distinction has tended 'to *grow stronger with time* rather than weaken'.[36] To be sure, recent scholarship questioning the validity of the public's taboo on the use of nuclear weapons (especially as tools to fight terrorism and nuclear proliferation), and policy debates about the use of tactical nuclear weapons – which surrounded the Trump administration's Nuclear Posture Review – demonstrates the prophetic nature of Brodie's 'firebreak theory'.[37] Besides, the 'fear' and 'aversion' associated with nuclear weapons have been used in studies to explain the role of emotion and cognition play in the misperceptions about weapons' effectiveness. Misperceptions can create an 'emotional springboard' for the process of reasoning that can influence nuclear non-use, deterrence, and nuclear taboos, which in turn informs policymaking.[38]

The escalation ladder metaphor, like any theoretical framework, has limitations. Actors would not necessarily, however, move sequentially and inexorably from the lower rungs ('subcrisis manoeuvring') to the higher rungs ('spasm or insensate war') – that is, rungs can be skipped and go up as well as down. Instead, there are many pathways and mechanisms – for instance, conventional vs nuclear, tactical vs strategic, or counterforce vs counter value – between low-intensity

Hersman, 'Wormhole escalation: The new nuclear age', *Texas National Security Review*, 2:3 (2020), pp. 91–109; Nicholas Miller and Vipin Narang, 'Is a new nuclear age upon us? Why we may look back on 2019 as the point of no return', *Foreign Affairs* (December 2019); and Andrew Futter and Benjamin Zala, 'Strategic non-nuclear weapons and the onset of a Third Nuclear Age', *European Journal of International Security* (2021), pp. 1–21.

[30]Kahn, *On Escalation: Metaphors and Scenarios*, p. 285.

[31]Sico van der Meer, 'Reducing Nuclear Weapons Risks: A Menu of 11 Policy Options', Policy Brief, Clingendael Netherlands Institute of International Relations (June 2018).

[32]Deterrence and coercion require the actor that is deterring to share information about the military balance with the actor that is being deterred. Alternatively, an actor may employ nuclear ambiguity to achieve the same deterrence effect – for example, the clandestine development of Russia's *Perimitr* (or 'Dead Hand' system). See Early and Gartzke, 'Spying from space', p. 5; and David E. Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (New York, NY: Anchor Books, 2009).

[33]Technological advances in AI technology and cyber capabilities, coupled with the increasingly commingled nature of the state's nuclear and conventional command and control systems, have enabled solutions to overcome the robustness of permissive action links and increase these vulnerabilities systems. Bruce Blair, *The Logic of Accidental Nuclear War* (Washington, DC: Brookings Institute, 1993).

[34]Kahn, *On Escalation*, p. 289.

[35]Ibid.

[36]Bernard Brodie, *Escalation and the Nuclear Option* (Santa Monica, CA: The Rand Corporation, 1965), p. 64.

[37]See, Darryl Press, Scott Sagan, and Benjamin Valentino, 'Atomic aversion: Experimental evidence on taboos, traditions, and the non-use of nuclear weapons', *American Political Science Review*, 107:1 (2013), pp. 188–206.

[38]Frank Sauer, *Atomic Anxiety: Deterrence, Taboo and the Non-Use of US Nuclear Weapons* (New York, NY: Springer, 2015), p. 176.

and all-out nuclear confrontation. Besides, states can be at different rungs or thresholds along the 'relatively continuous' pathways to war.[39] Despite its limitations, however, Kahn's 'escalation ladder' is a useful metaphorical framework to reflect on the possible available options (for example, a show of force, reciprocal reprisals, costly signalling, and pre-emptive attacks), progression of escalation intensity, and scenarios in a competitive nuclear-armed dyad. This article argues that the introduction of AI is creating new ways (or 'rungs') and potential shortcuts up (and down) the ladder, which might create new mechanisms for a state to perceive (or misperceive) others to be on a different rung, thus making some 'rungs' more (or less) fluid or malleable.

Strategic rivals require a balanced understanding of how the other views the escalation hierarchy. During a crisis or conflict, continuous feedback about an adversary's intentions, where it views itself on the escalation ladder, and how shifts in the scope or intensity of a situation (that is, kinetic, non-kinetic, or rhetorical) may be perceived.[40] Because of the inherently subjective nature of escalation, actions perceived as escalatory by one state can be misunderstood as thus by others.[41] What characteristics of AI technology may create new rungs on the escalation ladder that increase the inadvertent risk of escalating a conventional conflict to nuclear war?

An important feature of escalation theory is the notion of 'escalation dominance': when a nuclear power force an adversary to back down because the cost and risk associated with further escalation outweigh the perceived benefits.[42] The state that possesses a position of escalation dominance, *ceteris paribus*, has unique tactical advantages on a particular rung of the escalation ladder. Moreover, a state that has most to lose in a situation escalating, or fears escalation the least (that is, it is not the one doing the escalation), will axiomatically achieve an element of escalation dominance in a situation.[43] Escalation dominance is thus a function of several variables: (1) where two states are positioned on the escalation ladder; (2) an assessment of their respective capabilities (both offensive and defensive) on a particular rung; (3) each sides' perception of the probability and likely outcome of moving to a different rung; and (4) the perceived ability of one side of the other to effectuate this shift.[44] Furthermore, both sides' fear and aversion of intensifying a situation is also an important psychological variable of escalation dominance.

A new model of inadvertent escalation in the digital age

In his seminal work on inadvertent escalation, Barry Posen identifies the major causes of inadvertent escalation as the 'security dilemma', the Clausewitzian notion of the 'fog of war', and offensively orientated military strategy and doctrine.[45] This section applies Posen's framework to examine the effects of AI technology on the causes of inadvertent escalation. In the light of recent technological change, this section revisits Posen's analytical framework to examine the psychological underpinnings of security dilemma theorising (misperception, cognitive bias, and heuristics) to consider the how and why the novel characteristics of AI technology and the emerging digital information ecosystem may destabilise 'crisis stability' and increase inadvertent escalation risk.[46]

[39]Ibid., p. 38.

[40]Brecher, 'Crisis escalation', pp. 215–30.

[41]Greenhill and Krause (eds), *Coercion*, pp. 3–33.

[42]On the core logic of escalation dominance, see Kahn, *On Escalation*; Kahn, *Thinking About the Unthinkable in the 1980s* (New York, NY: Touchstone, 1985); and Matthew Kroenig, 'Nuclear superiority and the balance of resolve: Explaining nuclear crisis outcomes', *International Organization*, 67:1 (2013), pp. 141–71.

[43]Ibid., p. 290.

[44]Ibid.

[45]Posen, *Inadvertent Escalation*, pp. 12–27.

[46]'Crisis stability' refers to the presumption that control can be maintained during crisis or conflict to ensure that nuclear weapons are not used – that is, a situation where neither side has an incentive to strike first or pre-emptively. Cold War-era debates on the concept centred on how specific capabilities, postures, and military doctrine could escalate (either intentionally or inadvertently) to nuclear crisis or war. See Thomas Schelling, *Arms and Influence* (New Haven, CT and London, UK:

The security dilemma concept

According to 'security dilemma'[47] theorising, the relative ease of carrying out and defending against military attacks (the 'offence-defence balance'), and the ambiguity about whether a weapon is offensive or defensive (the 'offence-defence distinguishability'), can cause crisis instability and deterrence failure because these characteristics create fear and uncertainty about an adversary's intentions. That is, where they harbour benign (that is, non-aggressive/defensive) or malign (that is, aggressive/offensive) intent towards the other side.[48] In his seminal paper on the topic, Robert Jervis defines the security dilemma as the 'unintended and undesired consequences of actions *meant to be defensive* … many of how a state tries to increase its security decrease the security of others' – that is, one state's gain in security can inadvertently undermine the security of others.[49] Actors tend to perceive the accumulation of offensive capabilities (and attendant offensive strategies) by others as threatening, assuming the worst respond with countermeasures – for example, arms racing, crisis mobilisation, raising the alert status, and pre-emptive war.[50]

According to Jervis, it is 'the fear of being exploited' that 'most strongly drives the security dilemma'.[51] As we have noted, the fear (both rational and irrational) of conventional skirmishes crossing the nuclear Rubicon can be found in nuclear policymakers' statements as far back as the 1950s.[52] Therefore, security dilemma logic can be used to consider both peacetime spirals of political mistrust and shifts in the military balance, crisis stability dynamics, and escalation mechanisms once military conflict begins. The security dilemma concept is an inherently inadvertent phenomenon; that is, weapons indistinguishability (that is, offence vs defence) allows states to inadvertently (for rationale political or military reasons) threaten others, which can spark spirals of mutual distrust, strategic competition, and military action.[53]

There are several ways in which the security dilemma can act as a catalyst for inadvertent escalation that is likely to be compounded in the digital age.[54] First, while nuclear-armed states have a shared interest in avoiding nuclear war, they also place a high value on their nuclear forces – both vital security assets and as symbols of national prestige and status.[55] As a corollary, conventional weapons – devised by a nuclear power for defensive purposes or counterforce missions – may nonetheless be viewed by an adversary as an offensive threat to their nuclear survivability (discussed below).[56] Second, escalatory rhetoric and other aggressive responses by a threatened

Yale University Press, 1966); Richard Lebow, 'Clausewitz and nuclear crisis stability', *Political Science Quarterly*, 103:1 (1988), pp. 81–110; and Forrest Morgan, *Crisis Stability and Long-Range Strike: A Comparative Analysis of Fighters, Bombers, and Missiles* (Santa Monica, CA: Rand Corporation, 2013).

[47]For the seminal works on the 'security dilemma' see Herbert Butterfield, *History and Human Relations* (London, UK: Collins, 1951); John Herz, *Political Realism and Political Idealism: A Study in Theories and Realities* (Chicago, IL: University of Chicago Press, 1951); and Robert Jervis, 'Cooperation under the security dilemma', *World Politics*, 30:2 (1978), pp. 169–214.

[48]For seminal work on the concept, see Lynn-Jones, 'Offense-defense theory and its critics', pp. 660–91; Glaser and Kaufmann, 'What is the offense-defense balance and how can we measure it?', pp. 44–82; Van Evera, 'The cult of the offensive and the origins of the First World War', pp. 58–107; and Jervis, 'Cooperation under the security dilemma'. For studies on the 'cyber offense-defense balance', see Gartzke and Lindsay, 'Weaving tangled webs'; Ilai Saltzman, 'Cyber posturing and the offense-defense balance', *Contemporary Security Policy*, 34:1 (2013), pp. 40–63; Rebecca Slayton, 'What is the cyber offense-defense balance? Conceptions, causes, and assessment', *International Security*, 41:3 (2017), pp. 72–109; and Ben Garfinkel and Allen Dafoe, 'How does the offense-defense balance scale?', *Journal of Strategic Studies*, 42:6 (2019), pp. 736–63.

[49]Jervis, 'Cooperation under the security dilemma', pp. 169–70.

[50]Schelling, *Arms and Influence*.

[51]Jervis, 'Cooperation under the security dilemma', p. 172.

[52]Brodie, *Escalation and the Nuclear Option*.

[53]Jervis, 'Cooperation under the security dilemma', pp. 170, 193.

[54]Posen, *Inadvertent Escalation*, pp. 12–15.

[55]See Scott Sagan, 'Why do states build nuclear weapons? Three models in search of a bomb', *International Security*, 21 (1996), pp. 54–86; and Tannenwald, *The Nuclear Taboo*.

[56]See Futter and Zala, 'Strategic non-nuclear weapons and the onset of a Third Nuclear Age', pp. 1–21.

state (especially in situations of military asymmetry) may be misperceived as unprovoked malign intent and not as a response to the initiator's behaviour prompting action and reaction spirals of escalation. Third, and related, the state of heightened tension and the compressed decision-making pressures during a conventional conflict would radically increase the speed by which action and reaction spirals of escalation unravel. In the digital age, these dynamics would be further compounded, which would reduce the options and time for de-escalation and increase the risks of horizontal (the scope of war) and vertical (the intensity of war) inadvertent escalation.[57]

Security dilemma theorising also ties in with the concept of the 'capability/vulnerability paradox' in International Relations.[58] That is, one state's pursuit of a new resource to compete against other states introduces a vulnerability that is perceived as threatening.[59] This paradox suggests that when a state's military capabilities are dependent on a particular resource (that is, that may be exploited or dominated by an adversary), both sides have incentives for pre-emptive first strikes – the resource vulnerability and not the new capability, per se generate these destabilising outcomes.[60] Much like the security dilemma, therefore, the potential impact of the 'capability/vulnerability paradox' has increased in the context of technological advancements associated with the 'information revolution' in military affairs (that is, the centralisation of information and dependencies on digital information to conduct modern warfare), which now includes artificial intelligence.[61] As Pentagon insider Richard Danzig notes, 'digital technologies … are a security paradox: even as they grant unprecedented powers, they also make users less secure'.[62] In this sense, as a cause of inadvertent escalation, the security dilemma connects to the broader strategic and security literature on misperception, emotions, and cognition in various studies evince a strong qualitative connection between human psychology and war.[63] How might the confusion and uncertainty of war increase inadvertent risk?

[57]It is also plausible that improvements in battlefield awareness and decision-making AI-enabled tool afford commanders more time to make decision during combat. See Paul Scharre, 'Autonomous Weapons and Stability' (PhD Thesis, King's College London, 2020), available at: {https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.806777}; and James Johnson, 'Artificial Intelligence in nuclear warfare: A perfect storm of instability?', *The Washington Quarterly*, 43:2 (2020), pp. 197–211.

[58]Jacquelyn Schneider, "The capability/vulnerability paradox and military revolutions: Implications for computing, cyber, and the onset of war', *Journal of Strategic Studies*, 42:6 (2019), pp. 841–63.

[59]The 'capability/vulnerability paradox' in International Relations may also shift when the resource being pursued or threatened is less 'material' or its strategic value changes (for example, fossil fuels).

[60]As a helpful counterpoint, while the proliferation and dependency on the digital information ecosystem increase, the ubiquity of information and the intrinsic inability to control all information might create disincentives for first strike. Schneider, 'The capability/vulnerability paradox and military revolutions', p. 842.

[61]See, for example, Eliot A. Cohen, 'A revolution in warfare', *Foreign Affairs* (1996), pp. 37–54; Eliot A. Cohen, 'Change and transformation in military affairs', *Journal of Strategic Studies*, 27:3 (2004), pp. 395–407; Erik Dahl, 'Network centric warfare and operational art', *Defence Studies*, 2:1 (spring 2002), pp. 17–34; Arthur Cebrowski and John Garstka, 'Network-centric warfare: Its origin and future', *US Naval Institute Proceedings*, 124:1 (1998), pp. 28–35; and Michael Raska, 'The sixth RMA wave: Disruption in military affairs?', *Journal of Strategic Studies* (2020), available at: {https://doi:10.1080/01402390.2020.1848818}.

[62]Richard Danzig, 'Surviving on a diet of poisoned fruit: Reducing the National Security Risks of America's Cyber Dependencies', Center for a New American Security (21 July 2014), available at: {https://www.cnas.org/publications/reports/surviving-on-a-diet-of-poisonedfruit-reducing-the-national-security-risks-of-americas-cyber-dependencies}.

[63]See Robert Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976); Jacques Hymans, *The Psychology of Nuclear Proliferation: Identity, Emotions and Foreign Policy* (Cambridge, UK: Cambridge University Press, 2006); Janice G. Stein, 'Building politics into psychology: The misperception of threat', *Political Psychology*, 9:2 (1988), pp. 45–71; Barbara Farnham, *Avoiding Losses/Taking Risks: Prospect Theory and International Conflict* (Ann Arbor, MI: University of Michigan Press, 1994); Rose McDermott, James Fowler, and Oleg Smirnov, 'On the evolutionary origin of prospect theory preferences', *Journal of Politics*, 70:2 (2008), pp. 335–50; Philip Tetlock and Richard Boettger, 'Accountability: A social magnifier of the dilution effect', *Journal of Personality & Social Psychology*, 57:3 (1989), pp. 388–98; and Jonathan Mercer, 'Human nature and the first image: Emotion in international politics', *Journal of International Relations & Development*, 9 (2006), pp. 288–303.

Clausewitzian 'fog of war'

Inadvertent escalation risk can also be caused by the confusion and uncertainty (or 'fog of war') associated with gathering, analysing, and disseminating relevant information about a crisis or conflict – which has important implications for the management, control, and termination of war.[64] Traditional nuclear deterrence, crisis stability, and strategic stability theorising work off the questionable premise that actors are rational and presume that they are therefore 'able to understand their environment and coordinate policy instruments in ways that cut through the "fog of war"'.[65] In reality, misperception, miscalculation, accidents, and failures of communication 'events can readily escape control', and although these escalation dynamics are often unintended, potentially catastrophic events can result.[66] With the advent of the speed, uncertainty, complexity, and cognitive strains (see below) associated with conducting military operations on the digitised battlefield, the prospects that decision-making by 'dead reckoning' – dependent on the experience and sound judgement of defence-planners when information is scarce and autonomous bias proclivities loom large – can prevent command and control failures caused by the 'fog' seems fanciful.[67]

The confusion and uncertainty associated with the 'fog of war' can increase inadvertent risk in three ways: (1) it can complicate the task of managing and controlling military campaigns at a tactical level (or situational battlefield awareness); (2) it can further compound the problem of offence-defence distinguishability; and (3) it can increase the fear of a surprise or pre-emptive attack (or a 'bolt from the blue').[68] Taken together, these mechanisms can result in unintentional (and possibly irrevocable) outcomes and thus obfuscate the meaning and the intended goals of an adversary's military actions.[69] In short, the dense 'fog of war' increases the risk of inadvertent escalation because 'misperceptions, misunderstandings, poor communications, and unauthorized or unrestrained offensive operations could reduce the ability of civilian authorities to influence the course of war'.[70]

While Cold War decision-makers also faced extreme domestic and foreign pressures to decipher the adversary's intentions correctly, these pressures are being amplified in the current digitised information ecosystem – reducing the timeframe for decision-makers during a crisis, increasing the cognitive and heuristic burdens caused by information overload and complexity.[71]

[64]The strategic value of 'information' has long been recognised in military affairs: (1) as a force multiplier to alter the balance of power during war and peace; and (2) providing knowledge about an adversary's capabilities that can make political leaders more effective at recognising acceptable bargains and channels for cooperation, thus ensuring peace. Early and Gartzke, 'Spying from space', p. 5.

[65]See Bill Owens, *Lifting the Fog of War* (New York, NY: Farrar, Straus and Giroux 2000); and Jervis, Lebow, and Stein, *Psychology and Deterrence*.

[66]Robert Jervis, 'Arms control, stability, and causes of war', *Political Science Quarterly*, 108:2 (1988), pp. 81–110.

[67]The term 'dead reckoning' refers to analytical predictions, intuitions, predictions, or judgements of captains or pilots to navigate a ship derived from the environment's internal instruments and knowledge. Kahn, *On Escalation: Metaphors and Scenarios*, pp. 211–12. On 'autonomous bias', see Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick, 'Does automation bias decision-making?', *International Journal of Human Computer Studies*, 51:5 (1999), pp. 991–1006; and Mary L. Cummings, 'Automation bias in intelligent time-critical decision support systems', *AIAA 1st Intelligent Systems Technical Conference* (2004), pp. 557–62.

[68]Ibid., p. 20.

[69]The historical record is replete with examples of how misunderstandings or incomplete information about ongoing military operations can contribute to escalation – for example, China's entry into the Korean War; Germany's Blitz campaign on the British Isles in 1940; and the Cuban crisis. See Rosemary Foot, *The Wrong War: American Policy and the Dimensions of the Korean Conflict, 1950–1953* (Ithaca, NY: Cornell University Press, 1985); F. M. Sallagar, *The Road to Total War: Escalation in World War Two* (Santa Monica, CA: Rand Corporation, 1969); and Richard K. Betts, *Soldiers, Statesmen, & Cold War Crisis* (Cambridge, MA: Harvard University Press, 1977).

[70]Posen, *Inadvertent Escalation*, p. 22.

[71]Rebecca Hersman et al., 'Under the Nuclear Shadow: Situational Awareness Technology and Crisis Decision-Making', Center for Strategic and International Studies (18 March 2020), available at: {https://ontheradar.csis.org/analysis/final-report/}.

Cognitive pressures caused by the sheer volume of information flow (both classified and open sources) are being compounded by the degraded reliably and politicisation (or 'weaponisation') of information. These pressures can, in turn, make decision-makers more susceptible to cognitive bias, misperceptions, and heuristics (or mental shortcuts) to approach complex problem-solving – either directly or indirectly informing decision-making.[72]

While disinformation and psychological operations in deception and subversion are not a new phenomenon,[73] the introduction of new AI-enhanced tools in the emerging digital information ecosystem enables a broader range of actors (state and non-state) with asymmetric techniques to manipulate, confuse, and deceive.[74] Disinformation operations might erode the credibility of, and undermine public confidence in, a states' retaliatory capabilities by targeting specific systems (that is, command and control) or personnel (primarily via social media) who perform critical functions in maintaining these capabilities.[75] For example, cyberweapons – notably 'left of launch' operations – have been allegedly used by the United States to undermine Iranian and North Korean confidence in their nuclear forces and technological preparedness.[76]

The potential utility of social media to amplify the effects of a disinformation campaign was demonstrated during the Ukrainian Crisis in 2016 when military personnel's cell phones were the victims of Russian information operations.[77] However, the use of these novel techniques during a nuclear crisis and how they might impact the 'fog of war' is less understood and empirically untested.[78] During a nuclear crisis, a state might attempt, for instance, to influence and shape the domestic debate of an adversary (for example, shift preferences, exacerbate domestic-political polarisation, or coerce/co-opt groups and individuals on social media) in order to improve its bargaining hand by delegitimising (or legitimising) the use of nuclear weapons during an escalating situation, or bring pressure to bear on an adversary's leadership to sue for peace or de-escalate a situation – a tactic which may, of course, dangerously backfire.[79]

Moreover, a third-party (state or non-state) actor to achieve its nefarious goals could employ active information techniques (for example, spreading false information of a nuclear detonation, troop movement, or missiles leaving their garrison) during a crisis between nuclear rivals (for

[72]See Daniel Kahneman, *Thinking, Fast and Slow* (New York, NY: Farrar, Straus and Giroux, 2011); Martin Kaplan, Tatiana Wanshula, and Mark Zanna, 'Time pressure and information integration in social judgment', in Ola Svenson and John Maule (eds), *Time Pressure and Stress in Human Judgment and Decision Making* (Boston, MA: Springer, 1993), pp. 255–67; and Carsten De Dreu, 'Time pressure and closing of the mind in negotiation', *Organizational Behavior and Human Decision Processes*, 91:2 (2003), pp. 280–95.

[73]See Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York, NY: Farrar, Straus and Giroux, 2020).

[74]For example, see David Sanger, *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age* (New York, NY: Broadway Books, 2019); and Audrey K. Cronin, *Power to the People: How Open Technological Innovation is Arming Tomorrow's Terrorists* (New York, NY: Oxford University Press, 2019).

[75]Kristiina Müür, Holger Mölder, and Vladimir Sazonov, 'A comparative overview of online news', in Vladimir Sazonov et al. (eds), *Russian Information Warfare against the Ukrainian State and Defence Forces* (NATO Strategic Communications, 2016), pp. 70–99.

[76]See Riki Ellison, 'Left of Launch', Missile Defense Advocacy Alliance (2015), available at: {https://missiledefenseadvocacy.org/alert/3132/}.

[77]Marie Baezner and Patrice Robin, *Cyber and Information Warfare in the Ukrainian Conflict* (Zurich, Switzerland: Center for Security Studies, ETH Zurich, 2018).

[78]Harold Trinkunas, Herbert Lin, and Benjamin Loehrke, *Three Tweets to Midnight: Effects of the Global Information Ecosystem on the Risk of Nuclear Conflict* (Stanford, CA: Hoover Institution Press, 2020), pp. 150–2; and Johnson, 'Catalytic nuclear war in the age of artificial intelligence & autonomy'.

[79]Russia, for example, reportedly conducted an active information campaign against several NATO states to influence the public discourse and policymakers to destabilise NATO's missile defence operations in Europe. See Hege Eilersten, 'Russia's Ambassador warns: Missile shield will endanger Norway's borders', *High North News* (22 February 2017), available at: {https://www.highnorthnews.com/en/russias-ambassador-warns-missile-shield-will-endanger-norways-borders}.

example, India-Pakistan, US-China, Russia-NATO) to incite crisis instability.[80] Public pressures on decision-makers from these crisis dynamics, and evolving at a pace that may outpace events on the ground, might impel (especially thinned-skinned and inexperienced) leaders, operating under the shadow of the deluge of 24-hour social media feedback and public scrutiny, to take actions that they might not otherwise have.[81]

In sum, the emerging AI-enhanced digitised information environment, though not fundamentally altering the premise upon which actors (de-)escalate a situation, imperfect information, uncertainty, and risk associated with the 'fog of war', nonetheless, introduces novel tools of misinformation and disinformation and an abundance of information radically alters the cognitive pressures placed on decision-makers during crisis and conflict. Besides, decision-makers subjection to an abundance of disparate and often unverified information will indubitably influence actors' (collectively or individuals) policy preferences and perceptions. As a result, complicating the security dilemma challenge of understanding an adversary's capabilities, intentions, doctrine, and strategic thinking, with potentially profound repercussions for escalation dynamics. How might these pressures affect states' regional conflicts with different military doctrine, objectives, and attitudes to risk?

Offensive military doctrine and strategy

Because of a lack of understanding between (nuclear-armed and non-nuclear-armed) adversaries about where the new tactical possibilities offered by these capabilities figure on the Cold War-era escalation ladder, thus the pursuit of new 'strategic non-nuclear weapons' (for example, cyber weapons, drones, missile defence, precision munitions, counterspace weapons) increases the risk of misperception.[82] The fusion of AI technology into conventional weapon systems (that is, to enhance autonomous weapons, remote sensing for reconnaissance, improving missile guidance and situational awareness) is creating new possibilities for a range of destabilising counterforce options targeting states' nuclear-weapon delivery and support systems; for example, cyber NC3 'kill switch' attacks or tracking adversaries' nuclear-armed submarines and mobile missile launchers.[83]

Russia, the United States, and China are currently pursuing a range of dual-capable (conventional and nuclear-capable) delivery systems (for example, hypersonic guide vehicles, stealth bombers, and a range of precision munitions) and advanced conventional weapons (drones, space-based, and cyber weapons.[84]) that are capable of achieving strategic effects – that is, without the need to use nuclear weapons.[85] These lines are blurred further by the use of dual-use

[80]See Robert Ayson, 'After a terrorist nuclear attack: Envisaging catalytic effects', *Studies in Conflict & Terrorism*, 33:7 (2010), pp. 571–93; and Peter Hayes, 'Non-State Terrorism and Inadvertent Nuclear War', Nautilus Institute for Security and Sustainability Special Reports (18 January 2018).

[81]Trinkunas, Lin, and Loehrke, *Three Tweets to Midnight*, p. 152.

[82]'Strategic non-nuclear weapons' are also referred to as strategic conventional weapons or advanced conventional weapons. See James Acton, 'Russia and strategic conventional weapons: Concerns and response', *Nonproliferation Review*, 22:2 (2015), pp. 141–54; and Futter and Zala, 'Strategic non-nuclear weapons and the onset of a third nuclear age', pp. 1–21.

[83]See Andrew Futter, *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Washington, DC: Georgetown University Press, 2018), pp. 117–25.

[84]See US A*nnual Report to Congress: Military and Security Developments Involving the People's Republic of China 2020* (Washington, DC: Office of the Secretary of Defense, 2020); and US National Air and Space Intelligence Center (NASIC), *Ballistic and Cruise Missile Threat* (Washington, DC: Office of the Secretary of Defense, 2020).

[85]For example, Russia deploys dual-use ground-launched cruise missiles, India and Pakistan possess dual-payload ground-launched missiles. China's DF-26 intermediate-range ballistic missile is dual-capable, and the US, China, Russia, India, and Pakistan all have nuclear-capable aircraft capable of supporting conventional systems. See James M. Acton, 'Is it a Nuke? Pre-Launch Ambiguity and Inadvertent Escalation', Carnegie Endowment for International Peace (9 April 2020), available at: {https://carnegieendowment.org/2020/04/09/is-it-nuke-pre-launch-ambiguity-and-inadvertent-escalation-pub-81446}; Hans M. Kristensen, Robert S. Norris, and Julia Diamond, 'Pakistani nuclear forces, 2018', *Bulletin of the Atomic Scientists*, 74:5 (2018), p. 355; and Hans M. Kristensen and Matt Korda, 'Indian nuclear Forces, 2018', *Bulletin of the Atomic Scientists*, 74:6 (2018), p. 363, available at: {https://doi.org/10.1080/00963402.2018.1533162}.

command and control systems (that is, early warning, situational awareness, and surveillance) to manage conventional and nuclear missions.[86] Chinese analysts, for example, while concerned about the vulnerabilities of their command and control systems to cyberattacks, are optimistic about the deployment of AI augmentation (for example, ISR, intelligent munitions, and unmanned aerial vehicles) to track and target an adversary's forces, and will lower cost of (economic and political) signalling and deploying military forces.[87]

Advances in AI technology, in conjunction with the technologies it can enable (for example, remote sensing, hypersonic technology, and robotics and autonomy), increase the speed, precision, lethality, and survivability of strategic non-nuclear weapons, thus exacerbating the destabilising effects of these capabilities used by nuclear-armed rivals. Thus, opening new pathways for both horizontal and vertical inadvertent escalation.[88] Moreover, these technological advances have been accompanied by destabilising doctrinal shifts by certain regional nuclear powers (Russia, North Korea, and possibly China), which indicates a countenance of the limited use of nuclear weapons to deter an attack (or 'escalate to de-escalate'), in situations where they face a superior conventional adversary and the risk of large-scale conventional aggression.[89]

Furthermore, volatility in nuclear-armed states' civil-military relations can create internal pressures to pursue a more aggressive nuclear force posture or doctrine.[90] The assumption that new (and latent) regional nuclear-states will act in ways that reflect their interests as rational actors to avoid nuclear war, thus enhance deterrence (or 'rational deterrence theory'),[91] and crisis stability, understates the role of influential military organisations (that is, offensive-doctrinal bias, parochialism, and behavioural rigidities) in shaping nuclear doctrine that can lead to deterrence failure and nuclear escalation – despite national security interests to the contrary.[92] For instance, in a nuclear-armed state where the military officers influence the nuclear strategy, the adoption of offensive doctrines may emerge, which reflect volatile civil-military relations rather than strategic realities.[93]

In short, technological advancements to support states' nuclear deterrence capabilities will develop in existing military organisations imbued with their norms, cultures, structures, and invariably, mutually exclusive strategic interests. China's 'military-civil fusion' concept

[86]See James M. Acton, Li Bin, Alexey Arbatov, Petr Topychkanov, and Zhao Tong, *Entanglement: Russian and Chinese Perspectives on Non-Nuclear Weapons and Nuclear Risks*, ed. James M. Acton (Washington, DC: Carnegie Endowment for International Peace, 2017); and Hersman et al., 'Under the Nuclear Shadow'.

[87]See John Schaus and Kaitlyn Johnson, 'Unmanned Aerial Systems' Influences on Conflict Escalation Dynamics', Center for Strategic and International Studies (7 August 2018), available at: {https://aerospace.csis.org/unmanned-aerial-systems-influences-on-conflict-escalationdynamics/}.

[88]See Keir A. Lieber and Darryl G. Press, 'The new era of counterforce: Technological change and the future of nuclear deterrence', *International Security*, 41:4 (spring 2017), pp. 9–49; and Paul Bracken, 'The cyber threat to nuclear stability', *Orbis*, 60:2 (2016), pp. 188–203; Dean Wilkening, 'Hypersonic weapons and strategic stability', *Survival*, 61:5 (2019), pp. 129–48; and Michael Horowitz, Paul Scharre, and Alexander Velez-Green, 'A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence', arXiv (December 2019), available at: {https://arxiv.org/abs/1912.05291}.

[89]For example, Russia's military doctrine explicitly states it would consider nuclear weapons to respond to a large-scale conventional attack. While China maintains an official no-first-use policy, defence analysts continue to debate the veracity of this pledge, particularly in the event of conventional military aggression against a superior adversary in the Taiwan Straits or target China's dual-use military platforms. On Russia, see Alexey Arbatov, Petr Topychkanov, and Vladimir Dvorkin, *Entanglement: A New Security Threat: A Russian Perspective* (Washington, DC: Carnegie Endowment for International Peace, 2017), pp. 25–6. On China, see Caitlin Talmadge, 'Would China go nuclear? Assessing the risk of Chinese nuclear escalation in a conventional war with the United States', *International Security*, 41:4 (2017), pp. 50–92.

[90]See Scott Sagan, 'The perils of proliferation: Organization theory, deterrence theory, and the spread of nuclear weapons', *International Security*, 18:4 (1994), pp. 66–107.

[91]Several defensive-realist International Relations scholars have argued that self-interested rational unitary actors will likely behave in ways that reduce the dangers associated with the proliferation of nuclear weapons. Kenneth N. Waltz, 'Nuclear myths and political realities', *American Political Science Review*, 84:3 (1990), pp. 731–45.

[92]Sagan, 'The perils of proliferation', p. 102.

[93]States that have volatile civil-military relations are also more vulnerable to accidents involving nuclear operations. Ibid., pp. 98–9.

(a dual-use integration strategy) designed to co-opt or coerce Chinese commercial entities to support the technological development of the People's Liberation Army (PLA) to meet the needs of 'intelligentized warfare in the future' – is an important test case in a civilian-led initiative designed to drive military innovation in the pursuit of broader geostrategic objectives.[94] The impact of China's 'military-civil fusion' on civil-military relations remains unclear, however.[95]

US counterforce capabilities – to disarm an enemy without resort to nuclear weapons – used in conjunction with air and missile defences to mop up any residual capabilities after an initial attack will generate crisis instability and 'use it or lose it' pressures.[96] Chinese analysts, for example, have expressed concern that US advances in AI could overwhelm Chinese air defences, thus reducing the time available to commanders to respond to an imminent attack – for example, from the US autonomous AI-enabled Long Range Anti-Ship Missile (AGM-158C) designed to target 'high-priority targets'.[97] Furthermore, the increased optimism in states' ability to use AI-enhancements to find, track, and destroy others' nuclear forces enabled by AI technology (notably when military-capability imbalances exist) will be an inherently destabilising phenomenon.[98] What one side views as conventional operations might be viewed by the other side as a precursor to a disabling counterforce attack (for example, targeting dual-use command and control centres and air defences), thus increasing inadvertent escalation risk.[99]

Because of the asymmetry of interest at stake in a regional crisis involving the United States, the stakes will likely favour the defending nuclear-armed power.[100] According to 'prospect theory', a regional power would perceive the relative significance of a potential loss more highly than again.[101] That is, when prospect theory is applied to deterrence dynamics, leaders are inclined to take more risks (that is, risk-acceptant) to protect their positions, status, and reputations, than they are to enhance their position.[102] Thus, having suffered a loss, leaders are generally predisposed to engage in excessive risk-taking behaviour to recover lost territory – or other

[94]Suppose the Chinese government wants a technology that a particular commercial entity controls; extra-legal influence or coercion can compel the company to turn it over. In that case, forced technology transfers are not thought to occur routinely or considered effective. See Elsa B. Kania and Lorand Laskai, *Myths and Realities of China's Military-Civil Fusion Strategy* (Washington, DC: Center for a New American Security, 2021).

[95]For recent research on civil-military cooperation and technological innovation, see Maaike Verbruggen, 'The role of civilian innovation in the development of lethal autonomous weapon systems', *Global Policy*, 10:3 (2019), pp. 338–42.

[96]The United States and Russian nuclear doctrine maintain the option for counterforce operations to limit the damage it would suffer from a nuclear exchange or believe that the other side might launch a counterforce attack. Observers have also debated whether India and China are moving in the same direction. See, US Department of Defense, *Nuclear Posture Review* (Washington, DC: Office of the Secretary, 2018), p. 23; Christopher Clary and Vipin Narang, 'India's counterforce temptations: Strategic dilemmas, doctrine, and capabilities', *International Security*, 43:3 (2018/2019), pp. 7–52; and Talmadge, 'Would China go nuclear?', pp. 50–92.

[97]Zhang Yao, Wang Yonghai, Wang Jinghua, Li Manhong, Lu Ruimin, and Wang Liyan, 'Performance analysis and research of LRASM, the next generation of US anti-ship missile', *Aerodynamic Missile Journal* (15 July 2018), available at: {https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFQ&dbname=CJFDLAST2018&filename=FHDD201807006&v=MjI3NzZyV00xRnJDVVI3cWZZT1pzRmlybVdyN0FJeVhQYXJHNEg5bk1xSTlGWW9SOGVYMUx1eFlTN0RoMVQzcVQ=}; and John Keller, 'Air Force Asks Lockheed Martin to Build Three More LRASM Anti-Ship Missile Systems for High-Priority Targets', Military & Aerospace Electronics (8 December 2018), available at: {https://www.militaryaerospace.com/computers/article/16726716/air-force-askslockheed-martin-to-build-three-more-lrasm-anti-ship-missile-systems-for-highpriority-targets}.

[98]See Johnson, 'Artificial Intelligence in nuclear warfare', pp. 197–211.

[99]Posen, *Inadvertent Escalation*, pp. 1–27.

[100]David Ochmanek and Lowell H. Schwartz, *The Challenge of Nuclear-Armed Regional Adversaries* (Santa Monica, CA: RAND Corporation, 2008); and Avery Goldstein, 'First things first: The pressing danger of crisis instability in US-China relations', *International Security*, 37 (2013), pp. 49–89.

[101]See Jeffrey Berejikian, 'A cognitive theory of deterrence', *Journal of Peace Research*, 39:2 (2002), pp. 165–83; and Daniel Kahneman and Ames Tversky, 'Prospect theory: An analysis if decision making under risk', *Econometric*, 47 (March 1979), pp. 263–91.

[102]See Jack S. Levy, 'Loss aversion, framing, and bargaining: The implications of prospect theory for international conflict', *International Political Science Review*, 17:2 (1996), pp. 179–95.

position or reputational damage.[103] If, for instance, Chinese or North Korean leaders are faced with the prospect of an imminent attack on Taiwan, or Pyongyang view their regime survival at stake, they would likely countenance greater risks to avoid this potential (existential) loss.[104] Furthermore, these capabilities' crisis instability could also result from irrational behaviour derived from misperception, cognitive biases, or other emotional impulses, which makes nuclear escalation more likely.[105]

For example, Chinese analysts tend to overestimate the US's military AI capabilities relative to open-source reports – often citing outdated or inaccurate projections of US AI 'warfighting' budgets, development, and force posture.[106] The framing of Chinese discussion on US military AI projects is analogous to Soviet concerns about the missile gap with the US during the Cold War;[107] thus, risk compounding Beijing's fear that AI technology could be strategically destabilising.[108] In a world with imperfect (and asymmetric) information about the balance of power and resolve, and incentives to misrepresent and manipulate perceptions (exploit psychological dispositions and vulnerabilities) and emotions (strategic framing and fear appeals) of the information ecosystem, bargaining failure, and war are more likely.[109]

Given the confluence of secrecy, complexity, erroneous, or ambiguous intelligence data (especially from open-source intelligence and social media outlets),[110] AI-augmentation will likely exacerbate compressed decision-making and the inherent asymmetric nature of cyberspace information.[111] For example, using AI-enhanced cyber capabilities to degrade or destroy a nuclear-states command and control systems – whether as part of a deliberate coercive counterforce attack or in error as part of a limited conventional strike – may generate pre-emptive 'use it or lose it'

[103]For a recent study that finds elite actors' decision to go nuclear would are heavily influenced by how their affect their personal status and reputation, see Pauly B. C. Reid, 'Would U.S. leaders push the button? Wargames and the sources of nuclear restraint', *International Security*, 43:2 (November 2018), pp. 151–92.

[104]For example, several analysts warned about the inadvertent escalation risks associated with the US's Air-Sea Battle operational concept to counter China's anti-access, area-denial capabilities in the Western Pacific. See Joshua Rovner, 'AirSea Battle & Escalation Risks', Policy Brief No. 12, University of California Institute on Global Conflict and Cooperation (January 2012).

[105]Schelling, *Arms and Influence*.

[106]See, for example, Zhang Shen, Ji Zili, and Wang Wenhua, 'Overview of the development of US military intelligent weapons and equipment', *Military Digest* (1 September 2019), available at: {https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFQ&dbname=CJFDLAST2019&filename=JSWN201917015&v=MjQ4NTk4ZVgxTHV4WVM3RGgxVDNxVHJXTTFGckNVUjdxZllPWnNGaXJtVkwzQUx6N2NZTEc0SDlqTnFJOUVZWVI=}; and Zhang Shen, Ji Zili, and Wang Wenhua, 'US Military to Speed Up the Development of Smart Weapons', Defense Science & Technology Industry (15 August 2019), available at: {https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFQ&dbname=CJFDLAST2019&filename=ZGBG201908016&v=MTg5OTBoMVQzcVRyV00xRnJDVVI3cWZZT1pzRmlybVZydkxQeXJKYWJHNEg5ak1wNDlFWW9SOGVYMUx1eFlTN0Q=}.

[107]During the early years of the Cold War, this fear – though unfounded and likely exaggerated by US intelligence services – also cut the other way. The prospect of losing the nuclear arms race to the Soviets and reinvigorated the US government to restore US nuclear superiority. See Greg Thielmann, 'LOOKING BACK: The missile gap myth and its progeny', *Arms Control Today*, 41:4 (2011), pp. 44–8.

[108]Currently, China lags the United States in the development of military AI. Chinese AI engineers face significant technical obstacles in developing and deploying AI applications, including constraints on service members' technical literacy and the availability of data and computing power. Ryan Fedasiuk, *Chinese Perspective on AI & Future Military Capabilities* (Washington, DC: CSET Policy Brief, 2021).

[109]See James Fearon, 'Rationalist explanations for war', *International Organization*, 49:3 (1995), pp. 379–414; and Kelly Greenhill and Ben Oppenheim, 'Rumor has it: The adoption of unverified information in conflict zones', *International Studies Quarterly*, 61:3 (2017), pp. 660–76.

[110]For example, reports suggest that during the 2017 US-North Korean tensions, both Kim Jong Un and Donald Trump got much of their information from social media and other media outlets. Tom O'Connor, 'Like Trump, North Korea's Kim Jong Un gets his news from TV and Twitter', *Newsweek* (2 November 2017), available at: {https://www.newsweek.com/north-korea-kim-jong-un-relies-social-media-us-news-ignores-trump-701672}.

[111]See James Johnson, 'The AI-cyber nexus: Implications for military escalation, deterrence, and strategic stability', *Journal of Cyber Policy*, 4:3 (2019), pp. 442–60; and James M. Acton, 'Cyber warfare and inadvertent escalation', *Daedalus*, 149:2 (2020), pp. 133–49.

situations.[112] In a US-China conflict scenario, for instance, a US penchant for counterforce operations targeting adversaries' command and control, the comingled nature of China's (nuclear and conventional) missile forces, US and Chinese preference for the pre-emptive use of cyberweapons, and domestic-political pressures on both sides to retaliate for costly losses (either physical/kinetic or non-physical/political), increases the dangers of inadvertent escalation.[113]

These risks should give defence planners pause for thought using advanced conventional capabilities to project military power in conflicts with regional nuclear powers. In short, conventional doctrines and operational concepts could exacerbate old (for example, third-party interference, civil-military overconfidence, regime type, accidental or unauthorised use, or an overzealous commander with pre-delegation authority) and create new pathways (for example, AI-enhanced ISR and precision missile targeting and guidance, drone swarms, AI-cyberattacks, and mis/disinformation subversion) to uncontrollable inadvertent escalation. Missile defences and advanced conventional weapons are unlikely to prevent these escalatory mechanisms once battlefield perception shifts and the nuclear threshold is crossed.[114]

To the extent to which a state may succeed in limiting the damage from a nuclear exchange in using technologically enhanced (including AI) counterforce operations continues to be an issue of considerable debate.[115] However, from the perspective of inadvertent escalation, the efficacy of damage-limitation counterforce tactics is less significant than whether an adversary views them as such. Chinese and Russian fear that the United States is developing and deploying conventional counterforce capabilities – (especially cyberattacks on NC3 systems) to blunt their nuclear deterrence risks generating 'crisis instability' caused by 'use it or lose it' pressures – that is, pressures to use nuclear weapons before losing the capability to do so.[116]

Nuclear powers maintain different attitudes and perceptions on the escalation risk posed by cyberattacks and information warfare more generally, however. Chinese analysts, in particular, have expressed an acute awareness of the potential vulnerabilities of their respective NC3 systems to cyberattacks.[117] The United States has begun to take this threat more seriously, whereas Russian strategists, despite bolstering their cyber defences appear more sanguine and view information warfare as a continuation of peacetime politics by other means.[118] Consequently, the risk of inadvertent escalation caused by misperception and miscalculation will likely increase.[119]

[112]Posen, *Inadvertent Escalation*, pp. 13–14.

[113]Goldstein, 'First things first', pp. 49–89; and David Gompers and Martin Libicki, 'Cyber warfare and Sino-American crisis instability', *Survival*, 56 (2014), pp. 7–22.

[114]See Horowitz, Scharre, and Velez-Green, 'A stable nuclear future?'.

[115]See, for example, Lieber and Press, 'The new era of counterforce', pp. 9–49; Charles L. Glaser and Steve Fetter, 'Should the United States reject MAD? Damage limitation and US nuclear strategy toward China', *International Security*, 41:1 (2016), pp. 63–70; Austin Long and Brendan Rittenhouse Green, 'Stalking the secure second strike: Intelligence, counterforce, and nuclear strategy', *Journal of Strategic Studies*, 38:1–2 (2015), pp. 38–73; Rafael Loss and Joseph Johnson, 'Will artificial intelligence imperil nuclear deterrence?', *War on the Rocks* (19 September 2019), available at: {https://warontherocks.com/2019/09/will-artificial-intelligenceimperil-nuclear-deterrence/}; and Wilkening, 'Hypersonic weapons and strategic stability', pp. 129–48.

[116]For example, a cyberattack could be used to 'blind' or spoof an adversary's early warning systems in advance of launching a broader kinetic counterforce strike on its nuclear forces. See Acton, 'Cyber warfare & inadvertent escalation', pp. 133–49.

[117]See Fiona S. Cunningham and M. Taylor Fravel, 'Assuring assured retaliation: China's nuclear posture and US-China strategic stability', *International Security*, 40:2 (2015), pp. 15–23; Bracken, 'The cyber threat to nuclear stability', pp. 197–200; and Alexei Arbatov, Vladimir Dvorkin, and Sergey Oznobishchev, *Non-Nuclear Factors of Nuclear Disarmament: Ballistic Missile Defense, High-Precision Conventional Weapons, Space Arms* (Moscow: IMEMO RAN, 2010), available at: {https://www.files.ethz.ch/isn/144178/10002.pdf}.

[118]By contrast, NATO states tend to view information warfare as limited and tactical.

[119]See Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume I: Euro-Atlantic Perspectives* (Stockholm: Stockholm International Peace Research Institute, May 2019); Edward Geist and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Santa Monica, CA: RAND Corporation, 2018); and Michael Horowitz, Paul Scharre, and Alexander Valez-Green, 'A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence' (2019), available at: {https://arxiv.org/pdf/1912.05291.pdf}.

In what ways might the new tools and techniques emerging in digitised information exacerbate these dynamics?

The digitised information ecosystem, human psychology, and inadvertent risk

Misperceptions, cognitive bias, and the human psychological features of security dilemma theorising can also be used to elucidate the escalatory dynamics that can follow from inflammatory, emotionally charged, and other offensive public rhetoric (see, for example, fake news, disinformation, rumours, and propaganda) used by adversaries during crisis – or saber-rattling behaviour.[120] During, in anticipation of, or to incite a crisis or conflict, a state or non-state actor (for example, clandestine digital 'sleeper cells') could employ subconventional (or 'grey zone') information warfare campaigns to amplify its impact by sowing division, erode public confidence, and delaying an effective official response.[121]

The public confusion and disorder that followed a mistaken cell phone alert warning residents in Hawaii of an imminent ballistic missile threat in 2018 serve as a worrying sign of the vulnerabilities of US civil defences against state or non-state actors' seeking asymmetric advantages *vis-à-vis* a superior adversary – that is, compensating for its limited nuclear capabilities.[122] North Korea, for example, might conceivably replicate incidents like the Hawaii false alarm in 2018 in a disinformation campaign (that is, issuing false evacuation orders, issuing false nuclear alerts, and subverting real ones via social media) to cause mass confusion.[123]

During a crisis in the South China Seas or South Asia, for example, when tensions are running high, state or non-state disinformation campaigns could have an outsized impact on influencing crisis stability (dependent on the interpretation and processing of reliable intelligence) with potentially severe escalatory consequences. This impact would be compounded when populist decision-makers heavily rely on social media for information-gathering and open-source intelligence and thus more susceptible to social media manipulation.[124] *In extremis*, a populist leader may come to view social media as an accurate barometer of public sentiment, eschewing official (classified and non-classified) evidence-based intelligence sources, and regardless of the origins of this virtual voice – that is, from genuine users or fake accounts as part of a malevolent disinformation campaign. Consequently, the agenda-setting framing narrative of decision-makers during a crisis would instead be informed by a fragmented and politicised social media information ecosystem; amplifying rumours, conspiracy theories, and radical polarisation, which in turn, reduces the possibility of achieving a public consensus to inform and legitimatise decisions during a crisis. Such dynamics may also expose decision-makers to increased 'rhetorical entrapment' pressure whereby alternative policy options (viable or otherwise) may be overlooked or dismissed out of hand.[125]

Furthermore, increased public scrutiny levels – especially coupled with disinformation and public panic – could further increase political pressures on leaders whose electoral success determines their political survival.[126] Under crisis conditions, these dynamics may compromise diplomatic de-escalation efforts and complicate other issues that can influence crisis stability,

[120]See Robert Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976).

[121]Hersman, 'Wormhole escalation', pp. 91–109.

[122]Tim Starks, 'Hawaii missile alert highlights hacking threat to emergency services', *Politico* (16 January 2018), available at: {https://www.politico.com/newsletters/morning-cybersecurity/2018/01/16/hawaii-missile-alert-highlights-hacking-threat-to-emergency-systems-074411}.

[123]For example, in 2017, US forces in Korea received false SMS and Facebook messages ordering an evacuation. Kim Gamel, 'US forces in Korea warns of fake evacuation messages', *Stars & Stripes* (21 September 2017), available at: {https://www.stripes.com/news/pacific/us-forces-korea-warns-of-fake-evacuation-messages-1.488792}.

[124]Greg Sargent, 'Could Trump help unleash nuclear catastrophe with a single tweet?', *Washington Post* (26 December 2016), available at: {https://www.washingtonpost.com/could-trump-help-unleash-nuclear-catastrophe-with-a-single-tweet}.

[125]See Frank Schimmelfennig, 'The community trap: Liberal norms, rhetorical action, and the eastern enlargement of the European Union', *International Organization*, 55:1 (2001), pp. 47–80.

[126]Christopher Gelpi, 'Democracies in conflict: The role of public opinion, political parties, and the press in shaping security policy', *Journal of Conflict Resolution*, 61:9 (2017), pp. 1925–49.

including maintaining a credible deterrence and public confidence in a state's retaliatory capability and effective signalling resolve to adversaries and assurance to allies.[127] State or non-state disinformation campaigns might also be deployed in conjunction with other AI-augmented non-kinetic/political (for example, cyberattacks, deep fake technology, or disinformation campaigns via social media amplified by automated bots) or kinetic/military (see, for example, drone swarms, missile defence, anti-satellite weapons, or hypersonic weapons) actions to distract decision-makers – thus, reducing their response time during a crisis and conferring a tactical or operational advantage to an adversary.[128]

For example, in the aftermath of a terrorist attack in India's Jammu and Kashmir in 2019, a disinformation campaign (see, for example, fake news and false and doctored images) that spread via social media amid a heated national election,[129] inflamed emotions and domestic-political escalatory rhetoric, that in turn, promoted calls for retaliation against Pakistan and brought two nuclear-armed adversaries close to conflict.[130] This crisis provides a sobering glimpse of how information and influence campaigns between two nuclear-armed adversaries can affect crisis stability and the concomitant risks of inadvertent escalation. In short, the catalysing effect of costly signalling and testing the limits of an adversary's resolve (which did not previously exist) to enhance security instead increases inadvertent escalation risks and leaves both sides less secure.

The effect of escalatory imbued rhetoric in the information ecosystem can be a double-edged sword for inadvertent escalation risk. On the one hand, public rhetorical escalation can mobilise domestic support and signal deterrence and resolve to an adversary – making war less likely. On the other hand, sowing public fear, distrust (for example, confidence in the legitimacy and reliability of NC3 systems), and threatening a leader's reputation and image (for example, the credibility of strategic decision-makers and robustness of nuclear launch protocols) domestically can prove costly, and in turn, may inadvertently make enemies of unresolved actors. For example, following the Hague's Permanent Court of Arbitration ruling against China over the territorial disputes in the South China Seas in 2016, the Chinese government had to resort to social media censorship to stem the flood of nationalism, calling for war with US ally the Philippines.[131] Furthermore, domestic public disorder and confusion – caused, for example, by a disinformation campaign or cyberattacks – can in itself act as an escalatory force, putting decision-makers under pressure to respond forcefully to foreign or domestic threats, to protect a states' legitimacy, self-image, and credibility.[132]

These rhetorical escalation dynamics can simultaneously reduce the possibility for face-saving de-escalation efforts by either side – analogous to Thomas Schelling's 'tying-hands

[127]Rose McDermott, Anthony Lopez, and Peter Hatemi, '"Blunt not the heart, enrage it": The psychology of revenge and deterrence', *Texas National Security Review*, 1:1 (2017), pp. 68–89.

[128]Today, given the incipient nature of these technologies, there are active debates about the impact of social media on the dissemination and diffusion of (true and false) information on social media. For example, and contrary to conventional wisdom, while robots tend to accelerate the spread of both true and false news, false news (especially political in nature) spreads more than the truth because humans, not robots, are more predisposed to spread it. Soroush Vosoughi, Deb Roy, and Sinan Aral, 'A large-scale analysis of tweets reveals that false rumors spread further and faster than the truth', *Science*, 359:6380 (2018), pp. 1146–51.

[129]Most notably, disinformation spread via Facebook's WhatsApp that falsely claimed that a leader of the Indian National Congress Party had offered a bribe to the suicide bomber's family. Despite Facebook's efforts to contain the nefarious campaign, the misinformation was disseminated to more than 2.8 million Facebook users. Neha Thirani Bagri, 'Back story: When India and Pakistan clashed, fake news won', *Los Angeles Times* (15 March 2019), available at: {https://www.latimes.com/world/la-fg-india-pakistan-fake-news-20190315-story.html}.

[130]Social media disinformation campaigns can impact information flow in a crisis in two ways: it undermines the transmission of information from the decision-makers to the public and from the public to decision-makers. In combination, these dynamics can impair the timeliness, reliability, and accuracy of publicly disseminated information. Trinkunas, Lin, and Loehrke, *Three Tweets to Midnight*, pp. 69–73.

[131]Kenneth Tan, 'Chinese censors harmonize online posts calling for war following South China Sea ruling', *Shanghaiist* blog (13 July 2016), available at: {http://shanghaiist.com/2016/07/13/hague_ruling_censored/amp/}.

[132]Ibid., p. 69.

mechanism'.[133] During heightened tensions between the United States and North Korea in 2017, for instance, the Trump administration's heated war of words with Kim Jong Un, whether a madman's bluff or in earnest (or 'rattle the pots and pans') raised the costs of Kim Jong Un backing down (that is, with regime survival at stake), thus increasing inadvertent escalation risk, and simultaneously, complicating de-escalation.[134] Because of the fear that its nuclear (and conventional) forces are vulnerable to a decapacitating first strike, rhetorical escalation between a conventionally inferior and superior state is especially dangerous.[135] Research would be beneficial on how the contemporary information ecosystem might affect decision-making in different political systems.

Ultimately, states' willingness to engage in nuclear brinkmanship will depend upon information (and mis/disinformation), cognitive bias, and the perception of, and the value attached to, what is at stake. Thus, if one side considers the potential consequences of not going to war as intolerable (that is, regime survival, the 'tying-hands', or 'use it or lose it' pressures), then off-ramps, firebreaks, or other de-escalation measures will be unable to prevent crisis instability from intensifying.[136] Finally, to the extent to which public pressure emanating from the contemporary information environment affects whether nuclear war remains 'special' or 'taboo' will be critical for reducing the risk of inadvertent escalation by achieving crisis stability during a conventional war between nuclear-armed states. Future research would be beneficial (a) on how the digitised information ecosystem affects decision-making in different political regimes; and (b) the potential effect of asymmetry and learning in the distribution of countries with advanced AI-capabilities and dynamics associated with its adoption. Will nuclear-armed states with advanced AI-enabled capabilities treat less advanced nuclear peers that lack these capabilities differently? And how might divergences in states synthesis and adoption of military AI contribute to misperception, miscalculation, and accidents?

Policy implications

How can decision-makers mitigate the inadvertent escalation risks associated with AI and nuclear systems? Possible ways forward include, *inter alia*, arms control and verification, changes to norms and behaviour, unilateral measures and restraint, and bilateral and multilateral stability dialogue. AI technology is already raising a multitude of questions about warfare and shifts in the balance of power, which are challenging traditional arms control thinking.[137] Traditional arms control and non-proliferation frameworks of nuclear governance are not necessarily obsolete, however.[138] Instead, we will need to depart from conventional siloed, rigid, and stove-piped approaches and search for innovative frameworks and novel approaches to meet the challenges of the rapidly evolving dual-use technology, the linkages between conventional and nuclear weapons, and the informational challenges in the new nuclear age. An asymmetric arms control framework emphasises the importance of dynamism – allowing for mutual adjustment in force posture in ways that differ from the traditional 'like-for-like' approach to arms control – in designing such agreements would be a sensible starting point.[139]

[133]The 'tying-hands mechanism' refers to the idea that states seek to increase the credibility of their threats and resolve by taking costly actions (that is, 'go public' with its threats and demands) that increase the costs of backing down were the other side to counter-escalate but which would otherwise incur few costs. See Schelling, *Arms and Influence*.

[134]Trinkunas, Lin, and Loehrke, *Three Tweets to Midnight*, pp. 117–21.

[135]Morgan et al., *Dangerous Thresholds*, p. 105.

[136]Richard Ned Lebow, 'The deterrence deadlock: Is there a way out?', *Political Psychology*, 4:2 (1983), pp. 333–54.

[137]Amandeep Singh Gill, 'Artificial Intelligence and international security: The long view', *Ethics and International Affairs*, 33:2 (2019), pp. 169–79.

[138]Alexey Arbatov, *An Unnoticed Crisis: The End of History for Nuclear Arms Control?* (Moscow: Carnegie Moscow Center, June 2015), available at: {https://carnegieendowment.org/files/CP_Arbatov2015_n_web_Eng.pdf}.

[139]Heather Williams, 'Asymmetric arms control and strategic stability: Scenarios for limiting hypersonic glide vehicles', *Journal of Strategic Studies*, 42:6 (2019), pp. 789–813.

Recent discussion about AI technology (especially lethal autonomous systems) and arms control has focused on how military AI might be managed, restricted ('keeping humans in the loop'), or prohibited – for targeting and use of nuclear weapons.[140] AI could, perhaps counterintuitively, also offer innovative solutions to develop new and revise legacy arms control frameworks and contribute to non-interference mechanisms (NTM) for arms control verification – reducing the need for 'boots on the ground' inspectors in sensitive facilities.[141] For instance, AI object identification applications to augment satellite imagery of missile production facilities or test ranges, and pattern recognition tools (that detects anomalies from vast amounts of data),[142] might be used to support arms verification efforts, identify cheating behaviour under an arms control agreement, assess the nature of suspicious military movements, and in turn, enhance the credibility of future strategic arms control agreements.[143]Authentic, verified, and reliable open-source information should also be leveraged to support these gathering and analysis efforts.

The use of AI-augmentation to enhance states' early warning and detection systems to improve target identification might prevent false positives (or nuclear 'close calls'),[144] reduce bias, and improve the understanding of an adversary's actions (or reduce the 'fog of war'), thus lowering the risk of inadvertent escalation during a crisis.[145] Also, incorporating AI into early warning systems may be particularly stabilising for countries that do not have the advanced satellites, sensors, and forward-deployed radar systems that the US and Russia have developed to ensure missile launches are detected and assessed for threat potential.[146] AI-augmented early warning and detection systems could, for instance, offer Beijing improved transparency and confidence about US military operations in the Indo-Pacific (that is, to discern and discriminate between nuclear and conventional weapons systems in an incoming attack), thus reducing inadvertent escalation dynamics caused by miscalculation, false positives, or surprise.[147]

AI technology could also improve the safety of nuclear systems. For instance, it could increase the security and robustness of command-and-control cyber defences by identifying undetected vulnerabilities or other potentially undiscovered weaknesses. The US Defense Advanced Research Projects Agency (DARPA) has, for example, already begun to study the ways AI may

[140]See Wyn Bowen et al., 'The human side of verification: Trust and confidence', in *Trust in Nuclear Disarmament Verification* (London, UK: Palgrave Macmillan, 2018).

[141]For example, the Campaign to Stop Killer Robots, and the UN Convention on Certain Conventional Weapons (CCW) Group of Government Experts (GGE) on LAWS, to name a few.

[142]The CIA, for example, uses AI technology to support a range of pattern-recognition tasks, including: geo-locating images without the associated metadata, creating 3D models from satellite imagery, and inferring a building's function based on 'pattern-of-life' analysis. Kelley M. Sayler, 'Artificial Intelligence and National Security', Congressional Research Service (30 January 2019), p. 8, available at: {https://crsreports.congress.gov/product/details?prodcode=R45178}.

[143]For example, Rand Corporation has conducted studies on the use of AI technology to track mobile missiles. RAND has been exploring the use of AI to track mobile missiles, for example. See Paul K. Davis, *Applying Artificial Intelligence Techniques, 16; Security 2040: How Might Artificial Intelligence Affect the Risk of Nuclear War* (Washington, DC: Rand Corporation, 2018).

[144]See Patricia Lewis et al., *Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy* (London, UK: Chatham House, 2014).

[145]For example, to differentiate between the launch of a demonstration rocket and an ICBM (which have similar radar signatures) rapidly and accurately or to discern a conventional from a nuclear missile launch. Besides, AI-supported big data analytics could be used to collate and process electronic data (that is, signals, imagery, and open-source information) to identify patterns of behaviour and launch profiles unique to specific types of capabilities. Jessica Cox and Heather Williams, 'The unavoidable technology: How Artificial Intelligence can strengthen nuclear stability', *The Washington Quarterly*, 44:1 (2021), pp. 69–85.

[146]Ibid., pp. 73–5.

[147]For a contrary view that further automation of state's NC3 systems would be fundamentally destabilising, see Mark Fitzpatrick, 'Artificial Intelligence and nuclear command and control', *Survival*, 61:3 (2019), pp. 81–92; and James Johnson, 'Delegating strategic decision-making to machines: Dr. Strangelove redux?', *Journal of Strategic Studies* (2020), available at: {https://doi:10.1080/01402390.2020.1759038}.

be used to identify vulnerabilities in conventional military systems.[148] Success in efforts such as these might also help bolster nuclear deterrence and mitigate inadvertent (and accidental) escalation risk. AI could also support defence planners' design and manage wargaming and other virtual training exercises to refine operational concepts, test various conflict scenarios, and identify areas and technologies for potential development.[149] Thus, enabling participants to better prepare against adversaries in unpredictable, fast-moving environments and where unpredictable and counterintuitive human-machine and machine-machine interactions will inevitably take place.[150]

Finally, expanding the topics and approaches for bilateral and multilateral initiatives such as confidence-building measures, should include the novel non-kinetic escalatory risks associated with complexity in the AI and the digital domain (see, for example, dis/misinformation, deepfakes, information sabotage, and social media weaponisation) during conventional crises and conflict involving nuclear-armed states.[151] Today, AI technology is not currently integrated into states' nuclear targeting, command and control, or launch systems; thus, a narrow window exists for nuclear powers (the P5 members as well as India, Pakistan, Israel, and NATO states) to agree on new principles, practices, and norms (for example, banning attacks on nuclear-armed states' NC3 systems), and enshrine these into international law. For instance, within the framework of the United Nations Convention on Certain Conventional Weapons and tailored to the specific features of the technology and coordinated by the UN Conference on Disarmament.[152] Specific measures might include prohibiting or imposing limits on the fusion of AI technology in nuclear command and control systems, autonomous nuclear-armed missiles, and nuclear weapons launch decisions.[153]

Conclusion

To what extent might AI-enabled capabilities increase inadvertent escalation risk? In a global security environment characterised by great power strategic competition and regional strategic asymmetry (capabilities, domains, and stakes), new rungs, firebreaks, and thresholds on the escalation ladder are already challenging conventional assumptions of deterrence, strategic stability, and escalation. This article underscores the need for greater clarity and discussion on the specific characteristics of AI technology that may create new (or disrupt old) rungs on the metaphorical escalation ladder, and in turn, increase the risk of inadvertently transitioning crises between nuclear-armed (and especially regional) states from conventional to nuclear confrontation. The article builds on and adapts the foundational work on inadvertent escalation conducted at the end of the Cold War (on the cusp of the 'Second Nuclear Age'). Specifically, it examines the psychological underpinnings of escalation theorising to elucidate whether and how characteristics of AI technology, contextualised with the broader digital information ecosystem, might destabilise crisis stability, and increase inadvertent escalation risk.

[148]Johanna Curiel, 'Darpa cyber grand challenge recap: How bots will help lift the security game', *TechBeacon*, available at: {https://techbeacon.com/security/darpa-cyber-grand-challenge-recap-how-bots-will-help-lift-security-game}.

[149]Craig S. Smith, 'AI war games and the challenge of China', *Forbes* (12 June 2020), available at: {https://www.forbes.com/sites/craigsmith/2020/06/12/ai-war-games-and-the-challenge-ofchina/#603bc8946c74}.

[150]Paul K. Davis, *Applying Artificial Intelligence Techniques to Strategic-Level Gaming and Simulation* (Washington, DC: Rand Corporation Paper Series, P-7120, November 1985), available at: {https://www.rand.org/content/dam/rand/pubs/papers/2008/P7120.pdf}.

[151]See Michael C. Horowitz and Paul Scharre, 'AI and International Stability: Risks and Confidence-Building Measures', Center for a New American Security (12 January 2021), available at: {https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures}; and Michael C. Horowitz and Lauren Kahn, 'How Joe Biden can use confidence-building measures for military uses of AI', *Bulletin of the Atomic Scientists*, 77:1 (2021), pp. 33–5.

[152]Whether and how military AI and autonomous weapons can and should be regulated is highly contested issue among policymakers, scholars, and campaigning activists. See Frank Sauer, 'How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies', *Contemporary Security Policy*, 42:1 (2021), pp. 4–29.

[153]For a recent study on AI and arms control, see Matthijs M. Maas, 'How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons', *Contemporary Security Policy* (2019), pp. 285–311.

The article highlights three critical features of inadvertent escalation risk in the emerging AI-nuclear strategic nexus. First, while nuclear-armed states have a shared interest in avoiding nuclear war (expressed by the 'rational deterrence theory'), they also place a high value on their nuclear forces, which advanced conventional weapons enhanced by AI technology inadvertently threaten – especially in asymmetric military situations. The synthesis of AI technology into conventional weapon systems to enhance autonomy, remote sensing, missile guidance, and situational awareness creates new tactical possibilities for a range of novel destabilising conventional counterforce possibilities (for example, cyberattacks on NC3 and locating survivable nuclear retaliatory capabilities). In regional asymmetric crisis or conflict, AI-powered tools to find, track, and destroy a state's nuclear forces may be viewed by the other side as a precursor to a disabling counterforce attack, thus increasing incentives to escalate a conventional situation inadvertently. Further, these technological developments have been accompanied by destabilising doctrinal shifts by several regional nuclear powers (Russia, North Korea, and possibly China), compounding the problem of commingled dual-use conventional and nuclear weapons missions at a strategic level.

Second, escalatory rhetoric and other aggressive behaviour, amplified by the digital information ecosystem, might be misperceived as unprovoked malign intent – not as a response to the initiator's behaviour – leading to action and reaction spirals of potentially irrevocable inadvertent escalation. Cognitive and heuristic burdens caused by information overload and complexity will likely make decision-makers more susceptible to cognitive bias, misperceptions, and heuristics to approach complex problem solving. Also, new AI-enhanced tools are enabling a more comprehensive range of actors (state and non-state) with asymmetric techniques (for example, dis/misinformation and cyberattacks) to improve their bargaining hand by delegitimising (or legitimising) the use of nuclear weapons during an escalating situation – or suing for peace or de-escalating a situation. The study also demonstrates the impact of escalatory rhetoric in the information ecosystem could be a double-edged sword for inadvertent escalation risk at the political decision-making level. On the one hand, public rhetorical escalation can mobilise domestic support and signal deterrence and resolve to an adversary, thus making war less likely. On the other hand, sowing public fear, distrust, and threatening a leader's reputation and image domestically can inadvertently make enemies of unresolved actors.

Third, and related, the state of heightened tension, uncertainty, complexity, and compressed decision-making (or 'the fog of war') of modern digitised warfare will likely be dramatically increased with AI technology's infusion. That is, restricting the options and time available for de-escalation, compounding the problem of offence-defence distinguishability, and increasing the risks of both horizontal and vertical inadvertent escalation. Moreover, the increasing dependencies and concomitant vulnerabilities on digital technologies (especially AI) and information to conduct modern warfare create a new security paradox. This 'paradox' could create resource vulnerabilities that generate first strike and pre-emption incentives, predicated upon use it or lose it' pressures, whether real or illusory.

In today's nuclear multipolar world with great power techno-military competition (US-China and US-Russia) and regional asymmetric dynamics there is a political imperative to address the challenges for inadvertent escalation by engaging in broader strategic stability talks about the development of new and innovative normative frameworks.[154] Multipolarity – a function of political relations – exacerbates techno-military competition, which in turn, has important implications for strategic stability, deterrence, the security dilemma, and escalation dynamics described in this article. Against the backdrop of increasing populism, amplified and manipulated in the digital information ecosystem, long-term regional and global stability and nuclear security efforts

[154]James Johnson, 'Does the United States face a multipolar future? Washington's response through the lens of technology', in Benjamin Zala, *National Perspectives on a Multipolar Order Interrogating the Global Power Transition* (Manchester, UK: Manchester University Press, 2021), pp. 121–44.

will continue to be jettisoned in favour of short-termism, asymmetric and great power competition, political fragmentation, and asymmetric and great power competition – to gain the 'first-mover advantages' in areas such as AI, hypersonic weapons, and information warfare.[155]

When the problem arises because of technological creativity, and compliance and verification are concerned with abstaining from future breakthroughs and potential uses of AI, technological solutions are unlikely to be an adequate response – that is, the process is not reducible to structural processes and assumptions of 'rational' actors, agency matters.[156] Therefore, any solution to the emerging AI-nuclear strategic challenge must be a much a political as a technological one. Solutions will have to help adversaries 'escape an *irrational* situation where it is precisely *rational* [decision-making] behavior that may be most dangerous' – reducing perceived vulnerability in the short-term at the expense of future inadvertent risk.[157]

Acknowledgements. The authors would like to thank Matthew Bunn, Andrew Futter, Michael Smith, Benjamin Zala, and the anonymous reviewers for comments and feedback on the draft of this article.

James Johnson is a Lecturer in the Department of Politics and International Relations, University of Aberdeen. He is also an Honorary Fellow at the University of Leicester, a Non-Resident Associate on the ERC-funded 'Towards a Third Nuclear Age' project, and a Mid-Career Cadre with the Center for Strategic and International Studies (CSIS) Project on Nuclear Issues. He is the author of The US-China Military & Defense Relationship During the Obama Presidency and Artificial Intelligence and the Future of Warfare: USA, China & Strategic Stability (Palgrave Macmillan, 2018). His latest book project is entitled Artificial Intelligence: Nuclear Strategy and Risk in the Digital Age.

[155]James Johnson, 'The end of military-techno Pax Americana? Washington's strategic responses to Chinese AI-enabled military technology', *The Pacific Review* (2019), available at: {https://doi:10.1080/09512748.2019.1676299}.

[156]Futter and Zala, 'Strategic non-nuclear weapons and the onset of a Third Nuclear Age', pp. 18–19.

[157]Hakan Wilberg, Ib Damgaard Petersen, and Paul Smoker (eds), *Inadvertent Nuclear War* (New York, NY: Pergamon Press, 1993), p. 18.

Cite this article: Johnson, J. 2021. Inadvertent escalation in the age of intelligence machines: A new model for nuclear risk in the digital age. *European Journal of International Security* X: 1–23. https://doi.org/10.1017/eis.2021.23