Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?

James Johnson

James Martin Center for Nonproliferation Studies, Middlebury Institute of International Studies, Monterey, CA, USA.

Abstract

Will the use of artificial intelligence (AI) in strategic decision-making be stabilizing or destabilizing? What are the risks and trade-offs of pre-delegating military force (or automating escalation) to machines? How might non-nuclear (and non-state actors) state leverage AI to put pressure on nuclear states? This article analyzes the impact of strategic stability of the use of AI in the strategic decision-making process, in particular, the risks and trade-offs of pre-delegating military force (or automating escalation) to machines. It argues that AI-enabled decision support tools by substituting the role of human critical thinking, empathy, creativity, and intuition in the strategic decision-making process will be fundamentally destabilizing. In particular, if defense planners come to view AI's 'support' function as a panacea for the cognitive fallibilities and human analysis and decision-making. The article also considers the nefarious use of AI-enhanced fake news, deepfakes, bots, and other forms of social media by non-state actors and state proxy actors, which might cause states to exaggerate a threat from ambiguous or manipulated information, increasing instability.

Keywords Artificial intelligence; U.S.-China relations; nuclear security; deterrence policy; emerging technology; strategic stability

Introduction

Will the use of artificial intelligence (AI) in strategic decision-making be stabilizing or destabilizing? [1] How might synthesizing AI with nuclear command, control, and communications (NC3) early-warning systems impact the nuclear enterprise? The compression of detection and decision-making timeframes associated with the computer revolution is not an entirely new phenomenon.[2] During the Cold War, the United States and Soviets both automated their nuclear command-and-control, targeting, and early-warning detection systems to strengthen their respective retaliatory

[1] Recent progress in AI falls within two distinct fields: (1) 'narrow' AI and in particular the machine-learning AI sub-set; and (2) 'general' AI, which refers to AI with the scale and fluidity akin to the human brain. Most AI researchers anticipate that 'general' AI to be at least several decades away, if at all. Narrow AI is already utilized in the private sector, in particular, in data-rich research fields and applied sciences (e.g., predictive analytics for market research, consumer behavior, logistics, and quality control systems). See, Stuart Russell, and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014); Nils J. Nilsson, *The Quest for Artificial Intelligence*, (New York, NY: Cambridge University Press, 2010); and Stuart Russell, *Human Compatible* (New York, NY: Viking Press, 2019).

[2] Michael Horowitz, Paul Scharre, and Alex Velez-Green, *A Stable Nuclear Future? The Impact of Automation, Autonomy, and Artificial Intelligence* (Philadelphia: University of Pennsylvania, 2017).

capabilities against a first strike.[3] Technologies developed during the 1950s paved the way for modern undersea sensors and spaced-based communication, and over-the-horizon radar.[4] Further, many of the systems and concepts introduced in the 1960s are still used today.[5]

In the post-Cold War era, the emergence of nuclear multipolarity has created multifaceted escalation pathways to a nuclear confrontation involving nine nuclear-armed states, compared to the Cold War dyadic.[6] Sophisticated NC3 networks interact with nuclear deterrence through several key vectors: [7] (1) early warning satellites, sensors, and radars (e.g., to detect incoming missile launches); (2) gathering, aggregating, processing, and communicating intelligence for C2 planning (i.e., to send and receive secure and reliable orders and status reports between civilian and military leaders);[8] (3) missile defense systems as a critical component of nuclear deterrence and warfighting postures; and (4) monitoring, testing, and assessing the security and reliability of sensor technology, data, and communications channels, and weapon launch and platforms, used in the context of NC3.[9]

NC3 systems supply critical linkages between states' nuclear forces and their leadership, ensuring decision-makers have the requisite information and time needed to command and control (C2) nuclear forces. In short, NC3 systems are a vital pillar of the states' deterrence and communications, to ensure robust and reliable command and control over nuclear weapons under all conditions - and can have a significant impact on how wars are fought, managed, and terminated.[10] Because of the pivotal nature of

[3] In 2011, Commander-in-Chief of the Russian SRF, General S. Karakayev, confirmed in an interview with one of the central Russian newspapers that *Perimeter* exists and is operational. The system's characteristics and capabilities are unknown, however. Ryabikhin Leonid, "Russia's NC3 and Early Warning Systems," *Tech4GS* July 11, 2019.

[4] For example, the first Defense Support Program satellite for ballistic missile launch warning was launched in 1970, remains a core element of U.S. early warning infrastructure.

[5] Geoffrey Forden, Pavel Podvig, and Theodore A. Postol, "False Alarm, Nuclear Danger," *IEEE Spectrum* 37/3 (2000), 31-39.

[6] Michael Krepon, "Can deterrence ever be stable?" *Survival*, 57/3 (2015) 111-132.

[7] Modern nuclear command, control, and communications (NC3) systems include: early-warning satellites, radars, and sensors; facilities to collect and interpret early warning information; fixed and mobile networked command posts; and a communications infrastructure that includes landlines, satellite links, radars, radios, and receiving terminals in ground stations and aboard strike vehicles. See, Amy Woolf, *Defense Primer: Command and Control of Nuclear Forces* (Washington, DC: Congressional Research Service), December 11, 2018, 1.

[8] For example, DARPA's PETE (Professional, Educated, Trained, and Empowered) AI-enabled virtual assistant gathers, collates information, as well as liaising and executing orders from commanders. Peter W. Singer, *Wired for War* (New York, NY: Penguin Group, 2009), 359.

[9] Jon R. Lindsay, "Cyber Operations and Nuclear Weapons," *Tech4GS Special Reports*, June 20, 2019.

[10] Jeffrey Larsen, "Nuclear Command, Control, and Communications: U.S. Country Profile," *Tech4GS* Special Reports, August 22, 2019.

these systems to the nuclear enterprise, superior systems would likely outweigh asymmetries in arsenals sizes - and thus, put an adversary with less capable systems and more missiles at a disadvantage.

Nuclear security experts have cataloged a long list of computer errors, unstable or components, early warning radar faults, lack of knowledge about adversary's capabilities and *modus operandi* (especially missile defense systems), and human mistakes that led to nuclear accidents and demonstrated the limitations and potential for malicious interference of inherently vulnerable NC3 systems.[11] The risks and trade-offs inherent in NC3 systems since the Cold War-era, reflecting the complex social, emotional, heuristic, and cognitive evolution of human agents, making decisions amid uncertainty, will likely be amplified by the inexorable and ubiquitous complexity, uncertainty, and unpredictability that AI introduces. In particular, the military concept 'mission command.'[12] This concept holds that commanders' strategic-psychology (or Clausewitz's military 'genius') depends on the intuition, flexibility, and empathy of subordinates to implement the spirit of commander's intentions - especially in the context of uncertainty and incomplete information associated with modern warfare.[13]

AI-augmented systems operating at machine speed and reacting to situations in ways that may surpass humans' comprehension, might challenge the 'genius' of commanders and heuristics in strategic decision-making and raise broader issues about escalation control and the start of a slippery slope towards the abandonment of human moral responsibility.[14] That is, the uncertainties and unintended outcomes of machines interpreting human intentions, and making autonomous strategic decisions, in fundamentally non-human ways. A central risk posed by AI may not be the generation of bias, or decisions based on AI fuzzy logic, rather the temptation to act with

[11] For example, see Bruce Blair, *Strategic Command and Control: Redefining the Nuclear Threat* (Washington DC: Brookings Institution, 1985); Shaun Gregory, *The Hidden Cost of Deterrence: Nuclear Weapons Accidents* (London: Brassey's, 1990); Scott D. Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton, NJ: Princeton University Press, 1995); and Eric Schlosser, *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety* (New York: Penguin, 2014).

[12] Col. (Ret.) James D. Sharpe Jr. and Lt. Col. (Ret.) Thomas E. Creviston, "Understanding mission command," *U.S. Army*, July 10, 2013.

[13] A notable example of the failure of 'mission command' is the ill-timed ICBM test at the U.S. Vandenberg Air Force Base during the height of the 1962 Cuban Missile Crisis. Stephen J. Cimbala, *The Dead Volcano: The Background and Effects of Nuclear War Complacency*, (New York, NY: Praeger, 2002), 66.

[14] Wendell Wallach and Colin Allen, *Moral Machines* (New York, NY: Oxford University Press, 2009), 40.

confidence and certainty in response in situations that would be better managed with caution and prudence.[15]

Whereas the potential tactical and operational effects of AI are today qualitatively axiomatic,[16] its effect at a strategic level (especially for assessments of military power and strategic intention) remains uncertain, however.[17] On the one hand, future AI-augmented C2 systems may overcome many of the shortcomings inherent to human strategic decision-making during wartime, such as a susceptibility to invest in sunk costs, skewed risk judgment, cognitive heuristics, and group-think.[18] On the other hand, AI systems that enable planners to predict the production, commissioning, deployment, and use of nuclear forces by adversaries might *in extremis* undermine first-strike stability - or the premise of MAD - and make future nuclear war winnable.[19]

To date, much of the existing (albeit limited) literature has focused on how nuclear states' leverage of AI might impact stability - especially the fear of nuclear decapitation.[20] However, given the complex nuclear interactions between nuclear and non-nuclear (and non-state) actors, it is critical to expand the discussion to consider how AI may create new - or exacerbate existing - escalation pathways. How might variations such as regime type, nuclear doctrine, strategy, strategic culture, or force structure make states' more or less predisposed to developing AI in the nuclear domain?

The remainder of this article proceeds in two parts. Part one describes how defense planners might use AI in the strategic decision-making process. It examines the notion of human psychology to elucidate how and why militaries might use AI in the strategic decision-making process; despite commanders' conservative instincts and

[15] David Whetham and Kenneth Payne, "AI: In Defence of Uncertainty," *Defence in Depth*, December 9, 2019.

[16] Vincent Boulanin (ed.) *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. I Euro-Atlantic Perspectives* (SIPRI Publications, Stockholm: May 2019).

[17] Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies* 39/5-6, (2016), 793-819.

[18] Ben Connable, *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*, (Santa Monica, CA: RAND Corporation, 2012).

[19] For example, AI used in conjunction with autonomous mobile sensor platforms might compound the threat posed to the survivability of mobile ICBM launchers. See, Paul Bracken, "The Cyber Threat to Nuclear Stability," *Orbis* 60/2 (2016), 194.

[20] For example, see Vincent Boulanin (ed.) *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. I Euro-Atlantic Perspectives*); Edward Geist and Andrew Lohn, *How might artificial intelligence affect the risk of nuclear war?* (Santa Monica, CA: RAND Corporation, 2018); Mary L. Cummings, *Artificial intelligence and the future of warfare*, (London, UK: Chatham House, 2017); Todd S. Sechser, Neil Narang, and Caitlin Talmadge, "Emerging technologies and strategic stability in peacetime, crisis, and war," *Journal of Strategic Studies*, 42/6, 727-735; and Michael C. Horowitz, Paul Scharre, and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," December 2019, arXiv:1912.05291

humans' inherent distrust of machine-generated information. Next, it examines the risks and trade-offs of increasing - inadvertently or otherwise - the role of machines in the strategic decision-making process. Part two considers the implications for the nuclear enterprise of synthesizing AI with NC3 systems. Will this synthesis will be stabilizing or destabilizing? How might non-nuclear (and non-state actors) state leverage AI to place pressure on nuclear states?

AI strategic decision-making Magic 8-ball?

AI systems' recent success in several highly complex strategic games, has demonstrated insightful traits that have potentially significant implications for future military-strategic decision-making.[21] In 2016, for example, DeepMind's AlphaGo system defeated the professional Go master, Lee Sedol. In one game, the AI player reportedly surprised Sedol in a strategic move that "no human would ever do."[22] Three years later, DeepMind's AlphaStar system defeated one of the world's leading e-sports gamers at Starcraft II - a complex multiplayer game that takes place in real-time and in a vast action space with multiple interacting entities - and devised and executed complex strategies in ways a human player would unlikely do.[23] In short, existing rule-based machine learning algorithms would likely be sufficient to automate C2 processes further.

AI systems might undermine states' confidence in their second-strike capabilities, and potentially, affect the ability of defense planners to control the outbreak, manage the escalation, and terminate warfare. The central fear of alarmists focuses on two related concerns. First, the potentially existential consequences - i.e., dystopian *Terminator's* Skynet-like prophetic imagery - and existential consequences of AI surpassing human intelligence. Second, the possible dangers caused by machines absent human empathy (or other theory-of-the-mind emotional attributes), relentlessly optimize pre-set goals - or self-motivated future iterations that pursue their own - with

[21] DeepMind's AlphaStar victory represented a technical milestone in several ways: (1) using game-theory logic to continue exploring ways to improve and expanding its boundaries; (2) unlike games like chess or Go, operating in an imperfect information situation; (3) performing long-term planning in real-time; and (4) controlling a large and complex possibilities with combinatorial space possibilities (i.e., hundreds of units, personnel, and buildings), and in real-time. AlphaStar Team, "Alphastar: Mastering the Real-Time Strategy Game Starcraft II," *DeepMind Blog,* January 24, 2019.

[22] Cade Metz, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future," *Wired*, March 16, 2016.

[23] AI's technical milestones in a virtual environment would, however, unlikely be replicated in stochastic (i.e., randomly determined) and complex systems like NC3. See, AlphaStar Team, "Alphastar: Mastering the Real-Time Strategy Game Starcraft II," *DeepMind Blog*

unexpected and unintentional outcomes - or *Dr. Strangelove's* doomsday machine comparisons.[24]

Human commanders supported by AI, functioning at higher speeds, and compressed decision-making timeframes might, therefore, increasingly impede the ability - or the Clausewitzian 'genius' - of commanders to shape the action and reaction cycles produced by AI-augmented autonomous weapon systems. Existing rule-based machine learning algorithms would likely be sufficient to automate C2 processes further. For now, there is general agreement among nuclear-armed states that even if technological developments allow, decision-making that directly impacts the nuclear command and control should not be pre-delegated to machines - not least because of the explainability, transparency, and unpredictability problems associated with machine-learning algorithms.[25]

Psychologists have demonstrated that humans are slow to trust the information derived from algorithms (e.g., radar data and facial recognition software), but as the reliability of the information improves so the propensity to trust machines increases - even in cases where evidence emerges that suggests a machine's judgment is incorrect.[26] The tendency of humans to use automation (i.e., automated decision support aids) as a heuristic replacement for vigilant information seeking, cross-checking, and adequate processing supervision, is known as 'automation bias.' Despite humans' inherent distrust of machine-generated information, once AI demonstrates an apparent capacity to engage and interact in complex military situation (i.e., wargaming) at a human (or superhuman level), defense planners would likely become more predisposed to view decisions generated by AI algorithms as analogous (or even superior) with those of humans - even if these decisions lacked sufficiently compelling 'human' rational or fuzzy 'machine' logic.[27] Human psychology research has found that people are predisposed to do harm to others if ordered to do so by an authority figure.[28] As AI-

[24] Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," 814.

[25] See, Vincent Boulanin (ed.) *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. I Euro-Atlantic Perspectives*, 56-57.

[26] For example, see Linda J Skitka, Kathleen L Mosier, and Mark Burdick, "Does Automation Bias Decision-Making?," *International Journal of Human-Computer Studies* 51, no. 5 (1999), 991-1006; and Mary L Cummings, "Automation Bias in Intelligent Time-Critical Decision Support Systems," *AIAA 1st Intelligent Systems Technical Conference*, 2004, 557-562.

[27] AI experts predict that by 2040, AI systems may be able to play aspects of military war-games or exercises at superhuman levels. Edward Geist and Andrew J. Lohn, *How might artificial intelligence affect the risk of nuclear war?* 17.

[28] See, Marilynn B. Brewer, William B. Crando, *Social Psychology* (New York, NY: West Publishing Co. 1994).

enabled decision-makings tools are introduced into militaries, human operators may begin to view these systems as agents of authority (i.e., more intelligent and more authoritative than humans), and thus be more inclined to follow their recommendations blindly; even in the face of information that indicates they would be wiser not to.

This predisposition will likely be influenced, and possibly expedited by human bias, cognitive weaknesses (notably decision-making heuristics) assumptions, and the innate anthropomorphic tendencies of human psychology.[29] Experts have long recognized the epistemological and metaphysical confusion that can arise from mistakenly conflating human and machine intelligence, especially used in safety-critical high-risk domains such as the nuclear enterprise.[30] Further, studies have demonstrated that humans are predisposed to treat machines (i.e., automated decision support aids) that share task-orientated responsibilities as 'team members,' and in many cases exhibit similar in-group favoritism as humans do with one another.[31]

Contrary to conventional wisdom, having a human in the loop in decision-making tasks does also not appear to alleviate automation bias.[32] Instead, human-machine collaboration in monitoring and sharing responsibility for decision-making can lead to similar psychological effects that occur when humans share responsibilities with other humans, whereby 'social loafing' arises – the tendency of humans to seek ways to reduce their own effort when working redundantly within a group than when they work individually on a task.[33] A reduction in human effort and vigilance caused by these tendencies could increase the risk of unforced error and accidents.[34] In addition, a reliance on the decisions of automation in complex and high-intensity situations can make humans less attentive to - or more likely to dismiss - contradictory information,

[29] For example, U.S. Army investigators discovered that automation bias was a factor in the 2003 Patriot fratricides, in which Patriot air and missile defense operators mistakenly fired upon friendly aircraft during the second Iraq War. John K. Hawley, "Looking Back at 20 Years of MANPRINT on Patriot: Observations and Lessons," *Army Research Laboratory, ARL-SR-0158*, September 2007.

[30] David Watson, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence," *Minds and Machines* (2019) 29, 434.

[31] Clifford Nass, B.J. Fogg, and Youngme Moon, Y, "Can computers be teammates?" *International Journal of Human Computer Studies* 45, (1996), 669-678.

[32] Parasuraman, Raja, and Victor Riley, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, 52/3 (2010), 381-410.

[33] Ibid.

[34] One area of encouragement from recent studies on automation bias and errors is that participants who were made explicitly aware and received training on automation bias were less likely to make certain classes of errors. Linda J Skitka, Kathleen L Mosier, and Mark Burdick, "Does Automation Bias Decision-Making?" 991-1006.

and more predisposed to use automation as a heuristic replacement (or short-cut) for information seeking.[35]

The decision to automate nuclear capabilities might also be influenced by the regime type, political stability and legitimacy, and threat perceptions of a particular nuclear-armed state. An authoritarian nuclear-armed regime (i.e., China, North Korea, or Pakistan) that fears either an internal coup or foreign interference may elect to automate its nuclear forces so that only a small circle of trusted officials are involved in the nuclear enterprise.[36] China, for example, maintains strict controls on its nuclear command and control structures (i.e., separating nuclear warhead and delivery systems), and the evidence does not suggest Beijing has pre-delegated launch authority down the chain of command if a first strike decapitates the leadership. As a means to retain centralized command and control structures and strict supervision over the use of nuclear weapons, AI-enabled automation might become an increasingly amenable option to authoritarian regimes such as China.[37]

Moreover, a state that views its second-strike capabilities (including its NC3 systems) as vulnerable and insecure such as North Korea or perhaps China may be more inclined to automate its nuclear forces and launch postures. In short, non-democratic nuclear states with relatively centralized command and control structures, less confident in the survivability of their nuclear arsenal, and whose political legitimacy and regime stability is conditioned by the general acceptance of official narratives and dogma, would likely be more persuaded by the merits of automation, and less concerned about the potential risks - least of all the ethical, human cognitive, or moral challenges - associated with this decision. [1] Despite official Chinese statements supporting the regulation of military AI by global militaries, much of China's AI-related initiatives (e.g., the use of data for social surveillance to distill a social-credit

[35] See, Parasuraman, Raja, and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* 39/2 (June 1997), 230-53.

[36] During the Cold War, the Soviets developed a computer program known as 'VRYAN' designed to notify Soviet leaders of a pre-emptive U.S. nuclear strike. However, the data used to feed the system was often biased, and thus, propelled a feed-back loop that heightened the Kremlin's fear that the United States was pursuing first-strike superiority.

[37] Autocratic states may perceive an adversary's intentions differently from a democratic one. The belief that a regime's political survival (or legitimacy) is at risk might cause leaders to consider worst-case scenario judgments, and thus behave in a manner predicted by offensive realist scholars. Conversely, non-democratic leaders operating in closed political systems such as China might exhibit a higher degree of confidence or sanguinity in their ability to respond to perceived threats in international relations. Bias assessments from non-democratic regime's intelligence services might reinforce a leader's faith (or a false sense of security) in their diplomatic skill and maneuverability. See, Keren Yarhi-Milo, *Knowing the Adversary* (Princeton NY: Princeton University Press, 2014).

scoring system, and ubiquitous facial recognition policies) focus on the impact on social stability, and in particular, efforts to insulate the legitimacy of the regime against potential internal threats.[38]

By contrast, the political processes, accountability (especially elected leaders and head of state vis-à-vis public opinion), nuclear-launch protocols, nuclear strategy and doctrine, mature civil-military relations, and shared values between allies (i.e., U.S. and its NATO allies), in democratic societies should make them less predisposed - or at least more reticent and encumbered - in use of AI in the nuclear domain.[39] Perhaps the question to ask, therefore, is less *whether* AI will be integrated into NC3 systems, but rather by whom, to what extent, and at what cost to the nuclear enterprise?

A prediction revolution and automated escalation

At a theoretical level, the U.S. Defense Advanced Research Projects Agency's (DARPA) Knowledge-directed AI Reasoning Over Schemas (KAIROS) program demonstrates how NC3 systems infused with AI technology might function. KAIROS integrates contextual and temporal events of a nuclear attack into an analytics-based AI application, which can generate associated and prompt actionable responses.[40] KAIROS also highlights the need for AI-enhanced NC3 systems to handle the increasing entanglement of nuclear and conventional weapons and support systems. This blurring (or co-mingling) problem-set increases the potential for miscalculation and accidental escalation.[41] As a corollary, AI-enhanced early-warning detection systems must be able to reliably determine whether an imminent attack on its dual-use C2 systems (e.g., cyber offense) during a conventional conflict, is intended as a non-nuclear offensive campaign, or as a prelude to escalation to nuclear confrontation.

The biggest technical challenge for deploying a system like KAIROS is developing the ability to learn and adapt without the requirement for an iterative learning process - common in today's narrow AI applications such as Google Assistant, Google Translate language processing tools, and Google's AlphaGo supercomputer.

[38] For example, see Yuan Yi, "The Development of Military Intelligentization Calls for Related International Rules," *PLA Daily*, October 16, 2019.

[39] Brian W. Everstine, "DOD AI Leader Wants Closer Collaboration With NATO," Airforce Magazine, January 15, 2020.

[40] "Generating Actionable Understanding of Real-World Phenomena with AI," *DARPA* January 4, 2019.

[41] James M. Acton, "Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War," *International Security* 43/1 (2018), 56-99.

Further, the algorithms that power AI systems, like AlphaGo, are usually trained on vast datasets that are not available to support nuclear weapons.[42] Designing and training a machine learning algorithm for a nuclear early-warning, would, therefore, be almost entirely reliant on simulated data, which in the safety-critical nuclear domain would be extremely risky.[43]

Data limitations coupled with constraints on the ability of AI algorithms to capture the nuanced, dynamic, subjective accurately, and changeable nature of human commanders (or theory-of-the-mind functions) will mean that for the foreseeable future strategic decision-making will remain a fundamentally human endeavor.[44] That is, AI will continue to include some human agency - especially in collaboration with machines - to effectively manage the attendant issues associated with technological complexity and interdependence, avoiding, for now at least, the risks associated with pre-delegating the use of military force. Thus, AI technology used to support NC3 systems will continue to exhibit a similar penchant for cognitive bias and subjectively (e.g., attribution error, decision-making heuristics, path-dependency, and dissonance) that has long plagued the human foreign policy and national security decision-making process.[45]

While human agency should ensure that the role of AI in the nuclear domain is confined to a predominately tactical one, through the discharge of its 'support role,' it might nonetheless (and possibly inadvertently) influence strategic decisions that involve nuclear weapons. In other words, the distinction between the impact of AI at a tactical and strategic level is not a binary one.[46] Technology designed to augment

[42] As 'unsupervised' machine-learning techniques (as opposed to current supervised approaches) mature, the reliance on data and labeling (i.e., images, videos, and text) to support AI system's training environments is expected to decrease. See, Alexander Graves and Kelly Clancy, "Unsupervised learning: The curious pupil," *Deepmind*, June 25, 2019.

[43] NC3 systems must meet the 'always never' criteria, which is critical to deterrence. Nuclear weapons must always work when tasked to and never go off accidentally or without proper authorization. Moreover, NC3 systems must, under all circumstances, be able to execute a lawful order to employ nuclear force (known as 'positive control'). At the same time, NC3 must under all circumstances never allow the nuclear force to be used accidentally or by an illegitimate authority (known as 'negative control'). Jeffrey Larsen, "Nuclear Command, Control, and Communications: U.S. Country Profile," 10-11.

[44] The historical record demonstrates, human strategists rarely have a clear idea from the outset of what they are seeking to achieve through pursuing a particular strategic path, less still how these goals might be realized. Kenneth Payne, "Fighting On: Emotion and Conflict Termination," *Cambridge Review of International Affairs* 28/3 (August 2015), 480-97.

[45] Robert Jervis, *Perception and Misperception in International Politics*, (Princeton, N.J.: Princeton University Press), 1976.

[46] Kenneth Payne, *Strategy from Apes to Evolution Artificial Intelligence and War,* (Washington, DC: Georgetown University Press), 183.

autonomous weapon support systems ostensibly (e.g., intelligence, reconnaissance, and surveillance, and early-warnings systems) will nonetheless be involved in decisions that inform and shape strategic war-faring calculations.[47] The U.S. 2018 Nuclear Posture Review (NPR), for example, explicitly states that the DoD would pursue design support technologies (such as machine learning) to facilitate more effective and faster strategic decision-making.[48] In short, escalation at the tactical level could easily have *strategic effects*.

To support officers' construct operational plans, DARPA has designed AI-powered supports systems (Integrated Battle Command and 'Deep Green') that allow commanders to visualize, evaluate, anticipate an adversary's strategic intentions, and predict the impact of complex environments with changing parameters.[49] Chinese analysts have also begun to research the use of big-data and deep-learning AI techniques to enhance the processing speed and intelligence analysis of satellite images, to support the People's Liberation Army's (PLA) early warning capabilities and enable a 'prediction revolution' in future warfare.[50] In 2017, the PLA Rocket Force's Engineering University participated in an international workshop that was convened to focus on intelligent reasoning and decision-making.[51] Besides, China has also applied AI to wargaming and military simulations and researched AI-enabled data retrieval, and analysis from remote sensing satellites,[52] to generate data and insights that might be used to enhance Chinese early-warning systems, situational awareness, and improve targeting.[53]

While AI-enabled decision support tools are not necessarily destabilizing, this non-binary distinction could risk AI 'support' tools substituting the role of critical thinking, empathy, creativity, and intuition of human commanders in the strategic

[47] Ibid.

[48] U.S. Office of the Secretary of Defense, *Nuclear Posture Review* (Washington DC: Department of Defense, February 2018), 57-58.

[49] Today, for example, big data already enables probabilistic prediction of people's political attitudes, and activism, and their violent tendencies. See, Jakob Bæk Kristensen et al., "Parsimonious Data: How a Single Facebook Like Predicts Voting Behavior in Multiparty Systems," *PLOS One* 12/9, (2017); Petter Bae Brandtzaeg, "Facebook Is No "Great Equalizer": A Big Data Approach to Gender Differences in Civic Engagement Across Countries," *Social Science Computer Review* 35/1, (2017), 103-25.

[50] Jia Daojin and Zhou Hongmei, "The Future 20-30 Years Will Initiate Military Transformation," *China Military Online*, June 2, 2016.

[51] "Evidence Reasoning and Artificial Intelligence Summit Forum," December 26, 2017.

[52] For example, see, "Chinese commercial space start-ups launch two AI satellites in a hundred days," *Global Times*, November 26, 2018.

[53] From the open sources, no unambiguous evidence has emerged to suggest China has - or plans to in the near-future - use AI to augment its NC3 systems.

decision-making process. The danger of delegating (inadvertently or otherwise) moral responsibility to machines raises broader issues about the degree of trust and reliance placed with these systems. Ethicists emphasize the need for AI as 'moral agents' to exhibit reasoning - as opposed to Bayesian reasoning and fuzzy logic - through not only careful reflection and deliberation but also an aptitude to effectively and reliably simulate human emotions (especially empathy) for interacting socially with humans in many contexts.[54]

Unlikely as it may be that commanders would delegate (at least knowingly) authority of nuclear missile launch platforms (e.g., ballistic missile submarines (SSBNs), bombers, missile launch facilities, and transporter erectors-launchers (TELs)); nuclear delivery vehicles (e.g., intercontinental ballistic missiles (ICBMs), torpedoes, missiles, nuclear-armed long-endurance unmanned autonomous vehicles (UAVs)); or NC3 systems to machines, AI technology is expected to be more widely used to support decision-making on strategic nuclear issues.[55] Though computer programs, simulations, and data analysis are already used to inform human defense-planners, AI operating at superhuman speed and performing increasingly complex tasks is likely to accelerate this trend, as Google's AlphaGo now-infamous defeat of the Go world champion attested.[56] If defense planners come to view this 'support' function as a panacea for the cognitive fallibilities and human analysis and decision-making, however, the reliance on these systems could have destabilizing consequences. AI machine-learning algorithms are only as good as the data and information they are trained on and supplied during operations.[57] Moreover, because of the paucity of data available for AI to learn from in the nuclear domain, designing an AI-augmented support tool to provide early-warnings systems with reliable information on preemptive nuclear strikes would be extremely challenging.

Although AI systems can function at machine speed and precision in the execution of military force, algorithms adhering to pre-determined mission goals are

[54] Wendell Wallach and Colin Allen, *Moral Machines*, chapters 3 and 4.

[55] At least two nuclear-armed states (i.e., the United States and Russia) are considering the notion of using UAVs or UUVs for the delivery of nuclear payloads. See, Roland Oliphant, "Secret Russian radioactive doomsday torpedo leaked on television," *Daily Telegraph*, November 15, 2015; and Dave Majumdar, "USAF leader confirms manned decision for new bomber," *Flight International*, April 23, 2013.

[56] Darrell Etherington, "Google's AlphaGo AI Beats the World's Best Hsaauman Go Player," *TechCrunch*, May 23, 2017.

[57] Joseph Johnson, "MAD in an AI Future?" Center for Global Security Research, Lawrence Livermore National Laboratory, June 3, 2019.

unable to empathize with humans, which is necessary to determine or anticipate the intentions and behavior of an adversary. That is intentions communicated in the use of military action to signal deterrence or resolve (i.e., the willingness to escalate) during a crisis. Machines would likely be worse (or at least less reliable) at understanding human signaling involved in deterrence, in particular signaling de-escalation.[58] Not only would machines need to understand human commanders and human adversaries, but they must also be able to interpret an adversary AI's signaling and behavior. Thus, an AI algorithm that is optimized to pursue pre-programmed goals might misinterpret an adversary simultaneously signaling resolve who is while seeking to avoid conflict or deescalate a situation. Absent reliable means to attribute an actors' intentions, AI systems may convey undesirable and unintended (by human commanders) signals to the enemy, thus complicating the delicate balance between an actor's willingness to escalate a situation as a last resort and keeping the option open to step back from the brink.[59]

Counterintuitively, states may view the expanded automation of their NC3 systems as a way to manage escalation and enhance deterrence - signaling to an adversary that any attack (or the threat of one) would trigger nuclear escalation. Because of the difficulty of demonstrating, and thus, effectively signaling this automation posture before a crisis or conflict, this implicit threat would likely intensify crisis instability. In short, if a nuclear-armed state used automation to reduce its flexibility during a crisis, and without the ability to signal this to an adversary, it would be akin to Herman Kahn's notion of "tearing out the steering wheel" in a game of chicken without being able to throw it out the window.[60]

Furthermore, unwarranted confidence and reliance on machines - known as 'automation bias'- in the pre-delegation of the use of force during a crisis or conflict, let alone during nuclear brinksmanship, might inadvertently compromise states' ability to control escalation. [61] Overconfidence, caused or exacerbated by automation bias in

[58] Wong, Yuna Huh, John M. Yurchak, Robert W. Button, Aaron Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae, *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020).

[59] Signaling during wartime has also been a challenging balancing act, especially between different strategic cultures. For example, during the Vietnam War, strategic game theory influenced U.S. bombing decision planning, but this approach underestimated the role of uncertainty and unpredictable human psychology during warfare. Peter W. Singer, *Wired for War*, 305-306.

[60] Herman Kahn, *On Escalation: Metaphors and Scenarios*, (New York: Prager, 1965), 11.

[61] See, Parasuraman, Raja and Manzey, Dietrich, "Complacency and Bias in Human Use of Automation: An Attentional Integration. Human factors," 381-410; and Mary L. Cummings, "Automation Bias in Intelligent Time-Critical Decision Support Systems," 557-562.

the ability of AI systems to predict escalation and gauge intentions - and deter and counter threats more broadly - could embolden a state (especially in asymmetric information situations) to contemplate belligerent or provocative behavior; it might otherwise have thought too risky.[62] This kind of misplaced confidence might also reduce - or even eliminate - the psychological uncertainty that injects caution into defense planning, which might exacerbate escalation risks during a crisis or conflict. For example, China's substantial investment and strategic interest in AI-augmented decision support systems - part of a broader doctrinal emphasis on the notion of information dominance through scientific central-planning - to enable the PLA to respond more reliably and faster to a disarming attack, suggests that Chinese commanders may be susceptible to automation bias.[63]

During nuclear brinksmanship, the ultimate competition in risk-taking, the interaction between machines and human-strategic psychology (or the war of ideas), caused by the pre-delegation of escalation to autonomous weapons (or automated escalation), may increase the risk of misinterpreting an adversary's intentions, thereby increasing the risks associated with closing the damage-limitation window and undermining crisis stability and increasing first-strike incentives. Moreover, AI-controlled NC3 systems would be more vulnerable to subversion from cyber-attacks, which could increase this risk of inadvertent escalation caused by human or machine miscalculation or error – even if humans are kept 'in the loop.'[64]

Competitive pressures could result in the implementation of AI applications (both offense and defense) *before* they are sufficiently tested, verified, or technically mature, will make these systems more error-prone and susceptible to subversion - in particular, a cyber-attack.[65] Thus, even a well-fortified and fully trained AI system might remain vulnerable to subversion that would be difficult to detect and even harder

[62] For the foreseeable future, AI machine learning - and in particular the deep-learning subset - used to predict an adversaries' intentions will be highly correlative and dependent on a range of engineering factors that rely on human-compiled data on historical patterns, and the parameters of select modeling practices. Benjamin M. Jensen, et al., "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review*, (June 2019), 15.

[63] Lora Saalman, "Lora Saalman on How Artificial Intelligence Will Impact China's Nuclear Strategy," *The Diplomat*, November 7, 2018.

[64] There is a range of different ways to subvert AI systems, and given the embryonic nature of AI cyber-defense, the offense is likely to have the upper hand in this domain for the near future. Hyrum S. Anderson, Anant Kharkar, Bobby Filar, and Phil Roth, Evading Machine Learning Malware Detection, *blackhat.com*, July 20, 2017.

[65] James Johnson, "The AI-cyber nexus: implications for military escalation, deterrence, and strategic stability," *Journal of Cyber Policy*, 4/3, 442-460.

to attribute.[66] A clandestine cyber-attack that undermines effective information and communication flow would risk increasing the incentives for escalation.[67] Although system inputs and outputs can be observed, the speed and scale of machine learning mechanisms would make it difficult for operators to isolate, and thus, explain a particular machine-generated prediction or decision.[68]

A double-edged sword for stability

AI-augmented support systems and the expanded use of automation in NC3, more broadly, could, in several ways, improve confidence in the existing nuclear enterprise. Making accidents caused by human-error - especially false warnings - and the unauthorized use of nuclear weapons less likely, thereby enhancing strategic stability. First and foremost, machine learning algorithms and autonomous systems could be used to bolster NC3 defenses against both physical (e.g., kinetic attacks against C2 nodes) or non-kinetic cyber threats (e.g., offensive cyber, jamming attacks, and electromagnetic pulses generated by a high-altitude nuclear burst effort).

Second, AI-augmented communications systems could improve information flow and situational awareness, enabling militaries to operate at scale in complex environments, in particular situations with incomplete information. In this way, military technology like AI might expand the decision-making timeframe available to commanders during a crisis, a perspective that has been overlooked by global defense strategic communities. Whether these enhancements by centralizing the decision-making process and creating a new breed of so-called 'tactical Generals' micro-managing theater commanders from afar, will improve military effectiveness or exacerbate uncertainties and is, however, an open question.[69]

Third, machine learning techniques coupled with advances in remote sensing technology might enhance nuclear early warning and testing systems, making accidents caused by error less likely. Finally, automating several NC3 functions may reduce the

[66] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples," *arXiv*, May 24, 2016.

[67] During the formulation of Russia's most recent military doctrine, Russian strategists proposed that an attack on Russia's early warning systems would be interpreted as a sign of an impending nuclear attack. This proposal was not, however, included in the final version. Ryabikhin Leonid, "Russia's NC3 and Early Warning Systems," *Tech4GS*, 10.

[68] The historical record is replete with examples of false alerts and warnings from satellite and over-the-horizon radars. See Andrew Futter, *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Washington DC: Georgetown University Press, 2018), chapter 2.

[69] Peter W. Singer, "Robots and the rise of "tactical Generals," *Brookings*, March 9, 2009.

risk of human error caused by cognitive bias, repetitive tasks and fatigue.[70] For example, UAVs might replace signal rockets to form an alternative airborne communications network, especially useful in situations where satellite communication is not possible.[71] In aggregate, these enhancements could enable a variety of operations, including bolstering non-nuclear capabilities such as cyber, air defenses, and electronic jamming, improving target identification and pattern recognition systems, controlling autonomous platforms, and improving the way workforce and logistics are managed; to name but few.

For instance, DARPA's Real-time Adversarial Intelligence and Decision-making (RAID) machine learning algorithm is designed to predict the goals, movements, and even the possible emotions of an adversary's forces five hours into the future. RAID relies on a type of game theory that shrinks down problems into smaller games, reducing the computational power required to solve them.[72] Similarly, BAE Systems is working with DARPA to design cognitive-based machine learning algorithms and data models aimed to give space operators the ability to identify abnormal activities from vast data sets, to predict possible threats such as space-based launches and satellite movements.[73] Future iterations of these systems may be able to identify risks (including risks unforeseen by humans), predict when and where a conflict will break-out, and offer strategic solutions and alternatives, and, ultimately, map out an entire campaign.[74] As a corollary, AI systems might react to dynamic and complex combat situations more rapidly, learn from their mistakes, and burdened with fewer cognitive shortcomings than human commanders such as human emotion, heuristics, and group-think.[75]

[70] Studies have demonstrated over 180 types of human cognitive biases and limitations (e.g., working memory, attention, confirmation bias, and loss aversion). Buster Benson, "Cognitive Bias Cheat Sheet," *Better Humans*, September 1, 2016.

[71] See, James Johnson, "Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare," *The RUSI Journal* (2020) DOI: 10.1080/03071847.2020.1752026

[72] In early tests of RAID between 2004 and 2008, the system performed with greater accuracy and speed than human planners. Alexander Kott, and Michael Ownby, "Tools for Real-Time Anticipation of Enemy Actions in Tactical Ground Operations," *Defense Technical Information Center*, June 2005.

[73] Sandra Erwin, "BAE wins DARPA contract to develop machine learning technology for space operations," *Spacenews,* August 13, 2019.

[74] AI-assisted predictive policing has already had some notable successes in combating crime. For example, in Los Angeles, police claim to have reduced burglaries by 33% and violent crime by 21% through AI-driven predictive policing, and Chicago has established an algorithmically derived list of individuals considered most likely to commit crimes. Similarly, Japan, Singapore, and, most notably, China, have implemented similar systems. Keith Dear, "Artificial Intelligence and Decision-Making," *The RUSI Journal*, 164/5-6, 18-25.

[75] In early 2019, Deepmind's AlphaStar beat the world's leading e-sports gamers at Starcraft II, a complex multiplayer game that takes place in real-time and in a vast action space with multiple

While state-of-the-art technology such as AI, machine learning, big-data analytics, sensing technology, quantum communications, and 5G supported networks integrated with nuclear early-warning systems, might alert commanders of incoming threats faster; the greater precision and scalability afforded by these advances could, however, exacerbate escalation risks in two ways.[76] First, AI machine learning used as force multipliers for the cyber offense (e.g., data poisoning, spoofing, deepfakes, manipulation, hacking, and digital jamming) would be considerably more difficult for early warning systems to detect - or detect in time.[77]

An adversary could, for instance, target 'blind spots' in machine learning neural networks to nefariously manipulate data in such a way that both the human operator and AI would not recognize a change - known as data-poisoning (or data-pollution).[78] An AI machine learning generative adversarial networks (GANs), deepfake, coupled with a data-poisoning attack, might trigger an escalatory crisis between two or more nuclear states.[79] For example, an adversary (state or non-state) could a use an image or recording of military commander obtained from open-sources to generate and disseminate a deepfake containing false orders, intelligence, or geospatial imagery that at best generates confusion, and at worst aggravates a tense situation or crisis between rival nuclear powers (this theme is developed below).[80] In this sense, deepfakes will likely become (or already are) another capability in the toolkit of warfighters to wage campaigns of disinformation and deception - one that both sides may use of have used

interacting entities. This virtual environment would unlikely perform so well in stochastic and complex systems, however.

[76] For example, in 1983, a malfunctioning Soviet early warning system led to the detection of a nonexistent U.S. attack.

[77] Experts believe that once a deepfakes becomes widespread, even having a human involved may not suffice to determine the veracity or source of a specific post or site. Nautilus Institute, Technology for Global Security, Preventive Defense Project, "Social Media Storms and Nuclear Early Warning Systems: A Deep Dive and Speed Scenario Workshop," *NAPSNet Special Reports*, January 8, 2019.

[78] Paige Gasser, Rafael Loss, and Andrew Reddie, "Assessing the Strategic Effects of Artificial Intelligence - Workshop Summary," *Center for Global Security Research, Lawrence Livermore National Laboratory*, September 2018, 9.

[79] A recent workshop hosted by IISS demonstrated that malign manipulation of input data received by early-warning systems might not only subvert the output of AI systems in specific situations but also undermine the reliability of an entire algorithm network environment. In particular, if an attack was executed during the 'training' phase for such programs (e.g., pattern recognition or intelligence gathering and analysis software). Mark Fitzpatrick, "Artificial Intelligence and Nuclear Command and Control," *Survival*, 61/3 (2019), 81-92.

[80] For example, in 2019, non-state actors used AI voice mimicking software to generate a fake recording of British energy firm executives to conduct the world's first reported AI-enabled theft. Drew Harwell, "An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft," *Washington Post*, September 4, 2019.

against them.[81]

Second, in the unlikely event, an attack, subversion, or manipulation was successfully detected, threat identification (or attribution) at machine speed would be virtually impossible. Further, once an operation is executed, human operators would be unable to monitor a systems' decision calculus in real-time. Thus, the ability to effectively monitor and control escalation - or deescalate a situation - would be impaired. Even if nuclear early-warning systems eventually detected an intrusion, heightened levels of uncertainty and tension caused by an alert might impel the respective militaries to automate further and put their nuclear weapons on high alert status - to reduce the vulnerability of their strategic forces. In sum, asymmetric situations between adversaries could prompt states to shift their nuclear doctrine and postures (e.g., endorsing a doctrine of pre-emption or limited nuclear strikes), and expedite the reconciliation an expanded use of AI in the nuclear domain - even at the expense of control and stability.[82]

During a crisis, the inability of a state to determine an attacker's intent may lead an actor to conclude that an attack (threatened or actual) was intended to undermine its nuclear deterrent capabilities.[83] Conversely, a malfunctioning early warning system caused by a malicious attack could mean a nuclear-state is oblivious of an imminent nuclear attack, thus impeding it from responding appropriately due to degraded nuclear decision making. China's fear that the PLA's early-warning systems are inadequate to respond to a U.S. disarming first strike (or nuclear decapitation) has impelled Chinese planners to prioritize the mitigation of *false negatives* (i.e., misidentifying a nuclear weapon as a non-nuclear one) over *false positives* (i.e., misidentifying a non-nuclear weapon as a nuclear one), which the United States generally emphasizes. To be sure, *both* false positives and false negatives can cause misperceptions and mischaracterizations in ways that can exacerbate escalation risk.[84] This skewed

[81] Disinformation to deception and misinformation campaigns is a familiar aspect of warfare, perhaps most famously demonstrated by the efforts of the Allies during World War II's *Operation Bodyguard* to mislead the Axis regarding the location of what became the D-Day invasion of 1944. Jamie Rubin, "Deception: The Other 'D' in D-Day," *NBC News*, June 5, 2004.

[82] Russia, Pakistan, and perhaps China have reportedly indicated a possible willingness to use limited nuclear strikes to end a conventional war that is losing. Michael C. Horowitz, Paul Scharre, and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," 32-33.

[83] For example, even if the malware detected in an attack was only capable of espionage, a target may fear that it also contained a "kill switch" able to disable an early-warning system after activation.

[84] James M. Acton, *Is this a Nuke? Pre-launch Ambiguity and Inadvertent Escalation* (New York, NY: Carnegie Endowment for International Peace, 2020), 14.

assessment in the context of nuclear weapons ready to launch nuclear weapons at a moment's notice could precipitate worst-case scenario thinking, triggering inadvertent escalation.[85]

According to open sources, operators at the North American Aerospace Defense Command (NORAD) have less than three minutes to assess and confirm initial indications from early-warning systems of an incoming attack.[86] This compressed decision-making timeframe could put political leaders under intense pressure to decide to escalate during a crisis, with incomplete (and possibly false) information of a situation. Paradoxically, therefore, new technologies designed to enhance information (i.e., modernized NC3 systems supported by 5G networks, AI machine learning, big-data analytics, and quantum computing) might simultaneously erode precise and reliable information flow and communication, critical for effective deterrence.[87] During times of high-pressure crisis decision-making, actors tend to interpret unusual circumstances as threatening, even if an adversary's behavior has not, in fact changed. As a result, routine activities (e.g., troop movements) scrutinized in the context of an early-warning alert, may be viewed as more menacing than they might otherwise be.[88]

In addition to the nuclear interactions between nuclear-armed dyads, nefarious information manipulation by non-state actors (i.e., terrorists, rogue lone actors, insurgent groups, and criminals) or state proxy actors could also have destabilizing implications for the increasing nuclear interactions in the world between nuclear and non-nuclear states, *both* during times of war and peace.[89] AI-enhanced fake news, deepfakes, bots, and other malevolent social media campaigns might exploit human psychology that fuels people’s engagement with social media to influence public opinion.[90] Creating false narratives, amplifying false alarms, or drowning out opposing

[85] In recent years, some Chinse military publications have indicated that China may adopt a launch-on-warning alert status for its nuclear forces. Gregory Kulacki, *China’s Military Calls for Putting Its Nuclear Forces on Alert* (Cambridge, M.A.: Union of Concerned Scientists, 2016).
[86] “Is Launch Under Attack Feasible?” *Nuclear Threat Initiative*, August 4, 2016.
[87] For example, the China Aerospace Science and Industry Corporation has become active in the development of dual-use networks for quantum communications, which may be used to transmit classified military information between command and control centers and military units during combat. Raymond Wang, “Quantum Communications and Chinese SSBN Strategy,” *The Diplomat*, November 4, 2017.
[88] Robert Jervis *How Statesmen Think: The Psychology of International Politics,* (New Jersey, Princeton: Princeton University Press, 2017), 222.
[89] Paige Gasser, Rafael Loss, and Andrew Reddie, “Assessing the Strategic Effects of Artificial Intelligence - Workshop Summary,” 10.
[90] See, Bill Davidow, “Exploiting the Neuroscience of Internet Addiction,” *The Atlantic*, July 18, 2012.

views with destabilizing effects, especially in times of external tension and internal strife.[91]

False social media reports (e.g., reports of mobile missiles movements, real-time streaming of launches, the deployment of TELs, or possible false reports of detonations) might influence the threat sensors of nuclear early warning systems used to inform strategic decision-making. The level of sophistication (i.e., technical know-how and software) needed to execute these kinds of attacks is surprisingly low, with many programs (e.g., voice cloning and GANs software) available at a relatively low-cost (or often free) on the internet. This portends the diffusion and democratization of ever more sophisticated technology, amplifying the human pathologies (i.e., cognitive heuristics) that underlie this 'information cascade' phenomenon - human attraction to novel, negative information, and filter bubbles, which explain why technologies like deepfakes are especially prone to going viral and perpetuating destabilizing memes and falsehoods.[92]

In 2017, counterintelligence officials in South Korea received fake mobile and social media alerts with orders for U.S. military and DoD personnel to evacuate the Korean Peninsula.[93] Information attacks such as this suggest that non-state actors, state proxy actors - and possibly state actors - will inevitably attempt to use social media as a tool of war to provoke nuclear confrontation for political-ideological, religious, or for other malevolent goals; and with increasing levels of sophistication, stratagem, and AI-enhanced subterfuge.[94] AI-augmentation might also enable states (and non-state actors) to automate, accelerate, and scale synthetic social media accounts and content to support malevolent disinformation operations.[95]

[91] Between August 2017 and January 2018, six instances of social media playing a role in nuclear-prone conflicts occurred in the Asia-Pacific region alone. See, Nautilus Institute, Technology for Global Security, Preventive Defense Project, "Social Media Storms and Nuclear Early Warning Systems: A Deep Dive and Speed Scenario Workshop," 1.

[92] Danielle K. Citron and Robert Chesney, "*Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*," 107 California Law Review 1753 (2019).

[93] Dan Lamothe, "U.S. families got fake orders to leave South Korea. Now counterintelligence is involved," *The Washington Post,* September 22, 2017.

[94] AI systems can track individuals or groups,' online habits, knowledge, and preferences to calibrate specific messages (i.e., propaganda) to maximize the impact on that individual (or group) and minimize the risk that the information being conveyed is questioned. This information can then be used by AI systems in real-time to determine the messages' influence; and, in this way, learn to become more effective in its task. Stuart Russell, *Human Compatible*, 105.

[95] Alina Polyakova, "Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare," *Brookings*, November 15, 2018.

Furthermore, the amplification of false alarms, or the creation of false signals by social media (i.e., false positives and false negatives), could disrupt critical communication flows between military commanders - especially SSBNs on extended deterrence patrols - and their political leadership, and between allies and adversaries during crisis or conflict. Authoritarian regimes (i.e., China, North Korea, Pakistan, and Russia), whose political legitimacy and regime stability is conditioned or legitimized by the general acceptance of official narratives and dogma, tend to become empowered in when people's trust in truth (i.e., faith in what they see and hear) is undermined, and the vacuum is filled by the opinions of authoritarian regimes and leaders with authoritarian inclinations.[96]

These dynamics may also be compounded by human cognitive bias - people's tendency to filter information through the lens of pre-existing beliefs and values.[97] Research has shown that people tend to interpret ambiguous information as consistent with their pre-existing beliefs - dismissing information that contradicts these views - and accept information that allows them to avoid unpleasant choices.[98] A motivated authoritarian leader (or non-state actor) would be well-positioned to use AI-augmented tools (e.g., 'fake news' and 'deepfake' propaganda) to exploit this psychological weakness - to ensure the control and dissemination of false narratives and opinions.

Irregular or opaque communication flow between adversaries may also increase the risk of misperception and miscalculation, and to assume the worst of others' intentions.[99] Moreover, in asymmetric situations between nuclear-armed adversaries, inferior NC3 early warning systems (e.g., ISR systems without long-range sensors and less able to detect the subtle differences between nuclear and conventional delivery systems) could put leaders under intense pressure to launch a pre-emptive strike due to the perceived use-them-or-lose-them imperatives. In sum, absent robust, modern, and

[96] In contested information environments, errors in early warning systems and decision-making processes are characterized by a lack of data, ambiguous indicators, mixed signals, and conflicting sensor data inputs. Amidst this complexity and ambiguity, false signals frequently occur (and are even expected); not least because sensor systems may not be cross calibrated to provide cross-checking confirmation. Nautilus Institute, Technology for Global Security, Preventive Defense Project, "Social Media Storms and Nuclear Early Warning Systems: A Deep Dive and Speed Scenario Workshop," 12.

[97] Robert Jervis, *Perception and Misperception* (1976), 117-202.

[98] Michela Del Vicario, Antonio Scala, Guido Caldarelli, Eugene H. Stanley, Walter Quattrociocchi, "Modeling confirmation bias and polarization," *Sci Rep*, January, 11, 2017.

[99] See Robert Jervis and Mira Rapp-Hooper, "Perception and Misperception on the Korean Peninsula," *Foreign Affairs*, April 5, 2018.

reliable NC3 structures decisions to threaten the use of military force and escalate a situation could be premised on false, fabricated, or misperceived narratives.[100]

Accidental escalation might be set into motion as follows: During heightened tensions or crisis between State A and State B, a third-party actor or terrorist leaks false information (e.g., satellite imagery, 3D models, or geospatial data) into an open-source crowdsourcing platform, about the suspicious movement of State A's nuclear road-mobile TELs.[101] Because of the inability of State B to determine with confidence the veracity of this information, and with mounting public pressures to respond, State B escalates a situation on the belief it is the target of an unprovoked attack.[102] Asymmetries between adversaries NC3 systems and military capabilities would likely exacerbate the escalation mechanisms illustrated in this fictional scenario.[103] Taken together, increasingly sophisticated GANs, the problem of attribution in cyberspace, the inherently dual-use nature of AI, coupled with the exponentially complex nature of NC3 systems, and the compressed timeframe for strategic decision-making, will continue to lower the threshold for false-flag operations.[104]

Because of the perennial trade-offs between speed, precision, safety, reliability, and trust inherent in cognitive-psychological human-machine interactions, greater emphasis is now needed on how and based on what assumptions, AI systems are designed to replicate human behavior (i.e., preferences, inferences, and judgments). In this way, AI can begin to instill trust in how it reaches a particular decision about the use of military force. As emerging technologies, including AI and autonomy, quantum technology, and big-data analytics are synthesized with and superimposed on states' legacy NC3 systems - at various speeds and degrees of sophistication - new types of

[100] Several nuclear weapons states - North Korea, Pakistan, and India - have much less capable early warning systems compared to the United States, using fewer satellites and other long-distance sensors such as radars with limited coverage. North Korea does possess any long-range sensor systems.

[101] From an intelligence standpoint, nuclear solid-fuel missiles and tracked TELs reduces the ability of ISR systems to detect signs of launch preparation. Moreover, solid fuel also increases the speed missiles can be launched and reduces the number of support vehicles to support an operation.

[102] Alternative outcomes from this fictional scenario are, of course, possible. For example, counter-AI systems might uncover the source or false nature of the leak before it can do severe damage. State A might also be able to assure State B through backchannel or formal diplomatic communications of this falsehood. While, social media platforms have had some success in slowing down the ability of users to orchestrate manipulative and dangerous campaigns, once these operations (e.g., deepfakes and bots) go viral the ability to curtail them becomes inexorably problematic - for human operators or machines.

[103] Other WMD-related information that would likely be considered equally escalatory might include intelligence about the movement of support vehicles delivering liquid-fuel to prepare a liquid-fueled missile for launch, spikes in radiation levels, or the detection of chemical warfare agents, to name but a few.

[104] See, Herbert Lin, "Escalation Dynamics and Conflict Termination in Cyberspace," *Strategic Studies Quarterly*, 6/3, (Fall, 2012), 46-70.

errors, distortions, and manipulations (notably involving social media) appear more likely to occur. Important questions for policymakers include. Should nuclear early warning systems include (or perhaps ignore) social media in their threat assessments? Is there a third party that can provide real-time status of nuclear forces to serve as an independent (and trusted) reference to inform states early warning systems? If so, who should take the lead in creating and verifying it?

To reiterate a central theme of this article: while autonomous nuclear early-warning systems would allow planners to identify potential threats faster and more reliably than before; absent human judgment and supervision, coupled with, the heightened speed of warfare, and the inherent brittleness and vulnerability of machine learning systems, the risk of destabilizing accidents and false alarms (or 'false positives') would likely rise. In this way, this discussion speaks to the broader conversation about how divergences between states' nuclear strategy, force structure, and doctrine might affect how they view the use of AI in the nuclear enterprise – or the AI-nuclear dilemma.[105]

For example, the U.S. defense community is generally more concerned than its Chinese counterparts that a low-level conventional conflict could escalate to a strategic level, but less worried about the ability of U.S. forces to control escalation *above* the nuclear threshold.[106] Divergent Sino-U.S. attitudes about de-escalation during a crisis might increase the risk that a conventional conflict escalates to a high-level of intensity.[107] In sum, states today face similar contradictions, dilemmas, and trade-offs in the decision about whether or not to integrate AI and autonomy into the nuclear enterprise, as leaders continue to face in the quest for strategic stability more generally.[108]

Mitigating the negative consequences of AI for strategic stability

How can incentives be altered to enhance strategic stability? A prominent theme that runs through this article - and central to understanding the potential impact AI for

[105] Stephen D. Biddle, and Robert Zirkle, "Technology, Civil-Military Relations, and Warfare in the Developing World," *Journal of Strategic Studies* 19/2 (1996), 171-212.

[106] James Johnson, "Chinese Nuclear 'War-fighting:' "An Emerging Intense U.S.-China Security Dilemma and Threats to Crisis Stability in the Asia Pacific," *Asian Security*, 15/3 (2019), 215-232.

[107] Fiona S. Cunningham and M. Taylor Fravel, "Dangerous Confidence? Chinese Views on Nuclear Escalation," *International Security* (2019), 44/2, 106-108.

[108] Robert Jervis, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon* (Ithaca, NY: Cornell University Press, 1989).

strategic stability and nuclear security more broadly - is the fear that AI systems operating at machine-speed will push the pace of combat to a point where the actions of machine actions surpass the (cognitive and physical) ability of human decision-makers to control (or even comprehend) events.[109] Effective deterrence depends on the clear communication of credible threats (and consequence of violation) between adversaries, which assumes the sender and recipient of these signals share a familiar context allowing for mutual interpretation.[110]

Possible policy responses to push back against the threat posed to stability from AI in a multipolar can be broadly categorized into two broad categories. First, those that focus on enhancing debate and discussion between researchers, global defense communities, decision-makers, academics, and other political and societal stakeholders. As we have seen, the scope and nature of the use of AI technology, and digitized information more broadly, to affect change in political processes vary across types of political regimes.[111] The kind of complex interaction between stakeholders and actors in liberal democracies contrasts with authoritarian regimes like China, where the state deliberately leverages these technologies for explicitly political ends (e.g., surveillance, persuasion, deception, and social engineering).[112] Incompatible approaches to and uses of AI technology between democratic and authoritarian nuclear powers may undermine efforts to sustain public and open debate on AI and nuclear security.

Second, a range of specific policy recommendations for great military powers to negotiate and implement.[113] Success in these endeavors will require all stakeholders to be convinced of the need and the potential mutual benefits of taking steps towards the establishment of a coherent governance architect to institutionalize and ensure compliance with the design and deployment of AI technology in the military sphere.[114]

[109] James Johnson, "Artificial Intelligence: A Threat to Strategic Stability," *Strategic Studies Quarterly*, 14:1, 16-39.

[110] Jon, R. Lindsay, and Erik Gartzke, eds. 2019 *Cross-domain deterrence: Strategy in an era of complexity* (Oxford: Oxford University Press, 2019), 19.

[111] Evgeny Morozov, *The Net Delusion: The Dark Side of Internet Freedom* (New York, NY: PublicAffairs, 2012).

[112] Rachel Botsman, "Big data meets Big Brother as China moves to rate its citizens," *Wired*, October 21, 2017.

[113] Center for a New American Security, University of Oxford, University of Cambridge, Future of Humanity Institute, OpenAI & Future of Humanity Institute, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, (Oxford, UK: Oxford University, February 2018), 51-55.

[114] A good case in point of this kind of inclusiveness and broader debate is the recent work carried out by the U.S. Defense Innovation Broad (DIB). The DIB is an independent advisory committee tasked with engaging with industry, academia, and the private sector to recommend an ethics framework and principles to enhance DOD's ethical standards as outlined in the DOD's debut AI Strategy. David

In short, this process must be an inclusive and broad conversation about the character and nature of warfare in the digital age.

Debate & dialogue

To pre-empt and mitigate (or at least manage) the destabilizing and escalatory risks posed by the military AI, great powers must closely coordinate their confidence-building measures to pre-empt some the kind of risks to stability.[115] Specifically, great powers should establish an international framework for governance, norms, regulation, and transparency in the development and deployment of AI-augmented military capabilities.[116] Further, these frameworks will need to encompass not only the present but also potential future developments, in particular, what is and is not being baked into AI algorithms and how best to temper the public debate from becoming too fixated on killer robots and machine overlords.

Governments' will likely face challenges in these efforts for several reasons. First, AI R&D is very widely dispersed across geographic locations and inherently opaque. Second, the potentially destabilizing and accident-prone features of AI applications can be difficult to identify during the development stages by system engineers. Third, the unpredictability of AI may cause a liability-gap if AI acts in unforeseeable ways, creating legal challenges caused by unintentional or unpredicted harm.[117] Several existing frameworks that govern dual-use technologies such as space law, internet, and aviation standards might offer some useful insights for the exploration of AI regulation, demonstrating that even in highly contested military domains, international consensus and areas of compromise can be successfully found.[118]

Vergun, "Defense Innovation Broad Recommends AI Ethical Guidelines," *U.S. Dept of Defense*, November 1, 2019.

[115] Concerns relating to the security risks posed by emerging technology (especially LAWs) and maintaining meaningful human control, has led to a variety of initiatives, reports, and other explorative efforts including: reports by the International Committee of the Red Cross, the International Committee for Robot Arms Control, the United Nations Institute for Disarmament Research, and the adoption of a framework and guiding principles by the UN Convention on Certain Conventional Weapons (CCW).

[116] The 2011 Vienna Document on Confidence and Security Building Measures remains one of the foundational sources of transparency, which could be updated to incorporate AI and autonomous weapons. For instance, remotely operated or UCAVs could be included in Annex III of the Vienna Document, together with, combat aircraft and helicopters. Other transparency measures in this document that could also be relevant for LAWS include airbases visits, demonstration of new types of major weapon and equipment system; prior notification of certain military activities; and observation of certain military activities. Vienna Document 2011 on Confidence-and Security-Building Measures (Vienna: Organization for Security and Co-operation in Europe, 2011).

[117] Matthew U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law and Technology*, 29/2, (2016), 354-400.

[118] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (London, Palgrave Macmillan, 2019).

Furthermore, decision-makers must carefully consider the nuanced trade-offs between increasing degrees of technological complexity, and the vulnerabilities and potential new escalation vectors military AI could engender. It is not an immutable fact (or trajectory) that military systems will be imbued with nascent - and potentially error and accident-prone - iterations of AI. Instead, these decisions will be made by human policymakers, tasked with reflecting on these trade-offs, and ultimately, implementing these innovations into safety critical NC3 systems.

Similar to the cyber domain, resistance to these efforts will likely come from states who worry that in revealing their military AI (especially offensive) capabilities, they could upend the deterrence utility of these tools.[119] To be sure, the challenge of coordinating and implementing policies like these will require bold and visionary leadership; to circumvent the inevitable regional agendas, interdisciplinary resistance, and burgeoning security dilemmas between strategic rivals. Because of the rapid technological change in AI formal treaties associated with arms-control agreements that require lengthy and complicated negotiation and ratification processes, legal frameworks risk becoming obsolete before they come into effect. The historical record has demonstrated on several occasions that these kinds of challenges facing humanity can be overcome.[120]

Next, the think-tank community, academics, and AI research experts should pool their resources to investigate the implications of the military AI for a range of potential security scenarios including: (1) the extent to which AI machine learning could reinforce bias in systems, and the impact of this bias on AI-augmented weapons; (2) the implications of dual-use AI applications for co-mingled nuclear and non-nuclear weapons and cross-domain deterrence;[121] (3) how to prepare for and react to artificial general intelligence; (4) how might investments in R&D affect the offense and defense balance for AI-enhanced military systems; and (5) measures to mitigate, prevent, and manage offensive uses of AI - both by nuclear powers, non-nuclear powers, and non-

[119] During the Cold War, nuclear deterrence worked in large part because the Soviets and Americans both knew they possessed nuclear arsenals to destroy the other, coupled with confidence in the integrity of this capacity to respond to a first strike.

[120] Examples include the 1968 NATO conference at Garmisch, which established a consensus around the mounting risks from software systems, and the 1975 NIH conference at Asilomar that underscored the risks posed by recombinant DNA research. NATO; Sheldon Krimsky, *Genetic alchemy: The social history of the recombinant DNA controversy* (MIT Press, Cambridge, MA, 1962).

[121] For example, see Dima Adamsky, "Cross-Domain Coercion: The Current Russian Art of Strategy," *IFRI Proliferation Paper* 54 (2015), 1-43.

state entities.[122]

The U.S. National Security Commission on Artificial Intelligence (NSCAI) is a new bipartisan commission established by the John S. McCain National Defense Authorization Act for Fiscal Year 2019.[123] The Commission's early work represents a rare example of a collaborative effort between academia, civil society organizations, and the private sector, to highlight the opportunities and risks of utilizing AI for national security purposes. Because of the intrinsic dual-use nature of AI, this dialogue should also be expanded to include other stakeholders such as private-sector AI and cybersecurity experts, the commercial sector, ethicists, philosophers, civil society, and public opinion.[124]

The extent the integration of AI into dual-use systems such as AI, autonomy, and cyber, might influence actors' attitude to risk, the offense-defense balance, and perceptions of others' intentions and capabilities, would have profound implications for strategic deterrence, nuclear stability, and arms control.[125] A recent study investigated historical uses of dual-use technologies - biological and chemical weapons, space weapons, cryptography, internet governance, and nuclear technology - to derive insights with applications for AI dual-use risk management policies, such as export controls and pre-publication reviews. This analysis demonstrated the immense difficulty of establishing regulatory, legal, and normative frameworks for dual-use technologies.[126]

[122] The AI-related research finding is often not made public due to reasons related to intellectual property and broader national security concerns.

[123] In August 2018, Section 1051 of the Fiscal Year 2019 John S. McCain National Defense Authorization Act established the National Security Commission on Artificial Intelligence (NSCAI) as an independent Commission: "to consider the methods and means necessary to advance the development of artificial intelligence, machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States."

[124] While many public opinion polls have shown very negative views of the notion of autonomous weapons and AI overall, other studies have demonstrated that the level of negativity can vary significantly depending on how the question(s) are asked. Public opinion can also change dramatically over time, as seen with other previously emerging technologies such as the computer, the VCR, and the telephone. Research on the subject also indicated that the malleability of public opinion. Further, it is crucial to recognize that different communities and cultures will have varying abilities to make such adaptations (e.g., technological literacy, culture norms, and economic systems), which may pose challenges for implementing security policies in society at large. See, Rob Sparrow, "Ethics as a source of law: The Martens Clause and autonomous weapons," *ICRC Blog*, November 14, 2017.

[125] The White House's 2016 'Partnership on AI,' series of workshops on AI, and the 2017 'Beneficial AI' conference in Asilomar, and the 'AI Now' conference series are good examples of this kind of research collaboration.

[126] For example, the cautionary tale of ineffective efforts in the late-1990s to regulate cryptographic algorithms and cyber-network security tools through export controls. Karim K. Shehadeh, "The Wassenaar Arrangement and Encryption Exports: An Ineffective Export Control Regime that

AI & arms control

Can arms control agreements encompass emerging technologies like AI?[127] How might non-proliferation look like in AI? Scholarship on arms control and strategic stability has demonstrated that success in these efforts is predicated on the ability of states to delineate between weapon platforms clearly.[128] During the Cold War-era, most arms control advocates believed that reciprocal reductions in arms reduced the incentives for disarming first strikes, thus promoting strategic stability. In the context of emerging (and especially dual use) technologies, that assumption will be increasingly tested.[129] In particular, when the lines between dual-use capabilities and nuclear and non-nuclear are blurred (or co-mingled), arms control is more challenging, and strategic competition is more likely to emerge.[130]

Existing arms control frameworks, norms, and the notion of strategic stability, more broadly, will increasingly struggle to assimilate and respond to these f and interconnected trends. Because AI is intrinsically dual-use and non-monolithic and fluid in nature, future discussions must consider the implications of AI-related and AI-enabling technologies, including big data, supercomputers, remote sensor and recognition technology - both on the battlefield and at a societal level - robotics, hypersonic weapons, cyber, and quantum computing, amongst others.

Another complicating factor is that currently, there are no precise definitions or engineering methodology for formulating regulation or maintaining control over AI systems to ensure their safety - or the AI 'control problem.'[131] For example, AI experts believe that existing tools - such as reinforcement learning techniques - are unable to

Compromises United States Economic Interests," *American University of International Law Review*, 15/1 (1999), 271-319.

[127] For recent scholarship on the opportunities and pitfalls of efforts to prevent or contain the militarization of AI see, Matthijs M. Maas, "How viable is international arms control for artificial military intelligence? Three lessons from nuclear weapons, *Contemporary Security Policy* 40/3 (2019), 285-311.

[128] Arms control can contribute to arms race stability in three ways: (1) placing reciprocal limits on capabilities; (3) increasing transparency into an adversary's capabilities; and (3) reducing the likelihood of success in the event of military adventurism. Thomas C. Schelling and Morton H. Halperin, *Strategy and Arms Control* (New York: The Twentieth Century Fund 1961).

[129] Keir Lieber and Daryl G. Press, "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security*, 41/4 (2017), 9-49.

[130] Heather Williams, "Asymmetric Arms Control and Strategic Stability: Scenarios for Limiting Hypersonic Glide Vehicles," *Journal of Strategic Studies* 42/6 (2019), 789-813.

[131] The AI 'control problem' refers to the issue that under certain conditions, AI systems can learn in unexpected and counterintuitive ways, which engineers and operators may have expected or align with their goals. See, Stuart Russell, *Human Compatible*, 251.

resolve the risks posed to humans from AI-augmented autonomous systems.[132] Thus, a different kind of algorithm than currently exists will likely be needed, which prioritizes the safety, robustness, and interpretability of complex AI-infused military systems (e.g., NC3 early warning systems).

Whether AI applications in the military domain can be formally verified, for now, remains an unanswered question. The complexity of AI systems, and in particular, the difficulty of defining their properties for formal verification, makes them less amenable to verification compared to other types of technology.[133] For example, DARPA's Assured Autonomy Program uses machine learning algorithms to ensure the safety of autonomous cyber-physical systems. Because this program is designed to learn continuously throughout its lifespan, assurance and verification using traditional methods is very challenging.[134]

This challenge is further complicated by the increasingly cross-domain nature of modern deterrence and the asymmetries emerging in both nuclear and non-nuclear strategic arenas, including cyber, hypersonic weapons, space, and AI.[135] These concerns resonated in the 2018 U.S. NPR. The NPR emphasized that the coalescence of geopolitical tensions and emerging technology in the nuclear domain, in particular, unanticipated technological breakthroughs in new and existing innovations - especially affecting nuclear command and control - might change the nature of the threats faced by the United States and the capabilities needed to counter them.[136]

To improve strategic stability in an era of rapid technological change, great power strategic competition, and nuclear multipolarity, therefore, the formulation of future arms control frameworks will need to reflect these new shifting perspectives. Further, arms control efforts can no longer be restricted to bilateral engagement. Governments should also explore ways to increase transparency and accountability for

[132] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "Cooperative Inverse Reinforcement Learning," 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

[133] Kathleen Fisher, "Using formal methods to enable more secure vehicles: DARPA's HACMS program," ICFP 2014: Proceedings of the 19th ACM SIGPLAN International Conference on Functional Programming.

[134] Sandeep Neema, "Assured Autonomy,"

[135] Historical examples of asymmetric negotiations and engagement across military domains, including nuclear weapons, are relatively limited, an exception being an option for such facilitation during the negotiation of the INF Treaty that was ultimately abandoned. See, Jack Snyder, "Limiting Offensive Conventional Forces: Soviet Proposals and Western Options," *International Security* 12/4 (1988), 65-66.

[136] U.S. Department of Defense, *Nuclear Posture Review*, 14.

AI and national security, such as addressing the implications of deepfakes and lethal autonomous weapons.[137] To counter the threat posed by non-state actors using AI-enabled tools such as deepfakes to manipulate, deceive, or otherwise interfere with strategic decision-making systems in misinformation attacks, states should - in coordination with both allies and adversaries - continue to harden NC3 systems and processes (e.g., deepfake detection software to detect falsified information).[138] Towards this end, in 2017, NATO established a Strategic Communications Center of Excellence, which supports the development of best practices designed to raising awareness of the risks of disinformation posed by the nefarious dissemination of misinformation.[139] In the emerging deepfake arms race, the prospects for detection appear bleak, however.[140]

Furthermore, United States and China could reconvene the currently suspended Strategic Stability Dialogue (and possibly including Russia) to explore issues including: (1) the impact of AI integration with a range of military (including nuclear) capabilities; (2) the potential for 'new era' AI-infused counterforce and autonomous weapons to unhinge nuclear deterrence; (3) measures to mitigate the risks of inadvertent or accidental nuclear escalation; and (4) promoting collaborative research on AI's impact on international security and safety. It will be critical for all parties to acknowledge, and where possible seek clarification, on divergences in U.S. and Chinese nuclear doctrines, and attitudes to escalation and crisis management, and strategic stability.

Differences between nuclear powers' doctrine and force structure alluded to in this article, coupled with the intrinsic dual-use nature of AI-technologies, will likely narrow the scope for an international agreement on military-use AI - let alone in the strategic domain - at an official diplomatic level. That said, AI might be incorporated into ongoing broader Track 1.5/2 dialogue on cybersecurity and nuclear stability.[141] How impactful these unofficial dialogues are - that is, leading to tangible improvements in security policies - is an open question.

[137] Joshua New, "Why the United States Needs a National Artificial Intelligence Strategy and What It Should Look Like," ITIF December 4, 2018.

[138] For example, Matt Turek, "Semantic Forensics (SemaFor) Proposers Day," *Defense Advanced Research Projects Agency*, August 28, 2019.

[139] "NATO Takes Aim at Disinformation Campaigns," *NPR Morning Edition*, May 10, 2017.

[140] Will Knight, "The U.S. Military Is Funding an Effort to Catch Deepfakes and Other AI Trickery," *MIT Technology Review*, May 23, 2018.

[141] For example, see Michael O. Wheeler, "Track 1.5/2 Security Dialogues with China: Nuclear Lessons Learned," No. IDA-P-5135, *Institute for Defense Analyses*, 2014; and "Track 1.5 U.S.-China Cyber Security Dialogue."

The Chinese government has taken nascent steps to promote research and initiatives on the legal and ethical issues related to AI, including the exploration of rules, safety, regulation, and arms control measures to prevent the potentially destabilizing effects of AI.[142] Despite the backdrop of U.S.-China geopolitical tensions and strategic mistrust, collaboration on specific initiatives may create a foundation for improved understanding and transparency.[143] Strategic competition between great powers to reap the perceived first-mover advantages of AI, without resolving the control and safety issues, will likely become a negative-sum enterprise.[144] *In extremis*, the payoff for all parties might be "minus infinity."[145] Examples of possible arrangements that great powers might find mutually beneficial include a prohibition on the development or deployment of malicious software (i.e., adversarial AI) that targets an adversary's NC3 systems, as well as the use of AI technology to authorize the launch of nuclear weapons.[146] Unlikely as it may be that these agreements could - technically or politically - be verified for formal compliance purposes, a normative framework or understanding would be worthwhile exploring nonetheless.[147]

More broadly, global defense communities should actively invest in the development of AI cyber-defense tools (e.g., analyzing classification errors, automatic detection of remote vulnerability scanning, and model extraction improvements), AI-centric secure hardware (or 'counter AI'), and other fail-safe mechanisms and off-ramps (e.g., circuit breakers), to allow for de-escalation and prevent inadvertent escalation and unintentional consequences. Because multiple extraneous political and strategic factors will be central to any decision to move a situation up the rungs of the

[142] Some have viewed Chinese efforts to regulate the use of military AI internationally as a form of propaganda. Moreover, much of China's AI-related initiatives focus on the impact on social stability and the security of the regime against potential internal threats to its legitimacy. See Gregory C. Allen, "Understanding China's AI Strategy," *Center for a New American Security*, February 6, 2019.

[143] U.S.-China continuous and long-term engagement in scientific and technological collaboration in areas such as global health and climate change demonstrates the potential for cooperation on mutually beneficial policy issues. See, Jennifer Bouey, "Implications of US-China Collaborations on Global Health Issues," Testimony presented before the U.S.-China Economic and Security Review Commission," July 31, 2019.

[144] See, James Johnson, "The end of military-techno Pax Americana? Washington's strategic responses to Chinese AI-enabled military technology," *The Pacific Review* (2010).

[145] Stuart Russell, *Human Compatible*, 183.

[146] Other measures that can be taken to improve stability include reducing the number of nuclear weapons, taking arsenals off high-alert (or launch-on-warning) status, separating warheads from delivery systems (or de-matting warheads), shifting to a deterrent-only (or minimum deterrence) force posture, and adopting a no first use declaratory policy - as China and India does today.

[147] National Security Commission on Artificial Intelligence (NSCAI) Interim Report to Congress, November 2019, 46.

escalation ladder, these kinds of technical controls might not necessarily reduce the risk of escalation.

In addition, applying technological controls will continue to confront the issue of biases and assumptions that are pre-programmed (and often unwittingly) into states' AI algorithms.[148] How useful would existing tools be against vulnerabilities in AI systems? How to tailor these tools for AI systems across multiple military domains? Is there an equivalent to 'patching' in military AI systems? What kinds of policies might incentive, and ensure compliance with, meaningful reforms to existing hardware in the military sphere? How effective would off-ramps and firebreaks be in managing the escalation caused by AI? While these questions are challenging, and necessarily speculative for now, answers should become more evident as the technology matures.

Besides, today, there are no precise definitions or engineering methodology for formulating regulation, or maintaining control over AI systems to ensure their safety - or the AI 'control problem.'[149] For example, AI experts believe that existing tools are unable to resolve the risks posed to humans from AI-augmented autonomous systems.[150] Thus, a different kind of algorithm than currently exists will likely be needed, which prioritizes the safety, robustness, and interpretability of complex AI-infused military systems (e.g., NC3 early warning systems).

One of the most laudable efforts to date to adopt rules for acceptable behavior in the context of emerging technologies was expounded by the United Nations (UN), following the General Assembly resolutions on the topic. Explicitly, the UN expressed general concern that emerging technologies (especially cyber) might be used for nefarious purposes that are "inconsistent with the objectives of maintaining international stability and security," and the body proposed an expert panel to consider "possible cooperative measures to address them, including norms, rules, or principles"

[148] To protect U.S. consumers, the U.S. Senate recently proposed a new bill, which, if passed, would require companies to audit their machine-learning systems for "bias and discrimination,' and take corrective action promptly if such issues were identified. 116th Congress, 1st Session, S. 2065 "The Algorithmic Accountability Act of 2019," United States Government Publishing Office, April 10, 2019.

[149] The AI 'control problem' refers to the issue that under certain conditions, AI systems can learn in unexpected and counterintuitive ways, which may not always align with their goals set by their designers. See, Stuart Russell, *Human Compatible*, 251.

[150] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "Cooperative Inverse Reinforcement Learning."

of states.[151] In 2015, the expert panel articulated a set of core norms to "prevent the proliferation of malicious information and communication technology tools and techniques."[152]

While voluntary, non-binding this basic framework may potentially serve as a useful framework to inform any future arms control discussions between states on AI. Given today's inopportune geopolitical backdrop, the likelihood that Washington and Beijing (or Moscow) would accept international constraints on the use of technology that targets their respective NC3 and other critical infrastructure is remote, however. Nonetheless, the continued efforts by bodies like the UN, prominent commercial and national leaders, to discuss and promote such norms (e.g., clarifying red-lines and rules to enable restraint and tacit bargaining in the digital domain) remains of critical importance.[153] Important questions need to be considered, including: If a machine violates the International Humanitarian Law (IHL) and commits a war crime, for example, who will be held responsible? Should it be the programmer, the operator, the civilian authorities who gave the order, or the commander who decided to field the machine to fulfill that particular operation?[154]

Conclusion

This article considered the risks and trade-offs of increasing the role of machines in the strategic decision-making process, and the impact of synthesizing AI with NC3 early-warning systems for the nuclear enterprise. Despite the general agreement among nuclear-armed states that decision-making that directly impacts nuclear C2 architecture should not be pre-delegated to machines, once AI demonstrates an apparent capacity to engage at a superhuman level in the execution of strategic planning, defense planners may become more predisposed to view decisions generated by AI algorithms as

[151] UN General Assembly, "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security: Note by the Secretary-General," A/68/98, June 24, 2013.

[152] UN General Assembly, "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security: Note by the Secretary-General," A/70/174, July 22, 2015.

[153] A potential precedent for such high-level agreement is the 2015 U.S.-China agreement that prohibited the use of cyberspace for the theft of intellectual property. Notwithstanding the controversy about China's adherence to the agreement, it was generally accepted that Chinese cyberespionage in the United States subsided in the immediate aftermath of this deal. Adam Segal, "Is China Still Stealing Western Intellectual Property?" *Council on Foreign Relations*, September 26, 2018.

[154] The IHL's core objectives guiding the use of military force includes distinguishing between combatants and civilians; only based on military necessity; is it proportional in terms of the military gains versus the cost imposed on civilians and uses all practical precautions to help avoid tragedy.

analogous (or even superior) to humans. This predisposition would likely be influenced, or even expedited by, the anthropomorphic tendencies of human psychology. It would surely be a mistake to presume that an absolution of responsibility and accountability necessarily accompanies this transfer of authority.[155]

While human agency ensures that the role for AI in the nuclear domain is confined to a predominately tactical utility - through the discharge of its 'support role' - it could still nonetheless influence strategic decisions that involve nuclear weapons. That is, the distinction between the impact of AI at a tactical and strategic level is *not a binary one*: technology designed to augment autonomous tactical weapons ostensibly will be making decisions in the use of lethal force that informs and shape overarching strategic war-faring calculations.

This non-binary distinction could risk AI-powered decision support systems substituting the role of critical thinking, empathy, creativity, and intuition of human commanders in the strategic decision-making process. Unlikely as it is that defense planner would delegate - at least knowingly - the authority of missile launch platforms, delivery systems, or NC3 to machines, AI technology is expected to be more widely used to support decision-making on strategic nuclear issues - or decision support systems. In short, nuclear-states face a trade-off in not only whether to use AI-enabled decision-support tools, but also how these systems are calibrated to reflect states' risk tolerance (i.e., for false positives vs. false negatives) and confidence in their second-strike capabilities.[156] *Ceteris paribus*, a state more confident in its ability to retaliate in response to a first strike will be more inclined to design their NC3 systems in ways that do not over-rely on autonomous systems.

This will likely be a double-edged sword for stability in the nuclear enterprise. On the one hand, improvements could increase the states' confidence in their nuclear systems, reassure leaders that an adversary is not planning to launch a preemptive strike, and thus, improve strategic stability. For example, bolstering NC3 defenses against physical and cyber; improving information flow and situational awareness; enhancing nuclear warning and testing systems, thus, making accidents caused by errorless likely, and reducing the risk of human error caused by repetitive tasks and fatigue, and

[155] David Watson, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence," 434-435.
[156] Michael C. Horowitz, Paul Scharre, and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," 18.

expanded use of automation in NC3 could reduce the risk of accidents caused by unauthorized use.

On the other hand, these developments could increase escalation risks through two key vectors. First, machine learning enhanced cyber-attacks would be considerably more challenging to detect, and thus more effective. Second, in the unlikely event, an attack was successfully detected, threat identification at machine speed would be virtually impossible. In sum, the synthesis of AI - and other emerging technologies such as quantum communication, 5G networks, and big-data analytics - into nuclear early-warning systems could further compress the decision-making timeframe, create new network vulnerabilities, thereby eroding crisis stability.

During nuclear brinksmanship, the interaction between machines and human-strategic psychology might increase the risk of misperceiving an adversary's intentions, thus increasing first-strike incentives and undermining crisis stability. Furthermore, AI-supported NC3 systems would be more vulnerable to subversion from cyberattacks, which could increase the risk of inadvertent escalation - as a result of either human or machine error.[157] To reduce the perceived vulnerability of U.S. NC3 systems to cyberattacks, for example, the DoD recently proposed a substantial investment to upgrade its NC3 infrastructure.[158]

In addition to the nuclear interactions between nuclear-armed dyads, the use of AI-enhanced fake news, deepfakes, bots, and other malevolent social media campaigns by non-state actors, terrorists, and state proxy actors might also have destabilizing implications for the many nuclear interactions in the world between nuclear and non-nuclear states. In particular, false social media reports (e.g., reports of mobile missiles movements, real-time streaming of launches, the deployment of TELs, or possible false reports of detonations) may influence the threat sensors of nuclear early warning systems used to inform strategic decision-making. *In extremis*, nuclear confrontation could result from false, fabricated, or misperceived narratives. Because nuclear interactions increasingly involve the complex interplay of nuclear and non-nuclear (and non-state) actors, the leveraging of AI in this multipolar context will increasingly place destabilizing pressures on nuclear states.

[157] For example, the destruction of the Chinese embassy in Belgrade by the United States in 1999 illustrates that accidents in the military arena have broader and long-term geopolitical and geostrategic implications.

[158] U.S. Office of the Secretary of Defense, *Nuclear Posture Review*.

In combination, an overreliance on AI-enhanced systems (or automation bias) and the risk of false alarms (i.e., especially false positives) in cyberspace might cause states to exaggerate a threat posed by ambiguous or manipulated information, increasing instability. As new and increasingly complex technologies like AI are superimposed on states' legacy NC3 systems, more innovative types of errors, distortions, and manipulations seem more likely to occur. Future research would be beneficial on how states - both nuclear-armed and non-nuclear armed - might mitigate the potentially escalatory risks posed by AI and steer it to bolster strategic stability as the technology matures, and how non-state and third-party actors' pursuit of AI technology might threaten the strategic environment of nuclear-armed powers.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Dr. James Johnson is a Postdoctoral Research Fellow at the James Martin Center for Non-Proliferation Studies (CNS) at the Middlebury Institute of International Studies, Monterey. James holds a Ph.D. in Politics & International Relations from the University of Leicester, where he is also an honorary visiting fellow with the School of History & International Relations. He is the author of *The US-China Military & Defense Relationship during the Obama Presidency*. His latest book project is entitled, *Artificial Intelligence & the Future of Warfare: USA, China, and Strategic Stability*. James is fluent in Mandarin.

Contact jsjohnson@middlebury.edu

Bibliography

Acton, James M., "Escalation through Entanglement: How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War," *International Security* 43/1 (2018), 56–99.

Acton, James M., *Is this a Nuke? Pre-launch Ambiguity and Inadvertent Escalation* (New York, NY: Carnegie Endowment for International Peace, 2020).

Adamsky, Dima, "Cross-Domain Coercion: The Current Russian Art of Strategy," *IFRI Proliferation Paper* 54 (2015), 1-43.

Allen, Gregory C., "Understanding China's AI Strategy," *Center for a New American Security*, February 6, 2019, https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy

AlphaStar Team, “Alphastar: Mastering the Real-Time Strategy Game Starcraft II,” *DeepMind Blog,* January 24, 2019, https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

Anderson, Hyrum S., Kharkar, Anant, Filar, Bobby, and Roth, Phil, “Evading Machine Learning Malware Detection,” *blackhat.com*, July 20, 2017, https://www.blackhat.com/docs/us-17/thursday/us-17-Anderson-Bot-Vs-Bot-Evading-Machine-Learning-Malware-Detection-wp.pdf

Ayoub, Kareem, and Payne, Kenneth, “Strategy in the Age of Artificial Intelligence,” *Journal of Strategic Studies* 39/5-6, (2016), 793-819.

Biddle, Stephen D., and Zirkle, Robert, “Technology, Civil-Military Relations, and Warfare in the Developing World,” *Journal of Strategic Studies* 19/2 (1996), 171-212.

Botsman, Rachel, “Big data meets Big Brother as China moves to rate its citizens,” *Wired*, October 21, 2017, https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion

Bouey, Jennifer, “Implications of US-China Collaborations on Global Health Issues,” Testimony presented before the U.S.-China Economic and Security Review Commission,” July 31, 2019, https://www.rand.org/pubs/testimonies/CT516.html

Boulanin, Vincent (ed.) *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. I Euro-Atlantic Perspectives* (SIPRI Publications, Stockholm: May 2019).

Bracken, Paul, “The Cyber Threat to Nuclear Stability,” *Orbis* 60/2 (2016), 194, 1-16.

Brandtzaeg, Petter Bae, “Facebook Is No “Great Equalizer”: A Big Data Approach to Gender Differences in Civic Engagement Across Countries,” *Social Science Computer Review* 35/ 1 (2017), 103-25.

Brewer, Marilynn B., and Crando, William B., *Social Psychology* (New York, NY: West Publishing Co. 1994).

Bruce Blair, *Strategic Command and Control: Redefining the Nuclear Threat* (Washington DC: Brookings Institution, 1985).

Center for a New American Security, University of Oxford, University of Cambridge, Future of Humanity Institute, OpenAI & Future of Humanity Institute, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, (Oxford, UK: Oxford University, February 2018).

Cimbala, Stephen J., *The Dead Volcano: The Background and Effects of Nuclear War Complacency*, (New York, NY: Praeger, 2002).

Citron, Danielle K., and Chesney, Robert, "*Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*," 107 California Law Review 1753 (2019).

Connable, Ben, *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*, (Santa Monica, CA: RAND Corporation, 2012).

Cummings, Mary L., "Automation Bias in Intelligent Time-Critical Decision Support Systems," *AIAA 1st Intelligent Systems Technical Conference*, 2004, 557-562.

Cummings, Mary L., *Artificial intelligence and the future of warfare*, (London, UK: Chatham House, 2017).

Cunningham, Fiona S. and Fravel, M. Taylor, "Dangerous Confidence? Chinese Views on Nuclear Escalation," *International Security* (2019), 44/2, 106-108.

Davidow, Bill, "Exploiting the Neuroscience of Internet Addiction," *The Atlantic*, July 18, 2012, https://www.theatlantic.com/health/archive/2012/07/exploiting-the-neuroscience-of-internet-addiction/259820/

Dear, Keith, "Artificial Intelligence and Decision-Making," *The RUSI Journal*, 164/5-6, 18-25.

Del Vicario, Michela, Scala, Antonio, Caldarelli, Guido, Stanley, Eugene H., and Quattrociocchi, Walter, "Modeling confirmation bias and polarization," *Sci Rep*, January, 11, 2017 doi: 10.1038/srep40391

Erwin, Sandra, "BAE wins DARPA contract to develop machine learning technology for space operations," *Spacenews,* August 13, 2019, https://spacenews.com/bae-wins-darpa-contract-to-develop-machine-learning-technology-for-space-operations/

Etherington, Darrell, "Google's AlphaGo AI Beats the World's Best Human Go Player," May 23, 2017, *Techcrunch* https://techcrunch.com/2017/05/23/googles-alphago-ai-beats-the-worlds-best-human-go-player/

Everstine, Brian W., "DOD AI Leader Wants Closer Collaboration With NATO," Airforce Magazine, January 15, 2020, https://www.airforcemag.com/dod-ai-leader-wants-closer-collaboration-with-nato/

Fisher, Kathleen, "Using formal methods to enable more secure vehicles: DARPA's HACMS program," ICFP 2014: Proceedings of the 19th ACM SIGPLAN International Conference on Functional Programming, https://dl.acm.org/doi/10.1145/2628136.2628165

Fitzpatrick, Mark, "Artificial Intelligence and Nuclear Command and Control," *Survival*, 61/3 (2019), 81-92.

Forden, Geoffrey, Podvig, Pavel, and Postol Theodore A., "False Alarm, Nuclear Danger," *IEEE Spectrum* 37/3 (2000), 31-39.

Futter, Andrew, *Hacking the Bomb: Cyber Threats and Nuclear Weapons* (Washington DC: Georgetown University Press, 2018).

Gasser, Paige, Loss, Rafael, and Reddie, Andrew, "Assessing the Strategic Effects of Artificial Intelligence - Workshop Summary," *Center for Global Security Research, Lawrence Livermore National Laboratory*, September 2018.

Geist, Edward and, Lohn, Andrew, *How might artificial intelligence affect the risk of nuclear war?* (Santa Monica, CA: RAND Corporation, 2018).

Generating Actionable Understanding of Real-World Phenomena with AI," *DARPA* January 4, 2019, https://www.darpa.mil/news-events/2019-01-04

Graves, Alexander, and Clancy, Kelly, "Unsupervised learning: The curious pupil," *Deepmind*, June 25, 2019, https://deepmind.com/blog/article/unsupervised-learning

Gregory, Shaun, *The Hidden Cost of Deterrence: Nuclear Weapons Accidents* (London: Brassey's, 1990).

Hadfield-Menell, Dylan, Dragan, Anca, Abbeel, Pieter, and Russell, Stuart, "Cooperative Inverse Reinforcement Learning," 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, https://people.eecs.berkeley.edu/~dhm/papers/CIRL_NIPS_16.pdf

Harwell, Drew, "An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft," *Washington Post*, September 4, 2019, https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/

Horowitz, Michael C., Scharre, Paul, and Velez-Green, Alexander, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," December 2019, arXiv:1912.05291

Horowitz, Michael, Scharre, Paul, and Velez-Green, Alex, *A Stable Nuclear Future? The Impact of Automation, Autonomy, and Artificial Intelligence* (Philadelphia: University of Pennsylvania, 2017).

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg, "AI Safety Gridlocks," November 27, 2017.

Jensen, Benjamin M., et al., "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review*, (June 2019), doi.org/10.1093/isr/viz025

Jervis, Robert, and Rapp-Hooper, Mira, "Perception and Misperception on the Korean Peninsula," *Foreign Affairs*, April 5, 2018, https://www.foreignaffairs.com/articles/north-korea/2018-04-05/perceptionand-misperception-korean-peninsula.

Jervis, Robert, *How Statesmen Think: The Psychology of International Politics,* (New Jersey, Princeton: Princeton University Press, 2017.

Jervis, Robert, *Perception and Misperception in International Politics*, (Princeton, N.J.: Princeton University Press), 1976.

Jervis, Robert, *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon* (Ithaca, NY: Cornell University Press, 1989).

Jia Daojin and Zhou Hongmei, "The Future 20-30 Years Will Initiate Military Transformation," *China Military Online*, June 2, 2016, http://www.81.cn/jmywyl/2016-06/02/content_7083964.htm

Johnson, James, "Chinese Nuclear 'War-fighting:' "An Emerging Intense U.S.-China Security Dilemma and Threats to Crisis Stability in the Asia Pacific," *Asian Security*, 15/3 (2019), 215-232.

Johnson, James, "The AI-cyber nexus: implications for military escalation, deterrence, and strategic stability," *Journal of Cyber Policy*, 4/3, 442-460.

Johnson, James, "Artificial Intelligence: A Threat to Strategic Stability," *Strategic Studies Quarterly*, 14/1, 16-39.

Johnson, James, "Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare," *The RUSI Journal* (2020) DOI: 10.1080/03071847.2020.1752026

Johnson, James, "The end of military-techno Pax Americana? Washington's strategic responses to Chinese AI-enabled military technology," *The Pacific Review* (2020) DOI: 10.1080/09512748.2019.1676299

Johnson, Joseph, "MAD in an AI Future?" Center for Global Security Research, Lawrence Livermore National Laboratory, June 3, 2019, https://cgsr.llnl.gov/content/assets/docs/MAD-in-an-AI-Future.pdf

Joshua New, "Why the United States Needs a National Artificial Intelligence Strategy and What It Should Look Like," ITIF December 4, 2018, https://itif.org/publications/2018/12/04/why-united-states-needs-national-artificial-intelligence-strategy-and-what

Knight, Will, "The U.S. Military Is Funding an Effort to Catch Deepfakes and Other AI Trickery," *Technology Review*, May 23, 2018, https://www.technologyreview.com/2018/05/23/142770/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/

Kott, Alexander, and Ownby, Michael, "Tools for Real-Time Anticipation of Enemy Actions in Tactical Ground Operations," *Defense Technical Information Center*, June 2005, https://apps.dtic.mil/dtic/tr/fulltext/u2/a460912.pdf

Krepon, Michael, "Can deterrence ever be stable?" *Survival*, 57/ 3 (2015) 111-132.

Krimsky, Sheldon, Genetic alchemy: The social history of the recombinant DNA controversy (MIT Press, Cambridge, MA, 1962).

Kristensen, and Jakob Bæk et al., "Parsimonious Data: How a Single Facebook Like Predicts Voting Behavior in Multiparty Systems," *PLOS One* 12/9, (2017).

Lamothe, Dan, "U.S. families got fake orders to leave South Korea. Now counterintelligence is involved," *The Washington Post*, September 22, 2017, https://www.washingtonpost.com/news/checkpoint/wp/2017/09/22/u-s-families-got-fake-orders-to-leave-south-korea-now-counterintelligence-is-involved/

Larsen, Jeffrey, "Nuclear Command, Control, and Communications: U.S. Country Profile," *Tech4GS*, August 22, 2019, https://www.tech4gs.org/jeffrey-larsen.html

Leonid, Ryabikhin, "Russia's NC3 and Early Warning Systems," *Tech4GS* July 11, 2019, https://www.tech4gs.org/nc3-systems-and-strategic-stability-a-global-overview.html

Lieber, Keir, and Press, Daryl G., "The New Era of Counterforce: Technological Change and the Future of Nuclear Deterrence," *International Security*, 41/4 (2017), 9-49.

Lin, Herbert "Escalation Dynamics and Conflict Termination in Cyberspace," *Strategic Studies Quarterly*, 6/3, (Fall, 2012), 46-70.

Linda J Skitka, Kathleen L Mosier, and Mark Burdick, "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies* 51/5 (1999), 94-96.

Lindsay, Jon R., "Cyber Operations and Nuclear Weapons," *Tech4GS Special Reports*, June 20, 2019, *Tech4GS Special Reports*, June 20, 2019, https://www.tech4gs.org/nc3-systems-and-strategic-stability-a-global-overview.html

Lindsay, Jon, R., and Gartzke, Erik, eds. 2019 *Cross-domain deterrence: Strategy in an era of complexity* (Oxford: Oxford University Press, 2019).

Maas, Matthijs M., "How viable is international arms control for artificial military intelligence? Three lessons from nuclear weapons, *Contemporary Security Policy* 40/3 (2019), 285-311.

Majumdar, Dave, "USAF leader confirms manned decision for new bomber," *Flight International*, April 23, 2013, https://www.flightglobal.com/usaf-leader-confirms-manned-decision-for-new-bomber/109511.article

Metz, Cade, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future," *Wired*, March 16, 2016, https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

Morozov, Evgeny, *The Net Delusion: The Dark Side of Internet Freedom* (New York, NY: PublicAffairs, 2012).

Nass, Clifford, Fogg, B.J, and Moon, Youngme, Y., "Can computers be teammates?" *International Journal of Human Computer Studies* 45, (1996), 669-678.

National Security Commission on Artificial Intelligence (NSCAI) Interim Report to Congress, November 2019, https://epic.org/privacy/NSCAI-initial-report-073119.pdf

Nautilus Institute, Technology for Global Security, Preventive Defense Project, "Social Media Storms and Nuclear Early Warning Systems: A Deep Dive and Speed Scenario Workshop," *NAPSNet Special Reports*, January 8, 2019, https://nautilus.org/napsnet/napsnet-special-reports/social-media-storms-and-nuclear-early-warning-systems-a-deep-dive-and-speed-scenarios-workshop-report/

Nilsson, Nils J., *The Quest for Artificial Intelligence*, (New York, NY: Cambridge University Press, 2010)

Oliphant, Roland, "Secret Russian radioactive doomsday torpedo leaked on television," *Daily Telegraph*, November 15, 2015, https://www.telegraph.co.uk/news/worldnews/europe/russia/11991729/Secret-Russian-radioactive-doomsday-torpedo-leaked-on-television.html

Papernot, Nicolas, McDaniel, Patrick, and Goodfellow, Ian, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples," *arXiv*, May 24, 2016, https://arxiv.org/abs/1605.07277

Payne, Kenneth, "Fighting On: Emotion and Conflict Termination," *Cambridge Review of International Affairs* 28/3 (August 2015), 480-97.

Payne, Kenneth, *Strategy from Apes to Evolution Artificial Intelligence and War,* (Washington, DC: Georgetown University Press).

Polyakova, Alina, "Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare," *Brookings*, November 15, 2018, https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/

Raja, Parasuraman, and Riley Victor, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, 52/3 (2010), 381-410.

Raja, Parasuraman, and Riley Victor, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* 39/2 (June 1997), 230-53.

Rubin, Jamie, "Deception: The Other 'D' in D-Day," *NBC News*, June 5, 2004, http://www.nbcnews.com/id/5139053/ns/msnbc-the_abrams_report/t/deception-other-d-d-day/

Russell, Stuart, and Norvig Peter, *Artificial Intelligence: A Modern Approach*, 3rd edn (Pearson Education: Harlow, 2014).

Russell, Stuart, *Human Compatible* (New York, NY: Viking Press, 2019).

Saalman, Lora, “Lora Saalman on How Artificial Intelligence Will Impact China’s Nuclear Strategy,” *The Diplomat*, November 7, 2018, https://thediplomat.com/2018/11/lora-saalman-on-how-artificial-intelligence-will-impact-chinas-nuclear-strategy/

Sagan, Scott D., *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton, NJ: Princeton University Press, 1995).

Schelling, Thomas C., and Halperin, Morton H., *Strategy and Arms Control* (New York: The Twentieth Century Fund 1961).

Scherer, Matthew U., “Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies,” *Harvard Journal of Law and Technology*, 29/2, (2016), 354-400.

Schlosser, Eric, *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety* (New York: Penguin, 2014).

Sechser, Todd S., Narang, Neil, and Talmadge, Caitlin, “Emerging technologies and strategic stability in peacetime, crisis, and war,” *Journal of Strategic Studies*, 42/6, 727-735.

Segal, Adam, "Is China Still Stealing Western Intellectual Property?" *Council on Foreign Relations*, September 26, 2018, https://www.cfr.org/blog/china-still-stealing-western-intellectual-property

Col. (Ret.) Sharpe, James D. Jr., and Lt. Col. (Ret.), and Creviston, Thomas E., “Understanding mission command,” *U.S. Army*, July 10, 2013, https://www.army.mil/article/106872/understanding_mission_command

Shehadeh, Karim K., “The Wassenaar Arrangement and Encryption Exports: An Ineffective Export Control Regime that Compromises United States Economic Interests,” *American University of International Law Review*, 15/1 (1999), 271-319.

Singer, Peter W., “Robots and the rise of “tactical Generals,” *Brookings*, March 9, 2009, https://www.brookings.edu/articles/robots-and-the-rise-of-tactical-generals/

Singer, Peter W., *Wired for War* (New York, NY: Penguin Group, 2009).

Skitka, Linda J, and Mosier, Kathleen L., and Burdick, Mark, “Does Automation Bias Decision-Making?” *International Journal of Human-Computer Studies* 51/5 (1999), 991-1006.

Snyder, Jack, “Limiting Offensive Conventional Forces: Soviet Proposals and Western Options,” *International Security* 12/4 (1988), 65-66.

Sparrow, Rob, “Ethics as a source of law: The Martens Clause and autonomous weapons,” *ICRC Blog*, November 14, 2017, https://blogs.icrc.org/law-and-policy/2017/11/14/ethics-source-law-martens-clause-autonomous-weapons/

Turek, Matt, "Semantic Forensics (SemaFor) Proposers Day," *Defense Advanced Research Projects Agency*, August 28, 2019, https://www.darpa.mil/news-events/semantic-forensics-proposers-day

Turner, Jacob, *Robot Rules: Regulating Artificial Intelligence* (London, Palgrave Macmillan, 2019).

U.S. Office of the Secretary of Defense, *Nuclear Posture Review* (Washington DC: Department of Defense, February 2018), 57-58.

U.S. Office of the Secretary of Defense, *Nuclear Posture Review* (Washington DC: Department of Defense, February 2018), https://media.defense.gov/2018/Feb/02/2001872886/-1/-1/1/2018-NUCLEAR-POSTURE-REVIEW-FINAL-REPORT.PDF

UN General Assembly, "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security: Note by the Secretary-General," A/68/98, June 24, 2013, https://digitallibrary.un.org/record/753055?ln=en

UN General Assembly, "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security: Note by the Secretary-General," A/70/174, July 22, 2015, https://digitallibrary.un.org/record/799853?ln=en

Vergun, David, "Defense Innovation Broad Recommends AI Ethical Guidelines," *U.S. Dept of Defense*, November 1, 2019, https://www.defense.gov/Explore/News/Article/Article/2006646/defense-innovation-board-recommends-ai-ethical-guidelines/

Vienna Document 2011 on Confidence-and Security-Building Measures (Vienna: Organization for Security and Co-operation in Europe, 2011), https://2009-2017.state.gov/t/avc/cca/c43837.htm

Wallach, Wendell, and Allen, Colin, *Moral Machines* (New York, NY: Oxford University Press, 2009).

Wang, Raymond, "Quantum Communications and Chinese SSBN Strategy," *The Diplomat*, November 4, 2017, https://thediplomat.com/2017/11/quantumcommunications-and-chinese-ssbn-strategy/

Watson, David, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence," *Minds and Machines* (2019) 29, 417-440.

Wheeler, Michael O., "Track 1.5/2 Security Dialogues with China: Nuclear Lessons Learned," No. IDA-P-5135, *Institute for Defense Analyses*, 2014; and "Track 1.5 U.S.-China
Cyber Security Dialogue," https://www.ida.org/-/media/feature/publications/t/tr/track-152-security-dialogues-with-china-nuclear-lessons-learned/p-5135.ashx

Whetham, David, and Payne, Kenneth, “AI: In Defence of Uncertainty,” *Defence in Depth*, December 9, 2019, https://defenceindepth.co/2019/12/09/ai-in-defence-of-uncertainty/

Williams, Heather, “Asymmetric Arms Control and Strategic Stability: Scenarios for Limiting Hypersonic Glide Vehicles,” *Journal of Strategic Studies* 42/6 (2019), 789–813.

Wong, Yuna Huh, Yurchak, John M., Button, Robert W., Frank, Aaron, Laird, Burgess, Osoba, Osonde A., Steeb, Randall, Harris, Benjamin N., and Joon Bae, Sebastian, *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020).

Woolf, Amy, *Defense Primer: Command and Control of Nuclear Forces* (Washington, DC: Congressional Research Service), December 11, 2018, https://fas.org/sgp/crs/natsec/IF10521.pdf

Yarhi-Milo, Keren, *Knowing the Adversary* (Princeton NY: Princeton University Press, 2014).

Yuan Yi, “The Development of Military Intelligentization Calls for Related International Rules,” *PLA Daily*, October 16, 2019, http://military.workercn.cn/32824/201910/16/191016085645085.shtml

“Chinese commercial space start-ups launch two AI satellites in a hundred days,” *Global Times*, November 26, 2018, http://smart.huanqiu.com/ai/2018-11/13645096.html?agt=15422

“NATO Takes Aim at Disinformation Campaigns,” *NPR Morning Edition*, May 10, 2017, https://www.npr.org/2017/05/10/527720078/nato-takes-aim-at-disinformation-campaigns

116th Congress, 1st Session, S. 2065 “The Algorithmic Accountability Act of 2019,” United States Government Publishing Office, April 10, 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231